



FACULDADE DE MATEMÁTICAS

Trabalho Fin de Grao

INTRODUCCIÓN AOS MODELOS MIXTOS

Diego Losada González

2021/2022

UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

GRAO DE MATEMÁTICAS

Traballo Fin de Grao

INTRODUCCIÓN AOS MODELOS MIXTOS

Diego Losada González

Xullo, 2022

UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Traballo proposto

Área de Coñecemento: Estatística e Investigación Operativa
Título: Introducción aos Modelos Mixtos
Breve descripción do contido
<p>As estruturas xerárquicas de datos (estructuras multinivel) son frecuentes nas Ciencias Sociais, a Medicina ou a Bioloxía. Na formulación de estruturas xerárquicas asúmese que cada individuo pertence a un único grupo e o obxectivo é analizar as relacións a dous niveis: entre os grupos e dentro dos mesmos.</p> <p>A análise, tanto descritiva como inferencial, deste tipo de poboacións con estruturas complexas é o obxecto dos denominados modelos multinivel. Cabe sinalar que os modelos multinivel, terminoloxía que provén da Estatística Educacional, tamén se coñecen como modelos lineais xerárquicos ou modelos mixtos (Estatística, Bioestatística), modelos de efectos aleatorios ou modelos de coeficientes aleatorios (Econometría) ou modelos de compoñentes da varianza (Diseño de experimentos).</p> <p>Breve planificación:</p> <p>A modo de orientación, o traballo podería organizarse nas seguintes seccións:</p> <ul style="list-style-type: none">▪ Modelo de análise da varianza: ANOVA▪ Modelo de análise da varianza con efectos aleatorios: RANOVA▪ Introducción aos modelos multinivel con resposta continua.

Ademais, presentaremos diferentes modelos mixtos aplicados tanto a conxuntos de datos ou a datos simulados. Para iso, empregaremos o software estatístico libre R (https://www.r-project.org/).
--

Recomendacións

Outras observacións

Índice

Resumo	IX
Introdución	XI
1. ANOVA e ANCOVA	1
1.1. O modelo <i>ANOVA</i>	1
1.1.1. <i>ANOVA</i> como modelo linear xeral	1
1.1.2. Estimación dos parámetros	2
1.1.3. Análise da varianza e test F	5
1.1.4. Comparacións múltiples	10
1.2. O modelo <i>ANCOVA</i>	12
1.2.1. <i>ANCOVA</i> sen interacción	12
1.2.2. Estimación dos parámetros	13
1.2.3. <i>ANCOVA</i> con interacción	13
2. Introducción aos modelos mixtos	15
2.1. Datos multinivel	15
2.2. A necesidade de ter en conta os distintos niveis	17
3. RANOVA	19
3.1. Análise da varianza con efectos aleatorios	19


3.2. Estimación dos parámetros	21
3.2.1. Estimación da media global μ	21
3.2.2. Estimación das compoñentes da varianza	22
3.2.3. Predición dos efectos aleatorios	26
3.3. Contraste sobre os efectos grupais	32
4. Modelos mixtos con covariables relativas ao primeiro nivel	35
4.1. Modelo con intercepto aleatorio	35
4.2. Modelo con intercepto e pendente aleatorios	40
5. Modelos mixtos con covariables relativas ao segundo nivel	47
5.1. Variables contextuais composicionais	48
5.2. Variables contextuais globais	52
5.3. Interacción entre niveis	54
6. Conclusións	55
A. Código de R	57
A.1. Creación da base de datos <code>mates</code>	57
A.2. Introducción	58
A.2.1. Figura 1	58
A.3. Creación da base de datos <code>mates7</code>	59
A.4. ANOVA	60
A.4.1. Figura 1.1	60
A.4.2. Función <code>Bonferroni_taboa</code>	61
A.4.3. Test F e contrastes pareados	63
A.5. ANCOVA	63
A.5.1. Figura 1.2	63

A.6. RANOVA	65
A.6.1. VARCOMP e contraste sobre os efectos das escolas	66
A.6.2. Figura 3.5	66
A.6.3. Conxunto de datos u0df	66
A.6.4. Figura 3.3: Gráfico de eiruga	67
A.6.5. Figura 3.2	68
A.7. Modelos mixtos con covariables relativas ao nivel 1	71
A.7.1. Modelo con intercepto aleatorio e pendente fixa	71
A.7.2. Modelo con intercepto aleatorio e pendente fixa con SocialMin	71
A.7.3. Modelo con intercepto e pendente aleatorios e mais variable categórica . .	72
A.7.4. Figura 4.3	72
A.8. Modelos mixtos con covariables relativas ao nivel 2	76
A.8.1. Variable contextual composicional	76
A.8.2. Variable contextual global	77
Bibliografía	79

Resumo

No eido da Estatística, os modelos de regresión son a principal ferramenta empregada cando o que se precisa é estimar a relación entre variables aleatorias. En concreto, veremos como unha ou varias variables (que chamaremos variables explicativas) inflúen sobre outra variable (que chamaremos variable resposta).


Moitas bases de datos concernentes ao eido da Educación, a Medicina ou as Ciencias Medioambientais están xerarquicamente organizadas debido á propia natureza destas, de xeito que os individuos se atopan aniñados en grupos; como por exemplo, un conxunto de alumnas/os agrupadas/os por escolas. É obvio pensar que individuos clasificados nun mesmo grupo tenderán a ter un comportamento máis semellante que uns individuos calesquera de grupos diferentes, con menos información en común. Nestes casos, os modelos de regresión clásicos deixan de ser útiles e xorde a necesidade de ter en conta o efecto que producen estas agrupacións na variable resposta. As primeiras propostas para estudar este tipo de datos, sen ignorar as agrupacións existentes, son os modelos de análise da varianza (coñecidos como modelos *ANOVA*) ou modelos de análise da covarianza (coñecidos como modelos *ANCOVA*); mais estes modelos só son interesantes cando o que se quere é aplicar técnicas da Inferencia Estatística sobre certas características dos grupos presentes na base de datos.

Afondando aínda máis na análise de datos xerárquicos, os grupos presentes no conxunto de datos poden considerarse unha mostra aleatoria dunha poboación máis grande de grupos para facer Inferencia sobre os grupos en xeral. Neste caso, os modelos de regresión *ANOVA* e *ANCOVA* deixan de ser válidos, e xorden os denominados modelos mixtos ou modelos multinivel. Ao longo deste traballo introducíranse os modelos mixtos e pónerase de manifesto a súa utilidade para estudar bases de datos cunha estrutura de dous niveis, onde os individuos se atopan no primeiro nivel e están aniñados en grupos no segundo nivel, mediante a incorporación de efectos aleatorios. Para levar a cabo esta ilustración empregárase unha base de datos reais que será analizada empregando a ferramenta estatística .

Abstract

In the Statistical field, regression models are the main tool employed when estimating the relation among random variables is needed. In particular, we will see the effect of one or several variables (that will be denoted by explanatory variables) in another variable (that will be denoted by response variable).

A lot of databases concerning the fields of Education, Medicine or Environmental Sciences are hierarchically organized due to their own nature, so that individuals are organized in groups; for example, a set of students grouped by schools. It is obvious to think that individuals classified in a same group will tend to have a more similar behaviour than any individual from different groups, with less information in common. In these cases, classical regression models stop being useful and the necessity to take into account the effect produced by these groupings in the response variable arises. The first proposals to study this type of data sets, without ignoring the existing groupings, are the models of analysis of variance (*ANOVA*) or models of analysis of covariance (*ANCOVA*); but these models are only interesting when the goal is to apply Statistical Inference techniques on certain features of the groups present in the database.

Going even further in the analysis of hierarchical data, the present groups in the dataset can be considered a random sample of a bigger population of groups to make Inference about all groups in general. In this case, the regression models *ANOVA* and *ANCOVA* stop being valid, and the named mixed models or multilevel models arises. Throughout this work, mixed models will be introduced and it will be presented their utility in the study of databases with a structure of two levels, where individuals are on the first level and are nested in groups on the second level, by incorporating random effects. To carry out this illustration, it will be used a real database that will be analysed employing the statistical tool .

Introdución

Un modelo de regresión ten como obxectivo explicar a relación de dependencia existente entre unha variable Y , que se chamará variable resposta, e unha ou máis variables X , que se denominarán variables explicativas. O modelo máis sinxelo que se pode considerar é o que se coñece como **modelo de regresión linear simple**, que relaciona unha variable resposta continua Y cunha soa variable explicativa continua X mediante unha recta. Tal modelo pode expresarse como

$$Y = \beta_0 + \beta_1 X + \varepsilon, \quad (1)$$

onde β_0 é o intercepto, β_1 é a pendente e ε coñécese como o erro do modelo, que verifica a condición $\mathbb{E}(\varepsilon|X = x) = 0 \ \forall x \in \mathbb{R}$ e contén a variabilidade de Y que non se explica grazas a X .

Para poder aplicar técnicas de Inferencia Estatística sobre o modelo (1), necesitarase asumir as seguintes hipóteses:

- **Linearidade.** A función de regresión é unha recta (co cal ten senso escribir (1)).
- **Homocedasticidade.** A varianza do erro é a mesma para calquer valor de X , isto é, $Var(\varepsilon_i) = \sigma^2$ para todo $i = 1, \dots, n$, sendo n o número de observacións.
- **Normalidade.** Os erros están identicamente distribuídos do xeito $\varepsilon \in N(0, \sigma^2)$ ¹.
- **Independencia.** $\varepsilon_1, \dots, \varepsilon_n$ son mutuamente independentes².

A recta $y = \beta_0 + \beta_1 x$ é a liña de regresión teórica. Dada unha mostra aleatoria simple $(x_1, Y_1), \dots, (x_n, Y_n)$; para obter a recta de regresión axustada que aproxima a relación entre as

¹Con $Z \in N(\mu, \sigma^2)$ denotarase unha variable aleatoria continua que segue unha **distribución normal univariante** de media μ e varianza σ^2 , e a súa función de densidade vén dada por

$$f_Z(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left\{ -\frac{1}{2} \left(\frac{z - \mu}{\sigma} \right)^2 \right\}; \quad z \in \mathbb{R}.$$

Esta función ten forma de campá, é simétrica entorno a μ e ten puntos de inflexión $\mu - \sigma$ e $\mu + \sigma$. Se $\mu = 0$ e $\sigma = 1$, a distribución chámase **normal estándar univariante**.

²A independencia mutua é a situación de maior independencia, xa que cada par de subvectores aleatorios serán independentes entre si.

variables X e Y precísanse estimar os parámetros do modelo, que se veñen denotando por β_0 e β_1 . Así, se os estimadores dos parámetros fosen $\hat{\beta}_0$ e $\hat{\beta}_1$, para un valor de $X = x$; daríase a predición $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x$ para Y , obtendo así unha recta de predicións, precisamente a recta axustada polo modelo (1). Para o caso particular dun dato i -ésimo da mostra, teríase a predición $\hat{\beta}_0 + \hat{\beta}_1 x_i$, mentres que se observou Y_i , isto é, produciríase un erro de predición de $\hat{\varepsilon}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$, denominado **residuo** i -ésimo da regresión. A interpretación xeométrica destes residuos pode verse na Figura 1.

Estimaranse β_0 e β_1 mediante o **método de mínimos cadrados**, que consiste en escoller os estimadores $\hat{\beta}_0$ e $\hat{\beta}_1$ que minimicen a suma dos residuos ao cadrado. Así, $\hat{\beta}_0$ e $\hat{\beta}_1$ son tal que

$$\sum_{i=1}^n \hat{\varepsilon}_i^2 = \sum_{i=1}^n (Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 = \min_{\beta_0, \beta_1} \sum_{i=1}^n (Y_i - \beta_0 - \beta_1 x_i)^2.$$

Denótase a función a minimizar ou función obxectivo por

$$\phi(\beta_0, \beta_1) = \sum_{i=1}^n (Y_i - \beta_0 - \beta_1 x_i)^2.$$

Calculando ambas derivadas parciais, $\frac{\partial \phi}{\partial \beta_0}$ e $\frac{\partial \phi}{\partial \beta_1}$, e igualándoas a cero, obtéñense as denominadas **ecuacións normais**:

$$\begin{cases} -2 \sum_{i=1}^n (Y_i - \beta_0 - \beta_1 x_i) = 0, \\ -2 \sum_{i=1}^n (Y_i - \beta_0 - \beta_1 x_i) x_i = 0; \end{cases}$$

de onde se segue que os estimadores de mínimos cadrados son

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (Y_i - \bar{Y})(X_i - \bar{X})}{\sum_{i=1}^n (X_i - \bar{X})^2} = \frac{S_{XY}}{S_x^2} \text{ e } \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}.$$


Ademais, pódese consultar en [8] que as distribucións na mostraxe destes estimadores resultan ser

$$\hat{\beta}_1 \in N\left(\beta_1, \frac{\sigma^2}{n S_x^2}\right) \text{ e } \hat{\beta}_0 \in N\left(\beta_0, \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{n S_x^2}\right)\right),$$

onde \bar{x} e S_x^2 son a media e a varianza da mostra da variable explicativa X , respectivamente.

Ilustrarase o explicado no seguinte exemplo.

Exemplo 0.1. Raudenbush et al. [4] describen o estudo “High School and Beyond” de 1982 no que se consideran 7185 alumnos e alumnas cunha idade arredor dos 11 anos procedentes de 160 escolas estadounidenses, 70 das cales son de carácter relixioso, mentres que 90 pertencen á rede de escolas da Educación pública estadounidense.

Ao longo deste traballo considerarase un conxunto de datos que foi creado a partir da unión mediante a función `merge` de  dos conxuntos de datos `MathAchieve` e `MathAchSchool` incluídos no paquete `nlme` de Pinheiro et al. [24] e que non son máis que extraccións do estudo [4]. Na Táboa 1 amósase un estrato do conxunto de datos que se está a considerar, no cal se recollen as seguintes variables:

(i) **Variables relativas ao alumnado:**

- a) **Escola:** identificador da escola á que acode cada alumna/o, numeradas do 1 ao 160.
- b) **SocialMin:** variable dicotómica que toma os valores “Si” ou “Non” indicando se a/o estudante é membro dun grupo racial minoritario ou non, respectivamente.
- c) **Sexo:** variable categórica indicadora do sexo que toma os valores “Home” ou “Muller”.
- d) **StSE:** status socio-económico da familia á que pertence a/o alumna/o.
- e) **NotaMates:** nota acadada na materia de Matemáticas.

(ii) **Variables relativas ás escolas:**

- a) **MStSE:** media dos **StSE** das/os alumnas/os para cada centro educativo. Mide o status socio-económico promedio dos colexios.
- b) **Tamaño:** número de estudantes en cada escola.
- c) **Sector:** variable categórica indicando o carácter público ou privado da escola.
- d) **Particip:** proporción de estudantes da escola que participan no estudo académico.
- e) **AmbDiscrim:** medida do ambiente discriminatorio presente na escola.
- f) **MaioriaMin:** variable dicotómica que toma o valor 1 para as escolas con máis do 40 % das/os estudantes membros dun grupo racial minoritario, e 0 para o caso contrario.

Escola	SM	Sexo	StSE	NotaMates	MStSE	Tamaño	Sector	P	AD	MM
1	Non	Muller	-1.53	5.88	-0.43	842	Publica	0.35	1.6	0
1	Non	Muller	-0.59	19.71	-0.43	842	Publica	0.35	1.6	0
1	Non	Home	-0.53	20.35	-0.43	842	Publica	0.35	1.6	0

Táboa 1: Primeiras filas do conxunto de datos **mates** que se empregará ao longo de todo este traballo. Nótese que **SM** denota a **SocialMin**, **P** a **Particip**, **AD** a **AmbDiscrim** e **MM** a **MaioriaMin**.

Esta base de datos, que se denominará **mates**, foi deseñada coa finalidade de investigar o impacto de certas características relativas; tanto ao alumnado como á escola á que acode, na nota acadada en Matemáticas por cada estudante. Un dos obxectivos principais será determinar de que xeito a nota acadada por cada menor está relacionada co status socio-económico da súa familia. Para conseguilo, comezase tendo en conta unicamente estas dúas variables para axustar un modelo de regresión linear simple e irase aumentando a complexidade deste, asemade irá incrementándose tamén a efectividade e eficiencia dos modelos considerados.

Un exemplo de regresión simple móstrase na Figura 1, onde se representa o diagrama de dispersión da nota acadada en Matemáticas fronte ao status socio-económico das familias de 40 nenas e nenos escollidas/os arbitrariamente na base de datos orixinal.

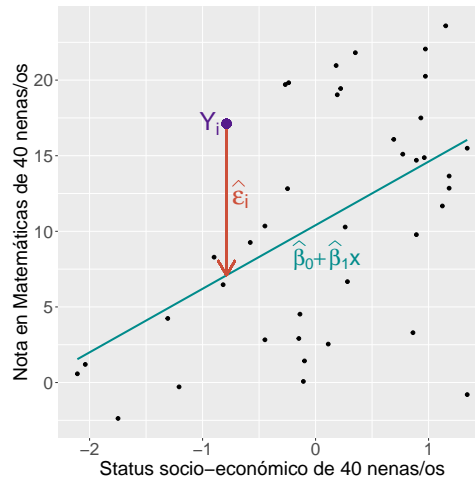



Figura 1: Diagrama de dispersión e recta axustada por mínimos cadrados (en azul) para a nota en Matemáticas de 40 nen/as/os de 11 anos escollidas/os aleatoriamente fronte ao status socio-económico destes. O segmento vertical vermello representa o residuo do dato i -ésimo (en violeta).

Pódese dar un paso máis e estender o modelo linear simple mediante a consideración de máis dunha variable explicativa continua que tamén sirva para explicar a variable resposta Y . Este modelo de regresión denomínase **modelo linear múltiple** e pode expresarse como

$$\mathbf{Y} = \mathbf{X}'\beta + \epsilon, \quad (2)$$

onde o erro ϵ segue a satisfacer as hipóteses de homocedasticidade, normalidade e independencia supostas no modelo linear simple. De novo por mínimos cadrados pódese estimar o vector de parámetros β , obtendo $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$, que segue unha distribución $\hat{\beta} \in N_p(\beta, \sigma^2(\mathbf{X}'\mathbf{X})^{-1})^3$, onde p é o número de variables explicativas a considerar.

Finalmente, o **modelo linear xeral** non é máis que a xeneralización de (2) e abrangue a consideración tanto de variables explicativas continuas como discretas. Pode atoparse máis información sobre estes modelos na obra de Faraway [8].

Ao longo deste traballo tratarase de estender este tipo de modelos a situacións nas cales a variable resposta Y esté medida en diferentes grupos, como no Exemplo 0.1, onde se fornecen os resultados en Matemáticas de alumnado de diferentes centros educativos e non é estraño pensar que as diferentes escolas teñan un “efecto” nas notas acadadas polas/os estudantes do centro. Ademais, todo o código de  empregado ao longo da totalidade do traballo atoparase dispoñible no repositorio *Modelos_Mixtos_con_R* de *GitHub* e será de carácter enteiramente reproducible. En gran parte atoparase tamén dispoñible no Anexo A.

³ $Z = (Z_1, \dots, Z_m) \in N_m(\mu, \sigma^2 I_m)$ presenta unha **distribución normal estándar multivariante** se cada unha das súas compoñentes Z_j ten distribución normal estándar univariante e son mutuamente independentes.

Capítulo 1

ANOVA e ANCOVA

Ao longo deste capítulo analizarase o **modelo de análise da varianza**, tamén coñecido como modelo *ANOVA* (acrónimo de *ANalysis Of VAriance*), que non é máis que un modelo de regresión no cal hai unha única variable explicativa, que ademais é discreta ou categórica. Logo incluíranse variables explicativas discretas e continuas asemade para construír o **modelo de análise da covarianza** ou modelo *ANCOVA* (segundo terminoloxía inglesa, *ANalysis of COVAriance*).

1.1. O modelo *ANOVA*

O modelo de análise da varianza ou *ANOVA* é semellante ao modelo de regresión linear (2), pero difire deste dado que considera unha variable explicativa discreta ou categórica.

1.1.1. *ANOVA* como modelo linear xeral

Supóñase que se ten unha variable continua Y medida en J grupos, isto é, J mostras independentes onde cada mostra está formada por variables independentes entre si e con idéntica distribución $N(\mu_j, \sigma^2)$. Denótase por μ_j á media da mostra da variable resposta relativa ao grupo j -ésimo, para cada $j \in \{1, \dots, J\}$. Ademais, o índice i será empregado para denotar o valor da variable resposta Y na i -ésima observación para o grupo j . Así disporase dunha mostra $\{Y_{ij}\}$ con $i = 1, \dots, n_j$ para cada grupo $j = 1, \dots, J$. Denótase tamén por $n = \sum_{j=1}^J n_j$ ao número de observacións totais da mostra. Nótese que por supoñerse todas as varianzas σ^2 iguais, o modelo de análise da varianza é homocedástico por definición.

O modelo *ANOVA* pode escribirse entón como un modelo de regresión do seguinte xeito:

$$Y_{ij} = \mu_j + \varepsilon_{ij}, \text{ con } i = 1, \dots, n_j \text{ e } j = 1, \dots, J; \quad (1.1)$$

onde os $\varepsilon_{ij} \in N(0, \sigma^2)$ son independentes para cada $j = 1, \dots, J$ e para cada $i = 1, \dots, n_j$.

Como xa se intúe en (1.1), o modelo *ANOVA* é un modelo linear xeral xa que pode ser expresado en notación matricial da forma $\mathbf{Y} = \mathbf{X}\beta + \boldsymbol{\epsilon}$, mediante as seguintes expresións:

$$\mathbf{Y} = \begin{pmatrix} Y_{11} \\ \vdots \\ Y_{n_1 1} \\ \vdots \\ Y_{1J} \\ \vdots \\ Y_{n_J J} \end{pmatrix}_{n \times 1}, \quad \mathbf{X} = \begin{pmatrix} 1 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{pmatrix}_{n \times J}, \quad \beta = \boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_J \end{pmatrix}_{J \times 1}, \quad \boldsymbol{\epsilon} = \begin{pmatrix} \varepsilon_{11} \\ \vdots \\ \varepsilon_{n_1 1} \\ \vdots \\ \vdots \\ \varepsilon_{1J} \\ \vdots \\ \varepsilon_{n_J J} \end{pmatrix}_{n \times 1}; \quad (1.2)$$

de onde se deduce que $\mathbb{E}(\mathbf{Y}) = (\mu_1, \binom{n_1}{\dots}, \mu_1, \mu_2, \binom{n_2}{\dots}, \mu_2, \dots, \mu_J, \binom{n_J}{\dots}, \mu_J)$ e que a matriz de covarianzas do modelo é $\text{Var}(\mathbf{Y}) = \text{Var}(\mathbf{X}\beta + \boldsymbol{\epsilon}) = \text{Var}(\boldsymbol{\epsilon}) = \sigma^2 I_n$, con I_n denotando a matriz identidade de dimensión $n \times n$.

1.1.2. Estimación dos parámetros

Os parámetros a estimar son as medias de cada grupo $j = 1, \dots, J$; é dicir, as compoñentes do vector $\boldsymbol{\mu}$ definido en (1.2).

Notación 1.1. Denotarase por $Y_{\bullet j} = \sum_{i=1}^{n_j} Y_{ij}$ á suma dos valores da variable resposta para o grupo j . Esta notación substitúe un índice efectuando a adición de todas as observacións obtidas ao recorrer ese índice e fixando os restantes. Empregarase esta notación ao longo do traballo. Nótese que, deste mesmo xeito, $\bar{Y}_{\bullet j} = \frac{1}{n_j} \sum_{i=1}^{n_j} Y_{ij} = Y_{\bullet j}/n_j$ denotaríase a media da mostra dos datos do grupo j , $Y_{\bullet\bullet} = \sum_{j=1}^J \sum_{i=1}^{n_j} Y_{ij}$ denotaríase a suma de todos os valores da variable resposta e $\bar{Y}_{\bullet\bullet} = \sum_{j=1}^J \frac{n_j}{n} \bar{Y}_{\bullet j}$ sería a súa media, así pois, a media na mostraxe de todos os datos.

Ao igual que para calquera modelo de regresión linear xeral, nos modelos de regresión con erro normal, os métodos de mínimos cadrados ou de máxima verosimilitude proporcionan o mesmo estimador para cada μ_j , daquela é indiferente que método se empregue para estimar os coeficientes do modelo *ANOVA*.

Conforme ao criterio de mínimos cadrados, a expresión a minimizar é:

$$Q = \sum_{j=1}^J \sum_{i=1}^{n_j} (Y_{ij} - \mu_j)^2 = \sum_{i=1}^{n_1} (Y_{i1} - \mu_1)^2 + \sum_{i=1}^{n_2} (Y_{i2} - \mu_2)^2 + \cdots + \sum_{i=1}^{n_J} (Y_{iJ} - \mu_J)^2.$$

Como cada parámetro só aparece nun sumatorio, para minimizar Q poderíase pensar en minimizar cada sumatorio por separado. Así, denotando $Q_j = \sum_{i=1}^{n_j} (Y_{ij} - \mu_j)^2$;

$$\frac{dQ_j}{d\mu_j} = -2 \sum_{i=1}^{n_j} (Y_{ij} - \mu_j) = 0 \Rightarrow \sum_{i=1}^{n_j} Y_{ij} = n_j \mu_j \Rightarrow \hat{\mu}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} Y_{ij} = \bar{Y}_{\bullet j}.$$



Deste xeito, o valor estimado para a media poboacional de cada grupo resultará ser a media da mostra de valores Y_{ij} asociados ao grupo j , como se podía intuír.

Para ver que efectivamente os métodos de máxima verosimilitude e de mínimos cadrados proporcionan os mesmos estimadores para todos os parámetros μ_j , con $j = 1, \dots, J$; basta con percatarse de que, dado que $Y_{ij} \in N(\mu_j, \sigma^2)$, a función de verosimilitude resulta ser

$$L(\mu_1, \dots, \mu_J; \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} \cdot \exp \left[-\frac{1}{2} \sum_{j=1}^J \sum_{i=1}^{n_j} \left(\frac{Y_{ij} - \mu_j}{\sigma} \right)^2 \right],$$

e efectivamente, maximizar L con respecto aos parámetros μ_j é equivalente a minimizar Q ou $\exp(Q)$.

Deste xeito, os residuos do modelo ANOVA, que seguen a ser a diferenza entre os valores observados e os axustados, resultan ser $\hat{\varepsilon}_{ij} = Y_{ij} - \hat{Y}_{ij} = Y_{ij} - \bar{Y}_{\bullet j}$. Así, os residuos representan a desviación dunha observación con respecto á media estimada do seu grupo. Unha propiedade importante é que os residuos $\hat{\varepsilon}_{ij}$ suman cero para cada grupo j , posto que $\sum_{i=1}^{n_j} \hat{\varepsilon}_{ij} = \sum_{i=1}^{n_j} (Y_{ij} - \bar{Y}_{\bullet j}) = \sum_{i=1}^{n_j} Y_{ij} - n_j \bar{Y}_{\bullet j} = 0$ para cada $j = 1, \dots, J$.

Exemplo 1.2. Para ilustrar o modelo ANOVA, seguirase empregando a base de datos presentada no Exemplo 0.1. Construírase un modelo en  que explique a nota acadada en Matemáticas (variable resposta) en función dunha única variable explicativa, que ademais sexa discreta. Unha situación interesante sería considerar unha variable auxiliar que permita determinar a pertenza ou non do alumnado considerado ás únicas 7 escolas cuxo 100 % do alumnado participou no estudo presentado en [4], coa finalidade de evitar nesgos de selección. Tales escolas denotaríanse por E1, E2, E3, E4, E5, E6 e E7. A base de datos creada mediante a consideración de unicamente estas sete escolas denominarase `mates7` e será a que se empregue de aquí en diante. A continuación amósase o código elaborado para definir o modelo, xunto co resumo do mesmo obtido coa función `summary` de .

```

anova_mates7 <- lm(NotaMates ~ Escola - 1)
summary(anova_mates7)

##
## Call:
## lm(formula = NotaMates ~ Escola - 1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -17.6957  -3.2007   0.7324   3.6733  10.1293
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## EscolaE1    18.1116     0.7533   24.04  <2e-16 ***
## EscolaE2    16.9639     1.0904   15.56  <2e-16 ***
## EscolaE3    19.7156     0.7138   27.62  <2e-16 ***
## EscolaE4    12.3102     0.8570   14.37  <2e-16 ***
## EscolaE5    18.4557     0.6619   27.89  <2e-16 ***
## EscolaE6    16.2323     0.7620   21.30  <2e-16 ***
## EscolaE7    14.8637     0.6505   22.85  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.997 on 300 degrees of freedom
## Multiple R-squared:  0.9219, Adjusted R-squared:  0.9201
## F-statistic: 506.1 on 7 and 300 DF,  p-value: < 2.2e-16

```

No apartado concernente á estimación dos coeficientes proporciónase, xunto a cada escola, a estimación $\hat{\mu}_j$ da media das notas de Matemáticas acadadas nese mesmo centro educativo “Ej” para un j entre 1 e 7; que como se viu, non é máis que a media da mostra das notas do correspondente colexio, $\bar{Y}_{\bullet j}$. Por outra banda, o nivel crítico asociado a cada unha destas estimacións non ten ningunha utilidade; pois é o p-valor asociado ao contraste de que tales estimacións son nulas, o que non ten sentido tratándose de variables enteiramente positivas (salvo datos illados).

Na Figura 1.1 móstrase o comportamento da nota acadada en Matemáticas para as diferentes escolas consideradas a través de diagramas de caixas. Ademais, engádese a media da variable resposta en cada grupo (liña punteada dentro do diagrama de caixa) e a media global (liña

descontinua azul). Unha primeira ollada amosa que as medias (liña punteada presente en cada caixa) concernentes á nota en Matemáticas quizáis difiran entre as diferentes escolas máis do que deberían se a causa fose soamente o azar.

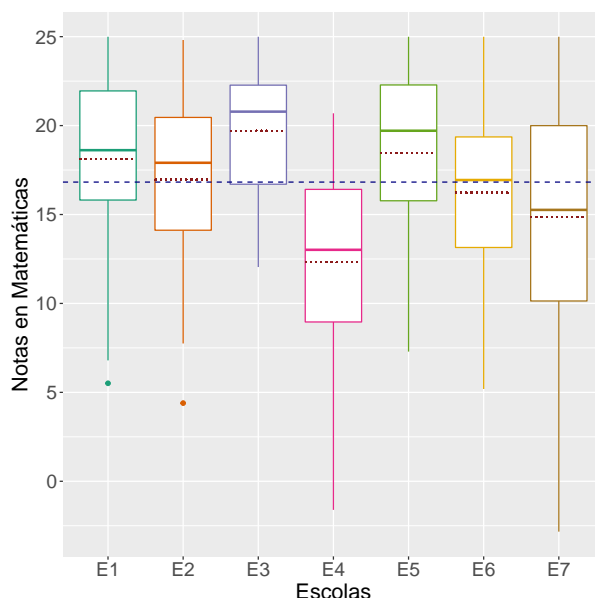


Figura 1.1: Diagramas de caixas da nota de Matemáticas nas 7 escolas con participación absoluta. A liña horizontal descontinua representa a media da mostra global, mentres que en cada caixa a raia punteada é a media da mostra de cada escola.

1.1.3. Análise da varianza e test F

Dado o modelo (1.1) con medias grupais μ_j onde $j = 1, \dots, J$; poderíase pensar se estas medias son todas iguais ou hai algunha que difire do resto. O caso particular de dous grupos, $J = 2$, non é máis ca un contraste de comparación de medias en poboacións normais, é dicir, o ben coñecido *t-test*¹. Estudaranse máis detalladamente os casos con $J \geq 3$. Para contrastar medias de máis de dous grupos de datos, poderíase pensar en facer os contrastes por pares, pero esta estratexia pode ser perigosa. Se temos varios grupos de datos e facemos varias comparacións; é probable que, co tempo, atopemos unha diferenza só por azar. En lugar disto, debería aplicarse un test integral, e é aquí onde xorde o denominado **test F**.

¹Dada unha variable $X \in N(\mu_X, \sigma_X^2)$ é coñecido que $\frac{\bar{X} - \mu_X}{S_{cX}/\sqrt{n}} \in T_{n-1}$, onde S_{cX} é a cuasivarianza da mostra X_1, \dots, X_n , e tal pivote permitirá realizar intervalos de confianza e contrastes de hipóteses para a media poboacional μ_X . Así, considerando a variable $X - Y \in N(\mu_X - \mu_Y, \sigma^2)$, o pivote convértese en $\frac{\bar{X} - \bar{Y}}{S_c/\sqrt{n}} \in T_{n-1}$; onde S_c^2 é a cuasivarianza da mostra $X_1 - Y_1, \dots, X_n - Y_n$, e permite, en particular, contrastar a diferenza das medias poboacionais. Este tipo de contrastes coñécense como *t-test* posto que a distribución do pivote é unha *t de Student*.

O *ANOVA* emprega un test de hipóteses simples para comprobar se as medias de varios grupos son iguais; isto é, se $\mu_1 = \dots = \mu_J$, ou se algunha das medias é distinta. O contraste a estudar é o seguinte:

$$\begin{cases} H_0 : \mu_1 = \dots = \mu_J, \\ H_a : \text{polo menos unha das medias difire do resto.} \end{cases} \quad (1.3)$$

Como cada ε_{ij} ten distribución normal, tamén $Y_{ij} \in N(\mu_j, \sigma^2)$; dado que Y_{ij} é unha función linear de ε_{ij} . Ademais, a independencia dos ε_{ij} implica a independencia dos Y_{ij} .

Nótese que, á vista das hipóteses do modelo *ANOVA* e baixo a hipótese nula H_0 , estaríase a dicir que a distribución da variable resposta é a mesma nos diferentes grupos considerados. Ademais, a variabilidade total das observacións Y_{ij} sen ter en conta os distintos grupos (baixo H_0) vén dada polas desviacións de cada valor observado respecto da media global $Y_{ij} - \bar{Y}_{\bullet\bullet}$. Pola contra, se temos en conta os grupos (baixo a hipótese alternativa H_a), a variabilidade debida ao erro vén dada polas desviacións de cada valor observado respecto da media estimada do seu respectivo grupo: $Y_{ij} - \bar{Y}_{\bullet j}$ ($= \hat{\varepsilon}_{ij}$). A diferenza entre ambas expresións coincide coa diferenza entre a media estimada de cada grupo e a media global, isto é,

$$(Y_{ij} - \bar{Y}_{\bullet\bullet}) - (Y_{ij} - \bar{Y}_{\bullet j}) = \bar{Y}_{\bullet j} - \bar{Y}_{\bullet\bullet}.$$

Así, pódese descompoñer a variabilidade total da variable resposta en dúas compoñentes:

$$Y_{ij} - \bar{Y}_{\bullet\bullet} = (\bar{Y}_{\bullet j} - \bar{Y}_{\bullet\bullet}) + (Y_{ij} - \bar{Y}_{\bullet j}),$$

e en consecuencia, efectuando o cadrado da expresión anterior teríase que

$$\begin{aligned} \sum_{j=1}^J \sum_{i=1}^{n_j} (Y_{ij} - \bar{Y}_{\bullet\bullet})^2 &= \sum_{j=1}^J \sum_{i=1}^{n_j} (\bar{Y}_{\bullet j} - \bar{Y}_{\bullet\bullet})^2 + \sum_{j=1}^J \sum_{i=1}^{n_j} (Y_{ij} - \bar{Y}_{\bullet j})^2 + \\ &+ \sum_{j=1}^J \sum_{i=1}^{n_j} 2(\bar{Y}_{\bullet j} - \bar{Y}_{\bullet\bullet})(Y_{ij} - \bar{Y}_{\bullet j}). \end{aligned}$$

Agora ben,

$$\sum_{j=1}^J \sum_{i=1}^{n_j} 2(\bar{Y}_{\bullet j} - \bar{Y}_{\bullet\bullet})(Y_{ij} - \bar{Y}_{\bullet j}) = 2 \sum_{j=1}^J \left((\bar{Y}_{\bullet j} - \bar{Y}_{\bullet\bullet}) \cdot \sum_{i=1}^{n_j} (Y_{ij} - \bar{Y}_{\bullet j}) \right) = 0;$$

xa que $\sum_{i=1}^{n_j} (Y_{ij} - \bar{Y}_{\bullet j}) = \sum_{i=1}^{n_j} Y_{ij} - n_j \bar{Y}_{\bullet j} = Y_{\bullet j} - Y_{\bullet j} = 0$, e entón dedúcese que

$$\underbrace{\sum_{j=1}^J \sum_{i=1}^{n_j} (Y_{ij} - \bar{Y}_{\bullet\bullet})^2}_{\text{variabilidade total}} = \underbrace{\sum_{j=1}^J n_j (\bar{Y}_{\bullet j} - \bar{Y}_{\bullet\bullet})^2}_{\text{variabilidade entre grupos}} + \underbrace{\sum_{j=1}^J \sum_{i=1}^{n_j} (Y_{ij} - \bar{Y}_{\bullet j})^2}_{\text{variabilidade dentro dos grupos}}. \quad (1.4)$$

O termo da esquerda serve como medida da variabilidade total das observacións Y_{ij} e denótase por VT . O primeiro termo na dereita de (1.4) denótase por VEG (variabilidade entre grupos), e o segundo por VDG (variabilidade dentro dos grupos). Deste xeito, (1.4) pode escribirse de forma equivalente como $VT = VEG + VDG$.

Correspondente a esta descomposición da suma total de cadrados, tamén podemos obter unha descomposición dos graos de liberdade asociados a cada sumando:

VT ten $n - 1$ graos de liberdade asociados. Hai en total n observacións $Y_{ij} - \bar{Y}_{\bullet\bullet}$, pero un grao de liberdade pérdese porque as desviacións non son independentes no sentido de que suman cero: $\sum_{j=1}^J \sum_{i=1}^{n_j} (Y_{ij} - \bar{Y}_{\bullet\bullet}) = Y_{\bullet\bullet} - Jn_j \bar{Y}_{\bullet\bullet} = Y_{\bullet\bullet} - Y_{\bullet\bullet} = 0$.


VEG ten $J - 1$ graos de liberdade asociados. Hai J desviacións da media estimada $\bar{Y}_{\bullet j} - \bar{Y}_{\bullet\bullet}$ de cada grupo j , pero un grao de liberdade pérdese igual que na VT , pois $\sum_{j=1}^J n_j (\bar{Y}_{\bullet j} - \bar{Y}_{\bullet\bullet}) = \sum_{j=1}^J n_j \bar{Y}_{\bullet j} - \bar{Y}_{\bullet\bullet} \sum_{j=1}^J n_j = \sum_{j=1}^J Y_{\bullet j} - n \bar{Y}_{\bullet\bullet} = Y_{\bullet\bullet} - Y_{\bullet\bullet} = 0$

VDG ten $n - J$ graos de liberdade asociados. Basta considerar a compoñente de VDG para cada grupo j : $\sum_{i=1}^{n_j} (Y_{ij} - \bar{Y}_{\bullet j})^2$, que é o equivalente á suma total de cadrados considerando só o grupo j , polo que ten $n_j - 1$ graos de liberdade asociados. Así, os graos de liberdade asociados a VDG son $(n_1 - 1) + (n_2 - 1) + \dots + (n_J - 1) = n - J$.

Nótese que

$$\sum_{j=1}^J \sum_{i=1}^{n_j} (Y_{ij} - \bar{Y}_{\bullet j})^2 = \sum_{j=1}^J (n_j - 1) \frac{(Y_{ij} - \bar{Y}_{\bullet j})^2}{n_j - 1} = \sum_{j=1}^J (n_j - 1) S_j^2;$$

onde S_j^2 denota a cuasivarianza de Y no grupo j .

Recompílese esta información na Táboa 1.1, que posteriormente verase que é un bosquejo da verdadeira saída de  ao realizar o contraste (1.3).

	Graos liberdade	Sumas de cadrados
Grupo	$J - 1$	$VEG = \sum_{j=1}^J n_j (\bar{Y}_{\bullet j} - \bar{Y}_{\bullet\bullet})^2$
Residuos	$n - J$	$VDG = \sum_{j=1}^J (n_j - 1) S_j^2$
Total	$n - 1$	$VT = \sum_{j=1}^J \sum_{i=1}^{n_j} (Y_{ij} - \bar{Y}_{\bullet\bullet})^2$

Táboa 1.1: Táboa de descomposición da variabilidade asociada a un modelo ANOVA.

A análise da varianza ANOVA céntrase en comparar as medias de cada grupo a través da análise da varianza entre grupos e dentro de cada grupo. Isto é, ANOVA considera simultaneamente moitos grupos e evalúa se as súas medias na mostraxe difiren máis do que se esperaría

da variación natural. Esta variabilidade é a media cadrática entre grupos e que se denotará por *MEG*. Por outra banda, para conseguir un valor de referencia sobre tanta variabilidade debe esperarse entre as medias da mostra emprégase unha estimación da varianza dentro dos grupos, o que se chamará erro cadrático medio ou media cadrática dentro dos grupos *MDG*.

As medias cadráticas obtéñense dividindo cada suma de cadrados polos seus graos de liberdade asociados:

$$MEG = \frac{VEG}{J-1} \quad \text{e} \quad MDG = \frac{VDG}{n-J}.$$

Para efectuar o contraste de igualdade entre as medias de todos os grupos simultaneamente, empregamos o **test F**, cuxo estatístico *F* adoptará a forma

$$F = \frac{MEG}{MDG} = \frac{VEG/(J-1)}{VDG/(n-J)},$$

que non é máis que o cociente entre a variabilidade entre grupos e a variabilidade dentro dos grupos.

Valores grandes de *F* aportan indicios a favor de H_a , xa que *MEG* tenderá a exceder *MDG* cando H_a é certa, xa que a variabilidade entre grupos é maior que a variabilidade dentro de cada grupo. É dicir, teríanse os grupos "separados", o cal é un indicativo de que non todas as medias son iguais. Valores pequenos de *F* aportan indicios a favor de H_0 , xa que *MEG* e *MDG* teñen o mesmo valor esperado baixo H_0 (ver [20, pp. 538–542]), isto é, baixo a hipótese nula de que todas as medias dos grupos son iguais calquera diferenza entre as medias da mostra débese unicamente ao azar. Para poder construír unha regra de decisión e a rexión crítica desta, necesítase coñecer a distribución do estatístico *F*.

Baixo H_0 , tense que todas as medias grupais μ_j son iguais e polo tanto todas as respostas Y_{ij} teñen a mesma distribución, e como consecuencia da hipótese de normalidade do modelo *ANOVA*, aplicando o Teorema de Cochran (pódese consultar en [20, pp. 92–93]) teríase que:

$$\text{Baixo } H_0, \quad \frac{VEG}{\sigma^2} \in \chi_{J-1}^2, \quad \frac{VDG}{\sigma^2} \in \chi_{n-J}^2; \quad \text{e son independentes}^2.$$

Polo tanto,

$$F = \frac{MEG}{MDG} = \frac{VEG/(J-1)}{VDG/(n-J)} = \frac{\frac{VEG/\sigma^2}{J-1}}{\frac{VDG/\sigma^2}{n-J}} \in F_{J-1, n-J} \quad (\text{Baixo } H_0).$$

Isto é, se H_0 é certa e as condicións do modelo se verifican, entón o estatístico *F* segue unha distribución **F de Snedecor**³ con graos de liberdade $J-1$ e $n-J$. Ademais, a rexión crítica


²Se Z_1, \dots, Z_m son variables aleatorias normais estándar e independentes; entón $X = Z_1^2 + \dots + Z_m^2 \in \chi_m^2$ segue unha distribución chi-cadrado con m graos de liberdade. É unha distribución non negativa, con media m e varianza $2m$.

³Se $X_1 \in \chi_{m_1}^2, X_2 \in \chi_{m_2}^2$ e son independentes, entón $F = \frac{X_1/m_1}{X_2/m_2} \in F_{m_1, m_2}$ e dise que ten distribución *F* de Snedecor con m_1 e m_2 graos de liberdade.


do test para un nivel de significación α será:


$$\text{Rexéitase } H_0 : \mu_1 = \dots = \mu_J \text{ se } F > f_{1-\alpha; J-1, n-J}, \quad (1.5)$$

onde $f_{1-\alpha; J-1, n-J}$ representa o cuantil de orde $1 - \alpha$ da distribución F de Snedecor con $J - 1$ e $n - J$ graos de liberdade.

Finalmente, para ver como se efectúan os cálculos do modelo de análise da varianza pódese seguir a organización da táboa ANOVA de , que aparece reflexada na Táboa 1.2.

	Graos liberdade	Sumas de cadrados	Medias cadráticas	F
Grupo	$J - 1$	$VEG = \sum_{j=1}^J n_j (\bar{Y}_{\bullet j} - \bar{Y}_{\bullet\bullet})^2$	$MEG = \frac{VEG}{J-1}$	$\frac{MEG}{MDG}$
Residuos	$n - J$	$VDG = \sum_{j=1}^J (n_j - 1) S_j^2$	$MDG = \frac{VDG}{n-J}$	
Total	$n - 1$	$VT = \sum_{j=1}^J \sum_{i=1}^{n_j} (Y_{ij} - \bar{Y}_{\bullet\bullet})^2$		

Táboa 1.2: Táboa de descomposición de variabilidade asociada a un modelo ANOVA obtida grazas ao entorno estatístico .

Exemplo 1.3. Aplicando a función `anova` de  ao modelo presentado no Exemplo 1.2, onde recórdese que a variable resposta é a nota acadada en Matemáticas e está medida en 7 escolas diferentes, pódese obter a táboa de descomposición da varianza que se amosa na Táboa 1.2.

```
anova(lm(NotaMates ~ Escola))

## Analysis of Variance Table
##
## Response: NotaMates
##           Df Sum Sq Mean Sq F value    Pr(>F)
## Escola      6 1569.3  261.558   10.475 1.49e-10 ***
## Residuals  300 7490.6   24.969
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Obsérvase que o valor do estatístico asociado ao test F é de 10.475. A partir del é sinxelo calcular o nivel crítico empregando (1.5), que resulta ser 1.49×10^{-10} . Agora ben, xa que o nivel crítico ou p-valor é menor que os niveis de significación habituais e, en particular, moito menor que $\alpha = 0.001$; que será o nivel de significación que se empregue ao longo de todo o traballo, séguese que existen evidencias estatisticamente significativas a favor da hipótese alternativa de que polo menos algunha media é diferente das demais, tal e como se intuía na Figura 1.1.

Agora, para estudar que escolas en particular teñen diferentes medias nas notas acadadas sería preciso efectuar un contraste de algún tipo a cada par de escolas; xorden así as **comparacións múltiples**.

1.1.4. Comparacións múltiples


Cando se rexeita a H_0 asociada a un test F , resulta interesante coñecer cales dos grupos que se están a estudar teñen diferentes medias. Para iso, poderíanse comparar as medias de cada par de grupos mediante un *t-test*; pero cando executamos tantos contrastes de comparacións de medias en poboacións normais o erro de tipo I (nivel de significación), que recórdese vén dado por $\alpha = \mathbb{P}(\text{erro tipo I}) = \mathbb{P}(\text{rexear } H_0 \mid H_0 \text{ é certa})$ incrementase, debido a que se os tests son independentes verifícase que

$$\begin{aligned} \mathbb{P}(\text{rexear algún test} \mid H_0 \text{ é certa}) &= 1 - \mathbb{P}(\text{non rexear ningún test} \mid H_0 \text{ é certa}) = \\ &= 1 - [1 - \mathbb{P}(\text{rexear } H_0 \mid H_0 \text{ é certa})]^K = \\ &= 1 - (1 - \alpha)^K, \end{aligned}$$

obtendo así unha función exponencialmente crecente en K , onde K é o número de comparacións consideradas. Así, se hai J grupos e todos son comparados dous a dous, entón $K = \frac{J(J-1)}{2}$. O problema resólvese empregando “un nivel de significación modificado”, isto é, por exemplo, a denominada **corrección de Bonferroni** para un nivel de significación α_s . Este método propón que para realizar contrastes múltiples sería máis axeitado empregar un nivel de significación máis estricto,

$$\alpha^* = \frac{\alpha_s}{K}.$$

Unha mellora deste método sería o **Bonferroni secuencial** ou **método de Holm** (pode verse en [17]), aínda que tamén existen outros procedementos para realizar estes contrastes pareados, como os **métodos Tukey** ou **Scheffé**; os cales poden consultarse en [20, pp. 574–580].

Exemplo 1.4. No caso particular das sete escolas estadounidenses, o número de contrastes a realizar sería de $K = 21$. Como se descoñece unha función directa de  que calcule os intervalos de confianza simultáneos polo método de Bonferroni, prográmase unha co nome de `Bonferroni_taboa`, cuxo código pode atoparse no Anexo A.4.2, que ademais indique directamente cales son as escolas con promedios diferentes nas notas de Matemáticas das/os súas/seus estudantes, tras verificar cales son os intervalos simultáneos que non conteñen o cero. Executando esta para a base de datos coa que se está a traballar e tendo en conta o nivel de significación de $\alpha = 0.001$ previamente considerado, obtense o seguinte:


```

Bonferroni_taboa(NotaMates, Escola, alpha_sen_CB = .001, ndixitos = 2)

## ##### Intervalos de confianza simultáneos
## # con corrección de Bonferroni: #####
##      dif      inf      sup
## 1-2 -1.15  -6.62  4.32
## 1-3  1.60  -2.68  5.89
## 1-4 -5.80 -10.51 -1.09
## 1-5  0.34  -3.79  4.48
## 1-6 -1.88  -6.30  2.54
## 1-7 -3.25  -7.36  0.86
## + .....
##
## ##### Grupos cuxas medias difiren: #####
## - Os grupos 1-4 difiren nas súas medias
## - Os grupos 3-4 difiren nas súas medias
## - Os grupos 3-7 difiren nas súas medias
## - Os grupos 4-5 difiren nas súas medias
##
## ## Nivel de significación empregado
##      coa corrección de Bonferroni:  0.001 / 21 = 2.38e-05

```

Á vista dos resultados anteriores, en concordancia co mostrado na Figura 1.1, os pares de escolas E1-E4, E3-E4, E3-E7 e E4-E5 teñen medias significativamente diferentes. Isto ten completo sentido, xa que a escola con peor nota media da mostra (E4) resulta ter unha nota media poboacional significativamente distinta ás das escolas E1, E3 e E5; que precisamente son as que teñen mellor media da mostra. Ademais, o colexio coa mellor media na mostraxe de todos (E3) tamén difire significativamente da escola coa segunda peor media da mostra, E7.

Nótese que, sería posible rexeitar a hipótese nula de (1.3) empregando a función `anova` de  e logo non identificar diferenzas nas comparacións pareadas. Obviamente isto non invalida a conclusión da análise ANOVA, soamente significa que non fomos capaces de identificar que grupos específicos difiren nas súas medias.

1.2. O modelo *ANCOVA*

A idea principal do modelo de análise da covarianza coñecido habitualmente como *ANCOVA* é aumentar o modelo de análise da varianza, que contén os efectos fixos dos grupos, engadindo unha ou máis variables continuas que estén relacionadas coa variable resposta Y . Nas seguintes liñas, consideraranse modelos de análise da covarianza que inclúan unha variable explicativa discreta e outra continua, o cal será suficiente para entender as cuestións que surxen ao incluír variables de ambos tipos. Por conseguinte, obterase unha recta de regresión distinta para cada grupo. Ademais, a variable explicativa pode influír de dous xeitos distintos na resposta; segundo a pendente das rectas varíe ou non entre os diferentes grupos falarase de modelos con interacción ou sen interacción, respectivamente.

Cada variable explicativa engadida ao estudo denomínase **variable concomitante**⁴. Claramente, a elección destas variables é importante, pois se non teñen relación coa variable resposta o modelo non aporta nada novo en relación co que achegaría a aplicación dun *ANOVA*. Ademais, para interpretacións máis claras dos resultados, unha variable concomitante debe observarse antes do estudo, e se se observa durante este; non debería estar influenciada polos grupos de ningún xeito (véxase [20, pp. 845–847]).

1.2.1. *ANCOVA* sen interacción

Seguindo coa notación introducida para o modelo *ANOVA*, denótase o valor da variable continua asociado coa i -ésima observación para o grupo j por X_{ij} . O modelo de análise da covarianza comeza co modelo (1.1) e simplemente engade outro termo (ou varios), reflectindo a relación entre a variable concomitante e a variable resposta. Como primeira aproximación emprégase unha relación linear,

$$Y_{ij} = \mu + \tau_j + \gamma X_{ij} + \varepsilon_{ij}, \text{ con } i = 1, \dots, n_j \text{ e } j = 1, \dots, J. \quad (1.6)$$

Aquí μ é a constante dada pola media global $\mu = \frac{1}{n} \sum_{j=1}^J n_j \mu_j$, τ_j é o efecto (fixo) do grupo j e γ é o coeficiente de regresión que explica a relación entre Y e X ; que non varía entre os grupos e por ese motivo se di que non hai interacción entre X e estes. Suporase tamén que $\varepsilon_{ij} \in N(0, \sigma^2)$ e que son independentes, para todo $j = 1, \dots, J$ e $i = 1, \dots, n_j$. Deste xeito, fixado o grupo, o modelo (1.6) non é máis ca unha recta de regresión con incercepto $(\mu + \tau_j)$ e pendente γ ; de modo que as rectas de regresión para os distintos grupos son rectas paralelas cuxo intercepto se desvía da media global nunha cantidade τ_j para cada grupo j . A ilustración da esquerda na Figura 1.2 concernente ás 7 escolas da base de datos *mates7* alude á natureza deste modelo.

⁴A denominación de variable concomitante é bastante frecuente no ámbito da Medicina. Outra opción sería empregar o termo covariable.

1.2.2. Estimación dos parámetros

Os parámetros a estimar son as desviacións τ_j de cada grupo j á media global, esta media global μ e mais a pendente común a todos os grupos, γ . Primeiramente, estímase a pendente empregando regresión particionada. Así, $\hat{\gamma}$ é a pendente da regresión simple resultante de considerar como variable resposta os residuos de Y sobre a variable explicativa discreta, e como variable explicativa os residuos da propia X sobre a mesma variable explicativa discreta; resultando

$$\hat{\gamma} = \frac{\sum_{j=1}^J \sum_{i=1}^{n_j} (Y_{ij} - \bar{Y}_{\bullet j})(X_{ij} - \bar{X}_{\bullet j})}{\sum_{j=1}^J \sum_{i=1}^{n_j} (X_{ij} - \bar{X}_{\bullet j})}.$$

A partir deste estimador, obter o resto xa é máis sinxelo, pois basta considerar $\hat{\mu} = \bar{Y}_{\bullet\bullet} - \hat{\gamma}\bar{X}_{\bullet\bullet}$ e $\hat{\tau}_j = \bar{Y}_{\bullet j} - \bar{Y}_{\bullet\bullet} - \hat{\gamma}(\bar{X}_{\bullet j} - \bar{X}_{\bullet\bullet})$.

A suma residual de cadrados deste modelo é

$$RSS = \sum_{j=1}^J \sum_{i=1}^{n_j} \hat{\varepsilon}_{ij}^2 = \sum_{j=1}^J \sum_{i=1}^{n_j} (Y_{ij} - \hat{\mu} - \hat{\tau}_j - \hat{\gamma}X_{ij})^2,$$


e entón a varianza do erro estímase por $\hat{\sigma}^2 = \frac{RSS}{n-J-1}$.

1.2.3. *ANCOVA* con interacción

Ao engadir interacción ao modelo (1.6) as rectas de regresión de Y sobre X non só mudan o intercepto, senón que a súa pendente tamén varía entre os diferentes grupos, en contraposición co modelo sen a interacción; no cal a pendente era a mesma en todas as rectas. Tal modelo pode escribirse como

$$Y_{ij} = \mu_j + \gamma_j X_{ij} + \varepsilon_{ij} = \mu + \tau_j + \gamma X_{ij} + \delta_j X_{ij} + \varepsilon_{ij}, \quad (1.7)$$

onde $j = 1, \dots, J$ e $i = 1, \dots, n_j$. Deste xeito, para construír o modelo *ANCOVA* con interacción bastaría con estimar J rectas de regresión linear simple totalmente independentes unhas doutras, xa que, fixado o grupo j , (1.7) é precisamente unha liña de regresión con intercepto $\mu_j = \mu + \tau_j$ e pendente $\gamma_j = \gamma + \delta_j$; tal e como se mostra na ilustración da dereita na Figura 1.2, onde se estiman 7 rectas independentes, unha para cada escola do conxunto de datos `mates7`.

Exemplo 1.5. Mediante a función `lm` de  axústanse dous modelos *ANCOVA*, un sen interacción e outro con interacción; para tratar de explicar a nota acadada na materia de Matemáticas polas/os alumnas/os das sete escolas con participación unánime fronte ao status socio-económico das familias destas/es. Na Figura 1.2 atópanse as gráficas asociadas a ámbolos dous modelos, e nas dúas se axustou unha recta diferente para cada escola. Na gráfica da esquerda as rectas varían na súa ordenada na orixe, mais todas as rectas teñen a mesma pendente. É a representación do

ANCOVA sen interacción. Pola contra, á dereita tense unha pendente diferente para cada escola; trátase do *ANCOVA* con interacción.

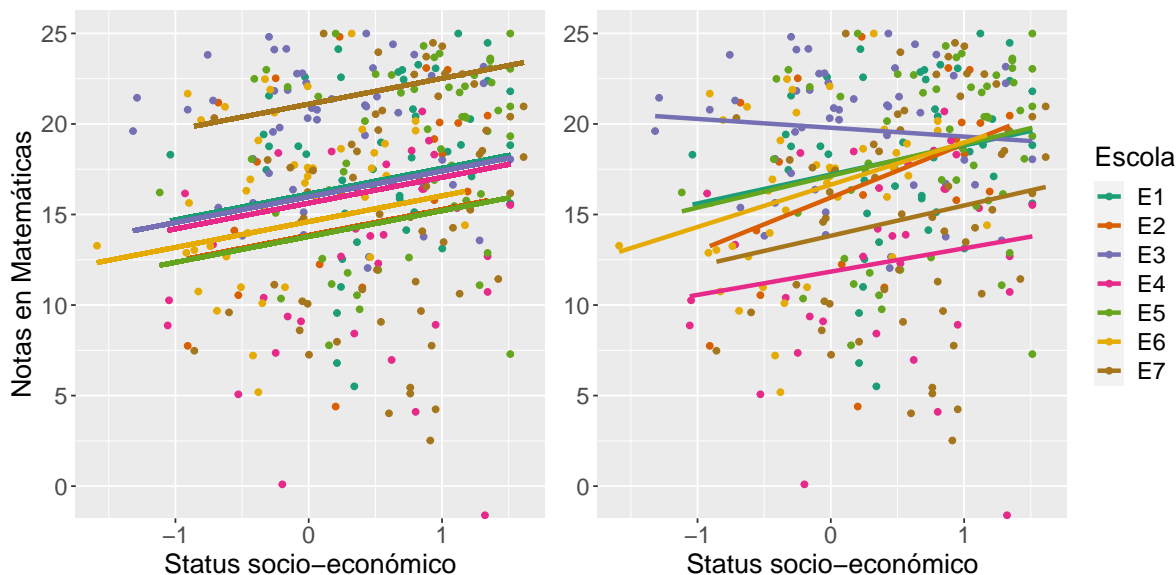



Figura 1.2: Á esquerda, un modelo de análise da covarianza sen interacción para as notas en Matemáticas fronte ao status socio-económico segundo as diferentes escolas. Á dereita incorpórase interacción ao modelo de análise da covarianza.

Analizando a significación dos coeficientes considerados en ambos modelos (facendo uso da función `anova` de  tal e como pode verse na Sección A.5), dedúcese ademais que o efecto das pendentes non é significativo, isto é, ao variar o status socio-económico (e mantendo fixas as demais características), as notas; sexan mellores ou peores, van mudar nunha cantidade semellante en todas as escolas. É importante salientar que isto non significa que o efecto dos nesgos sociais e económicos das/os estudantes nas notas acadadas por estas/es na materia de Matemáticas non varíe realmente duns centros educativos a outros. Tan só quere dicir que non varía ao considerar unicamente o efecto que causa o status socio-económico de cada menor en precisamente esas sete escolas.

Capítulo 2

Introdución aos modelos mixtos

Ao longo deste capítulo introducíranse os **datos multinivel** así como a necesidade de ter en conta as súas especiais características á hora de axustar modelos de regresión concernentes a datos deste tipo. Como resposta a esta necesidade natural xorden os **modelos mixtos**, tamén chamados modelos multinivel, modelos xerárquicos ou modelos de efectos aleatorios. Tal e como se verá máis adiante, o modelo mixto máis sinxelo de todos é o denominado modelo *RANOVA* ou modelo de análise da varianza con efectos aleatorios, que non é máis que a extensión natural do xa estudado modelo *ANOVA*.

2.1. Datos multinivel

Ao igual que no modelo *ANOVA*, que lémbrese viña dado por

$$Y_{ij} = \mu_j + \varepsilon_{ij}, \text{ con } i = 1, \dots, n_j \text{ e } j = 1, \dots, J; \quad (2.1)$$

seguirase supoñendo que hai un total de J grupos con n_j individuos no grupo j , e que o tamaño da mostra total é de $n = n_1 + n_2 + \dots + n_J$. Os datos desta forma chámanse datos multinivel ou datos xerárquicos¹, neste caso, cunha estrutura de dous niveis, onde os individuos se atopan no nivel 1 e están aniñados en grupos no nivel 2. O concepto de datos con estrutura xerárquica é xa antigo, aínda que cobrou gran importancia nos últimos anos debido ao desenvolvemento da computación, e á súa relevancia en ámbitos diversos como a Medicina (ver, a modo de exemplo, [34]), a Economía ou as Ciencias Medioambientais (ver, a modo de exemplo, [11]). Os datos con estrutura xerárquica poden ser aínda moito máis complexos que os vistos ata agora e contar con máis niveis diferentes de agrupamento (datos multinivel). Unha condición necesaria e suficiente

¹De aquí en diante entenderase por xerarquía unha organización na cal determinados individuos están agrupados en diferentes grupos, os cales poden estar agrupados, á súa vez, en máis grupos de niveis superiores.

para que un conxunto de datos sexa xerárquico, é dicir, que estea composto por datos multinivel, é que un individuo só pode pertencer a un único grupo do nivel 2, así como a un único grupo de cada un dos niveis posteriores. Do mesmo xeito, un grupo do nivel 2 só pode pertencer a un único grupo no nivel 3, así como a un único grupo de cada un dos niveis posteriores ao seu; e o mesmo ten que ocorrer para todos os niveis presentes no conxunto de datos en consideración.

Exemplo 2.1. Os datos multinivel son bastante frecuentes no eido educativo, de feito, no Capítulo 1 xa se viu unha estrutura de dous niveis na que as/os alumnas/os compoñen o nivel 1 e á súa vez están aniñados en escolas, as cales constitúen o nivel 2; aínda que tamén se poden dar estruturas de máis de dous niveis, segundo [15], en Educación é común atoparse con estruturas de 5 niveis (estudiante, clase, escola, distrito e área xeográfica). No conxunto de datos `mates7` xa empregado anteriormente téñense dispoñibles as variables

- nota en Matemáticas (que se denota por `NotaMates`),
- status socio-económico (que se denota por `StSE`),
- pertenza a un grupo racial minoritario (que se denota por `SocialMin`) e
- sexo da/o alumna/o (que se denota por `Sexo`);

relativas ao alumnado e que serían polo tanto as variables asociadas ao nivel 1, mentres que as variables

- media dos status socio-económico do alumnado de cada escola (que se denota por `MStSE`),
- número de estudantes de cada escola (que se denota por `Tamaño`),
- carácter público ou privado de cada escola (que se denota por `Sector`),
- proporción de estudantes de cada escola que participan no estudo [4] (que se denota por `Particip`),
- medida do ambiente discriminatorio de cada escola (que se denota por `AmbDiscrim`) e
- posesión de máis do 40 % das/os matriculadas/os en cada escola que sexan procedentes de grupos raciais minoritarios (que se denota por `MaioriaMin`);

serían as asociadas ao nivel 2, debido a que estas non varían entre todas as observacións que incumben a unha mesma escola. Ademais, como xa se adiantaba ao final da Introducción, as notas acadadas en Matemáticas polas/os estudantes probablemente estén enormemente influenciadas polas diferentes escolas.


Nesta situación, é importante destacar que a nota acadada na materia de Matemáticas por dúas/dous alumnas/os de distintos institutos que quizáis nin sequera se coñecen é totalmente independente. Agora ben, isto non ocorre con nenas e nenos que acoden diariamente á mesma escola, posto que terán bastante en común; dende o seu estilo de vida até a organización das clases de Matemáticas; pode ser incluso que lles imparta a materia o mesmo docente. Por conseguinte, a asunción de independencia feita nos modelos clásicos estudados ao longo do Grao non é asumible nesta situación; e resulta naturalmente necesario construír novos modelos que teñan en conta a característica inherente que é a xerarquía do conxunto de datos que se está a tratar.



2.2. A necesidade de ter en conta os distintos niveis

Cando os individuos forman grupos ou *clusters*, é obvio pensar que o máis probable é que individuos clasificados nun mesmo grupo resulten ter un comportamento máis semellante que uns individuos calquesquera de grupos diferentes e polo tanto con menos información en común. Tal e como afirma Goldstein [13], incluso cando os grupos se asignan aleatoriamente (no peor dos casos) a miúdo tende a haber diferenzas entre estes. Unha asunción típica na Estatística clásica, en particular no modelo de regresión dun só nivel presentado en (2), é que as observacións son independentes e identicamente distribuídas. Agora ben, se o conxunto de datos está formado por datos multinivel e aínda que non se teña en conta o efecto dos grupos á hora de construír un modelo de regresión, entón a hipótese de independencia non se verifica. Un xeito de ter en conta os efectos grupais sería incluír no modelo variables *dummy*² como variables explicativas, tal e como se fixo previamente para os modelos *ANOVA* e *ANCOVA* (modelos de efectos fixos) en (1.1) e (1.6), respectivamente; os cales poderían ser apropiados se o interese principal fose facer Inferencia sobre precisamente os grupos presentes na mostra. Así e todo xorden impedimentos cando o número de grupos é grande, xa que o número de parámetros a estimar medra considerablemente e o modelo de regresión pode non ser o suficientemente eficiente. Ademais, se o interesante non son precisamente eses grupos, senón que se consideran como unha mostra (aleatoria) dunha poboación máis grande e o que se quere é facer Inferencia sobre todos os grupos en xeral; por exemplo, se en lugar das 7 escolas o realmente importante é sacar conclusións arredor de tódalas escolas estadounidenses, entón os modelos de efectos fixos *ANOVA* e *ANCOVA* deixan de ser válidos. É aquí onde xorden os modelos mixtos ou modelos multinivel, en particular; os **modelos mixtos con variable resposta continua** que se abordarán nas seguintes seccións. Ademais, de aquí en diante, consideraranse unicamente datos xerárquicos cunha estrutura de dous niveis, onde os individuos se atopan no nivel 1 e están aniñados en grupos no nivel 2. As observacións

²As variables *dummy* son unhas variables ficticias que serven para indicar a posible pertenza dos individuos a cada grupo, tomando o valor 1 no caso de que un individuo pertenza a un determinado grupo, e o 0 en caso contrario.

entre niveis ou grupos distintos serán independentes, mentres que as observacións dentro dun mesmo grupo resultarán dependentes entre si posto que pertencen á mesma subpoboación. Por conseguinte, falarase de dúas **fontes de variación: entre grupos e dentro dos grupos** ou intra-grupos.

Para o axuste con  de modelos mixtos empregarase o paquete `lme4` (acrónimo de *Linear Mixed-Effects Models using 'Eigen' and S4*) que pode consultarse en Bates et al. [3], que é unha versión máis moderna do antigo `nlme` que contiña o estudo de [4] de onde se extraeu a información para crear o conxunto de datos `mates7`. O paquete `lme4` emprega métodos de álgebra linear máis eficientes (os do paquete `Eigen`), ademais de ser máis rápido computacionalmente e empregar menos memoria. Resultará de especial interese a función `lmer`, que serve para axustar modelos mixtos lineares.

No seguinte capítulo, comezarase explicando o modelo mixto máis sinxelo (o coñecido como modelo *RANOVA*), que non é máis que a extensión natural do modelo *ANOVA*; e deseguido engadiranse ao modelo covariables medidas no nivel 1 (Capítulo 4) e concernentes ao nivel 2 (Capítulo 5); asemade irase ilustrando o comportamento destes novos métodos na práctica empregando a base de datos `mates7` a prol de construír modelos máis sofisticados que permitan explicar mellor as notas acadadas polo alumnado na materia de Matemáticas e coñecer que parte das notas é debida ao propio alumnado e cal foi debido ao “efecto” da escola, no cal van implicitamente incluídas as modalidades de ensino, a formación do profesorado, etc. No Anexo A reproducirase parte do código de  empregado. Lémbrese que a totalidade do código de  atoparase ademais no repositorio *Modelos_Mixtos_con_R* de *GitHub* e será de carácter enteiramente reproducible, tanto o relativo á construción dos modelos que se estudarán a continuación, como o atinente a todas as figuras ilustradas ao longo da totalidade deste TFG.

Capítulo 3

RANOVA

Tal e como se adiantaba no capítulo anterior, as seguintes páxinas adicaranse a explicar a natureza do modelo mixto máis sinxelo de todos, que resulta ser a extensión natural do modelo *ANOVA* visto no Capítulo 1. De aquí en diante, seguirase supoñendo que se dispón dunha mostra aleatoria de datos multinivel, Y_{ij} , cunha estrutura de dous niveis, onde o segundo nivel está constituído por J grupos e o nivel 1 confórmano n_j individuos de cada grupo j , con $j = 1, \dots, J$. Xa se reparou en que existen ocasións nas que os grupos non teñen un interés intrínseco, senón que constitúen unha mostra (aleatoria) dun conxunto de moitos máis grupos e o que se quere é facer Inferencia sobre todos os grupos. Por exemplo, en canto á base de datos `mates7`, o que realmente interesa non son eses 7 colexios, senón toda a poboación de escolas dos Estados Unidos de América. Nestas circunstancias, o *ANOVA* con efectos fixos presentado en (1.1) deixa de ser relevante e é necesario extendelo a un modelo que incorpore efectos aleatorios, isto é, un modelo no cal as medias dos grupos deixen de ser constantes para converterse en variables aleatorias que seguen unha distribución normal cuxa media é a media xeral e cuxa varianza determina a capacidade de influencia de cada grupo.

3.1. Análise da varianza con efectos aleatorios

Neste contexto xorde o modelo *RANOVA* ou **modelo de análise da varianza con efectos aleatorios** (coñecido na literatura anglosaxoa como *Random effects ANOVA*), que non é máis que o modelo mixto ou multinivel máis sinxelo de todos e que pode escribirse como

$$Y_{ij} = \mu + u_j + \varepsilon_{ij}, \text{ con } i = 1, \dots, n_j \text{ e } j = 1, \dots, J; \quad (3.1)$$

onde μ é a media global, os $u_j \in N(0, \sigma_u^2)$ son independentes e identicamente distribuídos, os $\varepsilon_{ij} \in N(0, \sigma_\varepsilon^2)$ son independentes, e tamén u_j e ε_{ij} son variables aleatorias independentes entre si

para cada $j = 1, \dots, J$ e $i = 1, \dots, n_j$. Con σ_u^2 e σ_ε^2 denotamos as varianzas entre grupos e dentro dos grupos, respectivamente.

O modelo (3.1) é similar ao modelo *ANOVA* presentado en (1.1). A maior distinción é que no *ANOVA* as medias dos grupos μ_j son constantes, mentres que no *RANOVA* tense que $\mu_j = \mu + u_j \in N(\mu, \sigma_u^2)$, con $j = 1, \dots, J$, son variables aleatorias. Por ese motivo o modelo (3.1) é chamado modelo de análise da varianza con efectos aleatorios. Precisamente unha das vantaxes dos modelos mixtos é a habilidade de combinar os datos introducindo efectos aleatorios multinivel.

No *RANOVA* pénsase *a priori* nunha cantidade non fixa e sen límite de grupos, o interesante non son os μ_1, \dots, μ_J particulares do estudo, senón toda a posible poboación de μ_j , especialmente a media dos μ_j , μ ; e a variabilidade dos μ_j , medida por σ_u^2 . Mentres que σ_u^2 é unha medida directa da variabilidade dos μ_j , o efecto desta variabilidade sóese medir pola razón

$$VPC = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_\varepsilon^2} = \frac{\text{variabilidade entre grupos}}{\text{variabilidade total}}, \quad (3.2)$$

denominado **coeficiente de partición da varianza** (coñecido polas siglas en inglés VPC). Nótese que esta razón toma valores entre 0 (cando $\sigma_u^2 = 0$, logo $\mu_j = \mu$ para todo $j = 1, \dots, J$ e non hai diferenzas entre grupos, co cal (3.1) non é máis que unha regresión linear ordinaria) e 1 (cando $\sigma_\varepsilon^2 = 0$ ou $Y_{ij} = Y_j$ para todo i , é dicir, non hai diferenzas dentro de cada grupo). Nótese que o denominador de *VPC* representa a variabilidade da variable resposta Y , pois

$$\text{Var}(Y_{ij}) = \text{Var}(u_j + \varepsilon_{ij}) = \text{Var}(u_j) + \text{Var}(\varepsilon_{ij}) + 2\text{Cov}(u_j, \varepsilon_{ij}) = \sigma_u^2 + \sigma_\varepsilon^2,$$

xa que $\text{Var}(u_j) = \sigma_u^2$, $\text{Var}(\varepsilon_{ij}) = \sigma_\varepsilon^2$ e as variables aleatorias u_j e ε_{ij} son independentes entre si. Por conseguinte, o modelo *RANOVA* verifica a hipótese de homocedasticidade. En vista destas propiedades, a razón VPC mide a proporción da variabilidade total dos Y_{ij} que é explicada pola variabilidade dos μ_j , isto é, a proporción total de varianza atribuíble ás diferenzas entre grupos. Deste xeito, cando o cociente (3.2) toma valores próximos a cero, o efecto das diferenzas entre os grupos na variabilidade total é insignificante, e cando a razón é maior ou igual a 0.5 considerarase que unha cantidade considerable da variabilidade total é explicada polas diferenzas entre grupos.

Por outra banda, a covarianza entre dous individuos (que se denotarán por i e i') de distintos grupos (que se denotarán por j e j') é nula, isto é:

$$\begin{aligned} \text{Cov}(Y_{ij}, Y_{i'j'}) &= \text{Cov}(u_j + \varepsilon_{ij}, u_{j'} + \varepsilon_{i'j'}) = \\ &= \text{Cov}(u_j, u_{j'}) + \text{Cov}(u_j, \varepsilon_{i'j'}) + \text{Cov}(\varepsilon_{ij}, u_{j'}) + \text{Cov}(\varepsilon_{ij}, \varepsilon_{i'j'}) = 0, \end{aligned}$$

por causa de tódalas hipóteses de independencia asumidas previamente; mentres que para dous individuos i e i' dun mesmo grupo j , resulta que

$$\begin{aligned} \text{Cov}(Y_{ij}, Y_{i'j}) &= \text{Cov}(u_j + \varepsilon_{ij}, u_j + \varepsilon_{i'j}) = \\ &= \text{Cov}(u_j, u_j) + \text{Cov}(u_j, \varepsilon_{i'j}) + \text{Cov}(\varepsilon_{ij}, u_j) + \text{Cov}(\varepsilon_{ij}, \varepsilon_{i'j}) = \\ &= \text{Cov}(u_j, u_j) = \sigma_u^2. \end{aligned}$$

Deste xeito, a correlación entre dous individuos dun mesmo grupo resulta ser

$$p = \text{Cor}(Y_{ij}, Y_{i'j}) = \frac{\text{Cov}(Y_{ij}, Y_{i'j})}{\sqrt{\text{Var}(Y_{ij})}\sqrt{\text{Var}(Y_{i'j})}} = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_\varepsilon^2},$$

a que se denominará **correlación intra-grupos** de nivel 2. Ao longo deste traballo, na análise de datos reais denotarémolo por **correlación intra-escola**¹, ao igual que decide facelo Goldstein en [13]. Por conseguinte, en modelos con só dous niveis o VPC coincide co coeficiente de correlación intra-escola p ; que proporciona unha medida da homoxeneidade dos individuos dentro de cada grupo. En particular, segundo [16], os valores da correlación intra-escola no eido educativo soen estar entre 0.05 e 0.20. Finalmente, nótese que isto é certo para modelos de dous niveis e só para estes; é sinxelo darse de conta de que ao engadir outro nivel ao modelo, por exemplo as clases nas que están ubicadas/os as/os estudantes, a correlación p xa non tería a mesma expresión, senón unha notablemente máis complexa.

3.2. Estimación dos parámetros

No modelo de análise da varianza con efectos aleatorios, os erros de primeiro e segundo nivel son variables aleatorias $N(0, \sigma_u^2)$ e $N(0, \sigma_\varepsilon^2)$, respectivamente; e o seu comportamento queda caracterizado entón polas varianzas σ_u^2 e σ_ε^2 . Así, os parámetros a estimar no *RANOVA* son precisamente μ , σ_u^2 e σ_ε^2 .

3.2.1. Estimación da media global μ

Considéranse as medias grupais $\bar{Y}_{\bullet j} = \mu + u_j + \bar{\varepsilon}_{\bullet j}$, con $j = 1, \dots, J$. Así,

$$\begin{aligned} \mathbb{E}(\bar{Y}_{\bullet j}) &= \mathbb{E}(\mu + u_j + \bar{\varepsilon}_{\bullet j}) = \mu + \mathbb{E}(u_j) + \mathbb{E}(\bar{\varepsilon}_{\bullet j}) = \mu, \quad \text{e} \\ \text{Var}(\bar{Y}_{\bullet j}) &= \text{Var}(\mu + u_j + \bar{\varepsilon}_{\bullet j}) = \text{Var}(u_j + \bar{\varepsilon}_{\bullet j}) = \\ &= \text{Var}(u_j) + \text{Var}(\bar{\varepsilon}_{\bullet j}) + 2\text{Cov}(u_j, \bar{\varepsilon}_{\bullet j}) = \sigma_u^2 + \frac{\sigma_\varepsilon^2}{n_j}; \end{aligned}$$

onde a última igualdade é certa debido a que as variables aleatorias u_j e ε_{ij} son independentes entre si e mais ao teorema de Fisher². Deste xeito, a media da mostra de cada grupo é un estimador innesgado da media global; pero non é consistente, no sentido de que a varianza nunca será nula por moito que aumente o número de datos do grupo j . Para solventar o problema e mellorar

¹No ámbito da estatística este termo é frecuentemente coñecido por correlación intra-clase (ou *ICC*, empregando a notación de [9]), pero isto pode resultar confuso ao empregalo no eido educativo.

²Se $X = (X_1, \dots, X_n)$ é unha mostra aleatoria simple dunha poboación $N(\mu, \sigma^2)$, entón $\bar{X} \in N(\mu, \sigma^2/n)$ e $n\text{Var}(X)/\sigma^2 \in \chi_{n-1}^2$. A demostración do Teorema de Fisher pódese consultar en [19, p. 66].

o estimador, denotando $Var(\bar{Y}_{\bullet j}) = \sigma_u^2 + \frac{\sigma_\varepsilon^2}{n_j} = \Delta_j$, basta con considerar unha combinación das medias da mostra,

$$\hat{\mu} = \sum_{j=1}^J \frac{\Delta_j^{-1}}{\sum_{h=1}^J \Delta_h^{-1}} \bar{Y}_{\bullet j}; \quad (3.3)$$

onde Δ_j^{-1} é a “precisión” de $\bar{Y}_{\bullet j}$. Cando a precisión é igual en todos os grupos, o estimador obtido non é máis que o promedio global, isto é,

$$\hat{\mu} = \frac{\Delta^{-1} \sum_{j=1}^J \bar{Y}_{\bullet j}}{J \Delta^{-1}} = \frac{1}{J} \sum_{j=1}^J \bar{Y}_{\bullet j} = \bar{Y}_{\bullet\bullet} \text{ se } \Delta_j^{-1} = \Delta^{-1} \text{ para todo } j = 1, \dots, J.$$

Pola contra, se algunha precisión difire das outras; é necesario obter un estimador para $Var(\bar{Y}_{\bullet j}) = \Delta_j$, e este depende de σ_u^2 e σ_ε^2 ; cuxas estimacións serán as seguintes en procurar obterse.

3.2.2. Estimación das compoñentes da varianza

Analogamente ao feito para o modelo *ANOVA* en (1.4), considérase a seguinte descomposición da variabilidade da variable de interese Y :

$$\sum_{j=1}^J \sum_{i=1}^{n_j} (Y_{ij} - \bar{Y}_{\bullet\bullet})^2 = \sum_{j=1}^J \sum_{i=1}^{n_j} (Y_{ij} - \bar{Y}_{\bullet j})^2 + \sum_{j=1}^J n_j (\bar{Y}_{\bullet j} - \bar{Y}_{\bullet\bullet})^2; \quad (3.4)$$

e seguindo a mesma notación, (3.4) pode escribirse de xeito equivalente como:

$$\underbrace{VT}_{\text{variabilidade total}} = \underbrace{VDG}_{\text{variabilidade dentro dos grupos}} + \underbrace{VEG}_{\text{variabilidade entre grupos}}.$$

Para estimar a varianza de primeiro nivel σ_ε^2 , considérase $\hat{\sigma}_\varepsilon^2 = \frac{VDG}{N-J}$, onde segue a denotarse $N = \sum_{j=1}^J n_j$. Para a variabilidade entre grupos (ou de segundo nivel), se todos os grupos constan do mesmo número de observacións (datos balanceados), entón podemos considerar $S_u^2 = \frac{VEG}{n(J-1)}$, poñendo $n = n_j$ para un $j \in \{1, \dots, J\}$ calquera. Ocorre que $\mathbb{E}(\hat{\sigma}_\varepsilon^2) = \sigma_\varepsilon^2$, mentres que $\mathbb{E}(S_u^2) = \sigma_u^2 + \frac{\sigma_\varepsilon^2}{n} > \sigma_u^2$; en consecuencia, o estimador $\hat{\sigma}_\varepsilon^2$ é innesgado, mentres que non o é S_u^2 . Para anular o nesgo deste último, basta con considerar $\hat{\sigma}_u^2 = S_u^2 - \frac{\hat{\sigma}_\varepsilon^2}{n}$, establecendo $\hat{\sigma}_u^2 = 0$ no caso de que o resultado fose negativo, entendéndoo como que non hai efecto de grupos. A demostración destes feitos atópase deseguido e está baseada nos cálculos que fixo Kutner para o modelo *ANOVA* con efectos fixos, que poden consultarse en [20, pp. 538–542].

Proposición 3.1. *Baixo as hipóteses formuladas para o modelo RANOVA e dados $\hat{\sigma}_\varepsilon^2 = \frac{VDG}{N-J}$ e $S_u^2 = \frac{VEG}{n(J-1)}$, verifícase que*

$$(i) \quad \mathbb{E}(\hat{\sigma}_\varepsilon^2) = \sigma_\varepsilon^2,$$

$$(ii) \mathbb{E}(S_u^2) = \sigma_u^2 + \frac{\sigma_\varepsilon^2}{n}.$$

Demostración.

(i) Denotando a cuasivarianza das observacións nun grupo j por S_j^2 , tense que

$$\begin{aligned} \hat{\sigma}_\varepsilon^2 &= \frac{1}{N-J} \sum_{j=1}^J \sum_{i=1}^{n_j} (Y_{ij} - \bar{Y}_{\bullet j})^2 = \frac{1}{N-J} \sum_{j=1}^J \left[(n_j - 1) \frac{\sum_{i=1}^{n_j} (Y_{ij} - \bar{Y}_{\bullet j})^2}{n_j - 1} \right] = \\ &= \frac{1}{N-J} \sum_{j=1}^J (n_j - 1) S_j^2, \end{aligned}$$

e empregando que S_j^2 é un estimador innesgado da varianza dentro dos grupos (corolario inmediato do Teorema de Fisher), séguese que, efectivamente;

$$\mathbb{E}(\hat{\sigma}_\varepsilon^2) = \frac{1}{N-J} \sum_{j=1}^J (n_j - 1) \mathbb{E}(S_j^2) = \frac{1}{N-J} \sum_{j=1}^J (n_j - 1) \sigma_\varepsilon^2 = \sigma_\varepsilon^2.$$

(ii) Consideraranse datos balanceados, así pois, $n_j = n$ para todo j . Deste xeito,

$$S_u^2 = \frac{VEG}{n(J-1)} = \frac{1}{J-1} \sum_{j=1}^J (\bar{Y}_{\bullet j} - \bar{Y}_{\bullet\bullet})^2.$$

Agora ben,

$$\begin{cases} \bar{Y}_{\bullet j} = \mu + u_j + \bar{\varepsilon}_{\bullet j}, \text{ con } \bar{\varepsilon}_{\bullet j} = \frac{1}{n_j} \sum_{i=1}^{n_j} \varepsilon_{ij}, \text{ e} \\ \bar{Y}_{\bullet\bullet} = \mu + \bar{u}_{\bullet} + \bar{\varepsilon}_{\bullet\bullet}, \end{cases}$$

logo $\bar{Y}_{\bullet j} - \bar{Y}_{\bullet\bullet} = (u_j - \bar{u}_{\bullet}) + (\bar{\varepsilon}_{\bullet j} - \bar{\varepsilon}_{\bullet\bullet})$; e elevando ao cadrado e sumando por grupos séguese que

$$\sum_{j=1}^J (\bar{Y}_{\bullet j} - \bar{Y}_{\bullet\bullet})^2 = \underbrace{\sum_{j=1}^J (u_j - \bar{u}_{\bullet})^2}_{(a)} + \underbrace{\sum_{j=1}^J (\bar{\varepsilon}_{\bullet j} - \bar{\varepsilon}_{\bullet\bullet})^2}_{(b)} + 2 \underbrace{\sum_{j=1}^J (u_j - \bar{u}_{\bullet})(\bar{\varepsilon}_{\bullet j} - \bar{\varepsilon}_{\bullet\bullet})}_{(c)}.$$

Debido a que a esperanza é un operador linear, para calcular a esperanza do termo da esquerda pódese calcular a de todos estes sumandos por separado e logo sumalas.

(a) Analogamente ao feito en (i),

$$\sum_{j=1}^J (u_j - \bar{u}_{\bullet})^2 = (J-1) \frac{\sum_{j=1}^J (u_j - \bar{u}_{\bullet})^2}{J-1} = (J-1) S_{u_j}^2;$$

onde $S_{u_j}^2$ denota a cuasivarianza dos u_j e polo tanto

$$\mathbb{E} \left(\sum_{j=1}^J (u_j - \bar{u}_{\bullet})^2 \right) = (J-1) \mathbb{E}(S_{u_j}^2) = (J-1) \sigma_u^2.$$

- (b) Basta darse de conta de que $\sum_{j=1}^J (\bar{\varepsilon}_{\bullet j} - \bar{\varepsilon}_{\bullet\bullet})^2 / (J - 1)$ é unha varianza na mostra, e polo tanto un estimador innesgado da varianza da variable $\bar{\varepsilon}_{\bullet j}$; pero $\bar{\varepsilon}_{\bullet j}$ é precisamente a media de n erros independentes ε_{ij} ; e entón do Teorema de Fisher séguese que

$$\text{Var}(\bar{\varepsilon}_{\bullet j}) = \frac{\text{Var}(\varepsilon_{ij})}{n} = \frac{\sigma_\varepsilon^2}{n} \Rightarrow \mathbb{E} \left(\sum_{j=1}^J (\bar{\varepsilon}_{\bullet j} - \bar{\varepsilon}_{\bullet\bullet})^2 \right) = \frac{(J - 1)\sigma_\varepsilon^2}{n}.$$

- (c) Pola independencia entre si das variables aleatorias u_j e ε_{ij} e mais a linearidade da esperanza, pódese escribir

$$\mathbb{E} \left(2 \sum_{j=1}^J (u_j - \bar{u}_{\bullet}) (\bar{\varepsilon}_{\bullet j} - \bar{\varepsilon}_{\bullet\bullet}) \right) = 2 \sum_{j=1}^J \mathbb{E}(u_j - \bar{u}_{\bullet}) \mathbb{E}(\bar{\varepsilon}_{\bullet j} - \bar{\varepsilon}_{\bullet\bullet}).$$

Agora ben, xa que $\mathbb{E}(\varepsilon_{ij}) = 0$, entón $\mathbb{E}(\bar{\varepsilon}_{\bullet j}) = 0$ e $\mathbb{E}(\bar{\varepsilon}_{\bullet\bullet}) = 0$; co cal $\mathbb{E}(\bar{\varepsilon}_{\bullet j} - \bar{\varepsilon}_{\bullet\bullet}) = 0$ e a esperanza de (c) é nula.


Finalmente, pódese concluír que efectivamente

$$\mathbb{E}(S_u^2) = \frac{1}{J - 1} \mathbb{E} \left(\sum_{j=1}^J (\bar{Y}_{\bullet j} - \bar{Y}_{\bullet\bullet})^2 \right) = \sigma_u^2 + \frac{\sigma_\varepsilon^2}{n}.$$

□

Para poder establecer o estimador da varianza entre grupos σ_u^2 inicialmente proposto (S_u^2) supúxose que o número de observacións en cada grupo era o mesmo. En caso contrario, a estimación de σ_u^2 requerirá dunha estimación previa do efecto fixo (media global), que á súa vez precisa das estimacións da varianza. Neste caso, para obter as estimacións dos parámetros do *RANOVA* é necesario recorrer a procedementos iterativos.

Cando se estima un modelo multinivel mediante o método de máxima verosimilitude (ML na literatura inglesa), a dependencia entre os parámetros (efectos fixos e compoñentes da varianza) reflectida previamente para o caso particular do modelo *RANOVA* leva implícito recorrer a procedementos iterativos; primeiramente estímáanse os efectos fixos, por exemplo mediante o algoritmo *EM* (esperanza-maximización) que considera os efectos aleatorios como faltantes, ou ben empregando o método de mínimos cadrados xeralizados iterativo (*IGLS*, acrónimo de *Iterative Generalised Least Squares*), sobre os cales se pode atopar máis información en [28, pp. 41–42] ou en [25, pp. 164–165], respectivamente. Deseguido, empréganse as estimacións dos efectos fixos obtidas nesta primeira iteración para estimar as compoñentes da varianza, e á súa vez estas estimacións das compoñentes da varianza empréganse para obter unhas novas estimacións dos efectos fixos na seguinte iteración. Continuando deste xeito até que non haxa cambios significativos entre iteracións consecutivas obteranse as estimacións procuradas, tanto dos efectos fixos

como das compoñentes da varianza. Na Figura 3.1 amósase un bosquexo do que fai  cando se emprega o método ML para estimar os parámetros dun modelo mixto.

Pola contra, o método de máxima verosimilitude produce estimacións nesgadas dos parámetros aleatorios, posto que estimar os efectos fixos en primeiro lugar implica que toda a variabilidade da mostra debida a estes sexa ignorada, co cal as compoñentes da varianza son subestimadas, incrementándose así tanto os t-valores como os p-valores ou niveis críticos asociados. Isto é realmente preocupante en bases de datos pequenas, en concreto, cun número de grupos reducido, xa que a variabilidade da mostra dos efectos fixos tende a ser máis grande; asunto que se trata máis profundamente en [22], onde tamén se propón unha solución; o método de máxima verosimilitude restrinxida, coñecido polas súas siglas en inglés REML.

O método REML, ao contrario que ML, estima os efectos fixos e as compoñentes da varianza por separado. Ambos procedementos son asintoticamente equivalentes e igualmente complexos no relativo ao eido da computación. Para evitar a estimación simultánea que realiza ML, REML non fai máis que restrinxir a 0 os efectos fixos nun primeiro momento, permitindo estimar así as compoñentes da varianza separadamente. Finalmente, estímense os efectos fixos empregando as estimacións das compoñentes da varianza obtidas previamente, mediante unha modificación do método *IGLS* coñecida por *RIGLS* (acrónimo de *Restricted Iterative Generalised Least Squares*) e sen empregar procedementos iterativos. Toda a teoría concernente ao método *RIGLS* explícaa Goldstein en [12]. Na parte inferior da Figura 3.1 ilústrase o procedemento que se segue cando se emprega o método REML no contexto dos modelos multinivel.

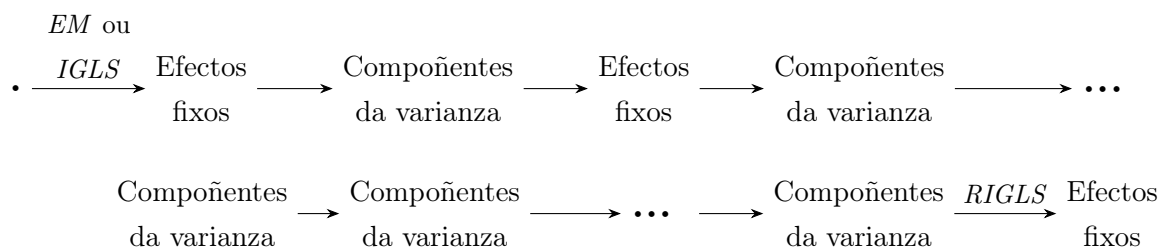


Figura 3.1: Comparación entre os procedementos *ML* (parte superior) e *REML* (parte inferior).

En definitiva, por unha banda, tal e como afirma [7] convén buscar a sinxeleza dos modelos; mentres que por outra banda, en canto ao procedemento por máxima verosimilitude, estimando os efectos fixos en primeiro lugar, ademais de ignorarse a variabilidade da mostra debida a estes, tampouco se teñen en conta os graos de liberdade que se perden estimándoos. É por iso que ao longo deste documento se decida traballar co método ML para os modelos mixtos máis sinxelos mentres que do Capítulo 4 en diante, ao engadir efectos aleatorios nas pendentes asociadas a covariables relativas ao nivel 1, asemade os graos de liberdade comezan a escasear, se procure empregar máis o método REML.

3.2.3. Predición dos efectos aleatorios

No *ANOVA* con efectos fixos as medias de cada grupo estimábanse polas medias na mostra, empregando para a estimación da media dun grupo j unicamente os individuos dese mesmo grupo j . Na análise da varianza con efectos aleatorios, pola contra, a media do grupo j sería $\beta_{0j} = \mu + u_j \in N(\mu, \sigma_u^2)$, que é unha variable aleatoria; e por esa razón se fala de predicións en lugar de estimacións. Ademais, no *RANOVA* uns grupos están ligados con outros (todos os u_j proceden da mesma distribución normal); o que permite obter unha predición da media de cada grupo máis eficiente que a que se obtería considerando só os individuos de cada grupo. Con esta notación, o modelo *RANOVA* introducido en (3.1) tamén se pode escribir como,

$$Y_{ij} = \beta_{0j} + \varepsilon_{ij}, \text{ con } \beta_{0j} \in N(\mu, \sigma_u^2) \text{ e } \varepsilon_{ij} \in N(0, \sigma_\varepsilon^2);$$

para $i = 1, \dots, n_j$ e $j = 1, \dots, J$, co que se consegue un modelo da mesma forma que a do modelo de análise da varianza con efectos fixos dado en (1.1), só que neste caso cun intercepto aleatorio. Para calcular a predición dos efectos aleatorios $\beta_{0j} = \mu + u_j$ empregarase a información proporcionada polos dous niveis de xerarquía do modelo.

Nivel 1: tendo en conta só este nivel, non quedaría outra que empregar o modelo para as medias da mostra de cada grupo $\bar{Y}_{\bullet j} = \beta_{0j} + \bar{\varepsilon}_{\bullet j}$; con $\bar{\varepsilon}_{\bullet j} \in N(0, \sigma_\varepsilon^2/n_j)$, onde a anterior distribución é consecuencia directa do Teorema de Fisher. Condicionalmente a β_{0j} , é dicir, se non fose aleatorio senón fixo, $\bar{Y}_{\bullet j}$ sería un estimador innesgado de β_{0j} ; e a varianza condicional sería $\text{Var}(\bar{Y}_{\bullet j}|\beta_{0j}) = \sigma_\varepsilon^2/n_j$ para cada grupo j .

Nivel 2: tendo en consideración agora só o nivel 2, é dicir, a información aportada soamente polos grupos; como $\beta_{0j} = \mu + u_j \in N(\mu, \sigma_u^2)$, entón podemos considerar como estimador para todos os β_{0j} o xa planteado en (3.3),

$$\hat{\mu} = \sum_{j=1}^J \frac{\Delta_j^{-1}}{\sum_{h=1}^J \Delta_h^{-1}} \bar{Y}_{\bullet j};$$

o cal é, condicionalmente a β_{0j} , nesgado; aínda que ao empregar información de todos os grupos e non só do j -ésimo a variabilidade podería reducirse considerablemente.

A predición óptima sería polo tanto a media ponderada de ámbolos dous estimadores, o proporcionado polo nivel 1 (individuos) e mais o proporcionado polo nivel 2 (grupos):

$$\hat{\beta}_{0j} = \underbrace{\lambda_j \bar{Y}_{\bullet j}}_{\text{nivel 1}} + \underbrace{(1 - \lambda_j) \hat{\mu}}_{\text{nivel 2}}; \text{ con } \lambda_j = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \hat{\sigma}_\varepsilon^2/n_j}.$$

O coeficiente λ_j denomínase **factor de fiabilidade** ou *shrinkage*. Ao medrar o cociente das varianzas estimadas entre grupos e dentro dos grupos $\hat{\sigma}_u^2/\hat{\sigma}_\varepsilon^2$ ou o tamaño do grupo j , n_j , aumenta

tamén λ_j ; ao mesmo tempo que o estimador $\hat{\beta}_{0j}$ se aproxima á predición baseada unicamente nos datos do grupo j considerado (nivel 1).

Así, a predición do efecto aleatorio é a seguinte,

$$\hat{u}_j = \hat{\beta}_{0j} - \hat{\mu} = \lambda_j \bar{Y}_{\bullet j} + (1 - \lambda_j) \hat{\mu} - \hat{\mu} = \lambda_j \underbrace{(\bar{Y}_{\bullet j} - \hat{\mu})}_{\text{residuos naive}} ;$$

onde a denominación de residuos *naive* é debida a que as diferenzas $(\bar{Y}_{\bullet j} - \hat{\mu})$ non son máis que os residuos do axuste de mínimos cadrados dunha regresión onde as observacións da variable resposta son precisamente os promedios dos grupos. Como $\lambda_j \leq 1$, os valores \hat{u}_j serán menores ou iguais que os residuos respecto á media. No caso de haber pouca información sobre os grupos ($\lambda_j \simeq 1$), entón as predicións que se fan dos efectos aleatorios \hat{u}_j están próximas aos residuos, isto é, ás diferenzas entre as medias do grupo e a media global.

Finalmente, nótese que, unha vez obtidas as predicións dos residuos de nivel 2, as estimacións dos residuos de nivel 1 poden obterse de xeito sinxelo a partir das anteriores,

$$\hat{\varepsilon}_{ij} = Y_{ij} - \hat{\beta}_{0j} = Y_{ij} - \hat{\mu} - \hat{u}_j.$$

Tal e como se contempla na Figura 3.2, resultan ser a distancia de cada dato á media do seu respectivo grupo.

Exemplo 3.2. Tal e como se foi adiantando, axustarase un modelo *RANOVA* sobre a nota de Matemáticas acadada por cada alumna/o (nivel 1) considerando que están aniñadas/os en escolas (nivel 2). Amósase a continuación a sintaxe da función `lmer`, que se empregará para o axuste de modelos mixtos. Debe terse en conta que a parte `(1 | Escola)` indica que se está a axustar un modelo mixto só con intercepto aleatorio en función dos grupos que establece a variable `Escola`. Ademais, o argumento `REML = FALSE` serve para implementar estimación por máxima verosimilitude (en lugar da estimación por máxima verosimilitude restrinxida ou *REML*), que lémbrese decidiuse empregala neste traballo para os modelos máis sinxelos, mentres non resulte ser un impedimento.

```
ranovamates = lmer(NotaMates ~ (1 | Escola), data = mates7, REML = FALSE)
summary(ranovamates)

## Linear mixed model fit by maximum likelihood ['lmerMod']
## Formula: NotaMates ~ (1 | Escola)
## Data: mates7
##
##          AIC          BIC    logLik deviance df.resid
##   1880.3    1891.5   -937.2   1874.3      304
```

```
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -3.5716 -0.6463  0.1723  0.7398  1.9971
##
## Random effects:
##   Groups   Name      Variance Std.Dev.
##   Escola   (Intercept) 4.699    2.168
##   Residual                24.967    4.997
## Number of obs: 307, groups:  Escola, 7
##
## Fixed effects:
##              Estimate Std. Error t value
## (Intercept)   16.684      0.872    19.13
```

A saída do `summary` (que amosa un resumo do modelo axustado) consta de tres partes, no comezo detállase o modelo e móstrase información sobre este, como criterios de axuste globais; entre eles o *AIC* e mais o *BIC*³. A segunda parte (*Random effects*) recolle as estimacións da varianza e da desviación típica dos coeficientes aleatorios, neste caso, dos residuos do nivel 2 e dos residuos do nivel 1. Así, $\hat{\sigma}_u^2 = 4.699$ e $\hat{\sigma}_\varepsilon^2 = 24.967$. Debaixo amósase o número total de observacións das que se dispón, xunto co número de grupos de cada nivel superior do modelo. Nestas circunstancias, como só se está a considerar un nivel superior e este é o relativo ás escolas; manifesta que se teñen en conta os efectos de exactamente sete escolas. A parte final é a saída dos efectos fixos (*Fixed effects*), táboa que contén a estimación, o erro típico estimado e o t-valor para todos os parámetros fixos do modelo. Neste caso, a media global estímase por $\hat{\mu} = 16.684$ mentres que a local para cada escola j estímase por $\hat{\mu} + \hat{u}_j$, onde \hat{u}_j é o residuo de tal colexio.

Nótese que na saída anterior non se amosa o p-valor para o intercepto fixo. Isto será así para todos os coeficientes fixos sempre que se traballe coa función `lmer` e é debido a que a distribución exacta dos estatísticos dos tests baixo a hipótese nula (sen efectos fixos) non se coñece. No seguinte, baseándose no criterio de Roback e Legler presentado en [27], considerarase que t-valores⁴ con valor absoluto maior que 1.96 indican evidencias estatisticamente significativas a favor da hipótese alternativa de que o parámetro é distinto de 0. Isto é debido a que nunha normal estándar o 95 % dos valores baixo a curva atópanse entre -1.96 e 1.96 , e entón debido ao Teorema Central do Límite (pódese consultar en [23, p. 157]), este número pódese empregar

³O Criterio de Información de Akaike (*AIC*) e o Criterio de Información de Bayes (*BIC*) son medidas globais do modelo que teñen en conta o axuste e á vez compensan o exceso de parámetros. Pode atoparse máis información para ámbolos dous criterios nos documentos orixinais de Akaike [1] e [2].

⁴Cociente entre a estimación dun parámetro e a súa desviación típica estimada.

para construír intervalos de confianza do 95 %. Finalmente, se o intervalo de confianza obtido non contén ao 0, concluírase que non hai evidencias de que o parámetro sexa nulo.

Claramente, unha escola con $\hat{u}_j > 0$ ten notas por enriba da media global, mentres que os colexios con $\hat{u}_j < 0$ quedarán por debaixo da media global. Máis adiante farase Inferencia sobre estes residuos para determinar que diferenzas sobre a media global se poden considerar significativas e cales debidas á casualidade; de xeito semellante aos contrastes feitos para o ANOVA.


Por outra banda, o coeficiente de partición da varianza estimado resulta ser

$$\widehat{VPC} = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \hat{\sigma}_\varepsilon^2} = \frac{4.699}{4.699 + 24.967} = 0.1584.$$

Polo tanto, un 15.84 % da variabilidade do modelo explícase por diferenzas entre as escolas (nivel 2) e o 84.16 % restante, por diferenzas dentro das propias escolas, isto é, entre o alumnado (nivel 1). Ademais, o coeficiente de correlación entre dous/dúas alumnas/os escollidas/os ao azar en calquera escola (correlación intra-escola) será de $\hat{\rho} = 0.1584$. O feito de que este valor se ubique máis próximo a 0.20 que a 0.05 tamén o explica [16], posto que este conclúe que as correlacións intra-escola tenden a ser maiores en cursos menos avanzados; e nestas circunstancias estase a tratar con alumnado de arredor de 11 anos.

Na Figura 3.2 ilústrase a verdadeira natureza do modelo *RANOVA*. A media global das notas acadadas, que se estima por $\hat{\mu} = 16.684$, represéntase cunha liña negra, mentres que as medias para as escolas E3, E4, E5 e E7 ($\hat{\mu}_j$ con $j = 3, 4, 5, 7$) evócanse mediante liñas descontinuas. Tamén se exemplifican algúns residuos, tanto relativos ao nivel 1 como ao nivel 2, ademais dos datos asociados a estes.

Unha particularidade do *RANOVA* é que os residuos de primeiro nivel $\hat{\varepsilon}_{ij}$ (en negro) resultan ser a distancia de cada dato á media do seu respectivo grupo, que ademais segue unha distribución normal, posto que $(\mu + u_j) \in N(\mu, \sigma_u^2)$ para $j = 1, 2, \dots, 7$. Os residuos de segundo nivel, claro está, non son máis que a distancia da media global $\hat{\mu}$ á respectiva media local $\hat{\mu}_j$ para algún $j = 1, \dots, 7$; e veñen representados polas frechas máis curtas da esquerda da gráfica.

Na Táboa 3.1 calculada coa axuda de  amósanse as predicións obtidas para os erros a nivel escola u_j , seguidas das súas desviacións típicas estimadas $\hat{\sigma}(u_j)$ e da súa posición ao ordealas de menor a maior, ademais das estimacións das medias locais $\hat{\beta}_{0j}$ para as notas de cada escola, dos residuos *naive* (diferenzas entre as medias grupais da mostra e a estimación da media global) e dos valores que toma o factor de fiabilidade ou *shrinkage* λ_j ; respectivamente.

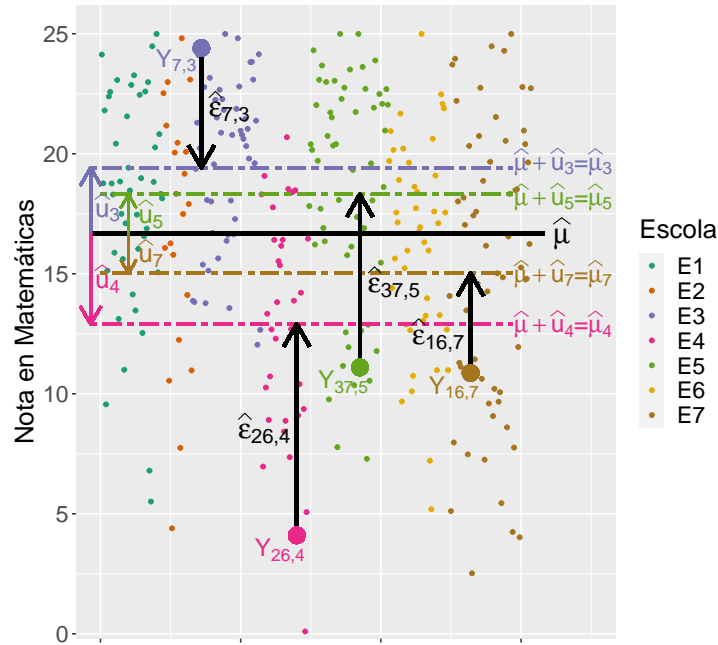



Figura 3.2: Ilustración do modelo de análise da varianza con efectos aleatorios para a nota de Matemáticas nas 7 escolas con participación absoluta, xunto coa media global, algunhas medias locais e mais algúns residuos concernentes a ambos niveis (alumnado e escolas).

Escola	\hat{u}_j	$\hat{\sigma}(u_j)$	Posición	$\hat{\beta}_{0j}$	$\bar{Y}_{\bullet j} - \hat{\mu}$	λ_j
E1	1.274	0.712	5	17.958	1.428	0.892
E2	0.224	0.974	4	16.907	0.280	0.798
E3	2.735	0.678	7	19.419	3.032	0.902
E4	-3.782	0.797	1	12.901	-4.374	0.865

Táboa 3.1: Cabeceira do conxunto de datos consistente nas predicións dos erros a nivel escola e mais algúns outros parámetros asociados ao modelo *RANOVA* construído, obtido grazas a .

Así, como \hat{u}_3 é o valor máis grande, a escola E3 é a que mellor promedio de notas en Matemáticas tende a ter. Analogamente, o colexio E4 é o que peores notas acada; aínda que a última afirmación é menos sólida, posto que o erro da escola E4 ten unha desviación típica estimada lixeiramente máis alta que o do centro E3 (0.797 fronte a 0.678).

Co obxectivo de manifestar a incerteza das predicións dos \hat{u}_j debida á variabilidade da mostra, represéntanse as estimacións ordeadas de menor a maior xunto cuns intervalos de confianza de nivel 95 % para estes. Este tipo de representacións coñécense como gráfico de eiruga⁵.

⁵Para conxuntos de datos cun número de grupos de nivel 2 máis elevado, o gráfico da Figura 3.3 conta cos intervalos de confianza moito máis apegados e verdadeiramente semella unha eiruga (ver exemplo en [29, p. 67]).

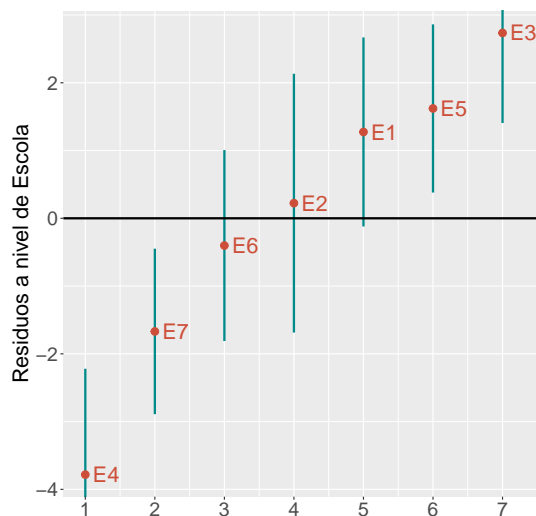


Figura 3.3: Gráfico de eiruga para as predicións dos erros a nivel escola xunto con intervalos de confianza asociados a estes.

As escolas cuxos intervalos de confianza non intersecan o cero teñen un efecto sobre as notas de Matemáticas significativamente distinto ao efecto promedio das escolas en xeral. Así, os centros con intervalos por enriba do 0 son os mellores; mentres que os colexios con intervalos por debaixo do 0 son peores.

Reparando na Figura 3.3 séguese que 3 escolas se atopan sobre a media, 2 son mellores e outras 2 son peores que a media. Ademais, en concordancia co que se agoiraba contemplando a Táboa 3.1, semella que a escola E4 ten notas peores ao resto, ao igual que o colexio E3 semella ser o mellor. Pola contra, as aseveracións realizadas non son tan claras; cando varios intervalos de confianza se sobrepoñen uns sobre outros, non se está en condicións de afirmar que tales residuos de segundo nivel u_j son diferentes. Nótese que a afirmación de que a escola E4 é peor que o resto é a máis segura, no sentido de que hai tres escolas cuxos intervalos de confianza se superpoñen co da escola E3, mentres que só o dunha se solapa co do colexio E4.

Nótese que no caso de querer facer calquer contraste múltiple, como estudar que escolas son significativamente distintas da escola E3; ao igual que na Sección 1.1.4 é necesario recorrer a niveis de significación modificados e empregar métodos como os de Bonferroni, Holm ou Tukey.

Finalmente, que a escola E3 teña o factor de fiabilidade $\lambda_3 = 0.902$ quere dicir que a predición elaborada polo modelo *RANOVA* ($\hat{\beta}_{0,3} = 19.419$) está relativamente próxima á predición baseada unicamente nos datos da escola E3 ($\bar{Y}_{\bullet,3} = 19.716$). De feito, a predición do efecto aleatorio é $\hat{u}_3 = \lambda_3(\bar{Y}_{\bullet,3} - \hat{\mu})$; e polo tanto \hat{u}_3 é un 90.2 % do residuo *naive* desa mesma escola.

3.3. Contraste sobre os efectos grupais

Cando o VPC se atopa próximo a cero, así pois, σ_u^2 é case nulo, parece haber evidencias de que realmente non hai ningunha diferenza verdadeira entre os distintos grupos. Para contrastar esta afirmación pódese empregar un contraste de hipóteses a fin de comprobar se o efecto das diferenzas entre os grupos é nulo ou difire máis do que se podería esperar só por azar. O contraste a estudar é o seguinte:

$$\begin{cases} H_0 : \text{non hai diferenzas entre os grupos,} \\ H_a : \text{hai diferenzas entre os grupos.} \end{cases} \iff \begin{cases} H_0 : \sigma_u^2 = 0, \\ H_a : \sigma_u^2 \neq 0. \end{cases} \quad (3.5)$$

Como $u_j \in N(0, \sigma_u^2)$, se a hipótese nula é certa entón necesariamente todos os u_j son nulos. Logo, o contraste anterior pode pensarse como un contraste entre dous modelos diferentes, que ademais están aniñados⁶:

$$\begin{cases} H_0 : Y_{ij} = \mu + \varepsilon_{ij}, \\ H_a : Y_{ij} = \mu + u_j + \varepsilon_{ij}. \end{cases} \quad (3.6)$$

en ambos casos con $i = 1, \dots, n_j$ e $j = 1, \dots, J$. En canto ao modelo baixo a hipótese nula, coñécese habitualmente como **modelo dun só nivel para a media**, posto que tamén se podería expresar como $Y_i = \mu + \varepsilon_i$ con $i = 1, \dots, n$; é dicir, sen ter en conta os grupos e reflectindo simplemente a desviación de cada dato con respecto á media global. Tales diferenzas entre o valor da variable resposta Y e a media global veñen dadas por ε_i , que non é máis que o erro para cada individuo i no modelo; e estes seguen unha distribución normal de media cero e unha varianza σ^2 , isto é, $\varepsilon_i \in N(0, \sigma^2)$ para cada dato $i = 1, \dots, n$. A varianza σ^2 simboliza a variabilidade ao redor da media; deste xeito, se fose cero, todos os puntos terían o mesmo valor. Do mesmo xeito, canto máis grande é a varianza, máis grandes son as desviacións sobre a media.

Na Figura 3.4 ilústrase a natureza do modelo dun só nivel para a media, en concreto, para o caso particular das sete escolas. É evidente a diferenza imperante entre ámbolos dous modelos do contraste (3.6). Mentres que na Figura 3.2 concernente ao *RANOVA* os residuos ao nivel alumnado resultaban ser a distancia de cada dato á media do seu respectivo grupo; no modelo baixo a hipótese nula, ao non reparar no efecto das escolas, os residuos (de cor negra) non son máis que as distancias de cada dato á media global μ , tal e como se amosa na Figura 3.4.

⁶Dous modelos dinse aniñados se o modelo máis complexo se pode construír a partir do máis simple engadindo un ou máis parámetros. Por exemplo, considérense tres modelos $M1$, $M2$ e $M3$. $M1$ considera como variable explicativa a $X1$, $M2$ considera a $X1$ e a $X2$ e $M3$ considera a $X1$ e a $X3$. $M1$ e $M2$ están aniñados, así como tamén $M1$ e $M3$; pero non o están $M2$ e $M3$.

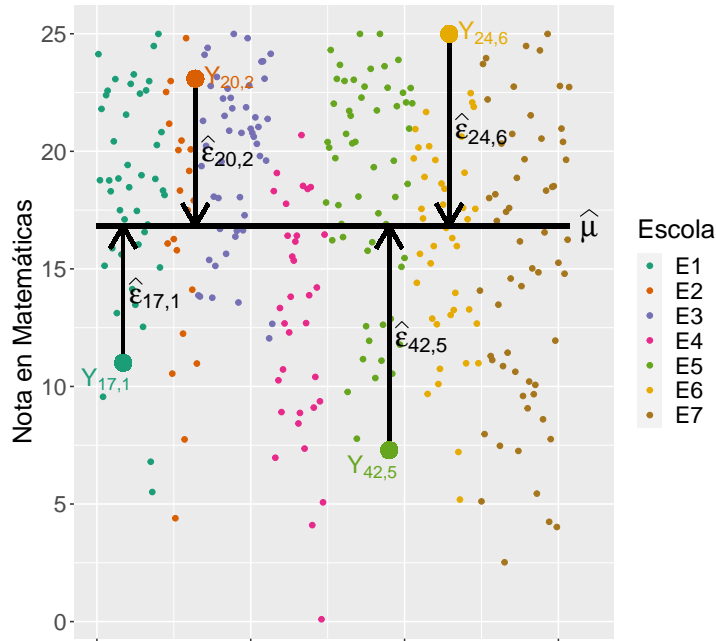


Figura 3.4: Modelo dun só nivel para a media (global), sen considerar o efecto das escolas, xunto con algúns residuos (de cor negra) e cos seus respectivos datos da cor correspondente á escola.

Agora ben, xa que ámbolos dous modelos están aniñados, á hora de construír un estatístico para realizar o contraste (3.5), podemos comparar os dous modelos de (3.6) mediante un **test de razón de verosimilitudes**, isto é, empregando o estatístico


$$LR = -2 \log \left(\frac{L_1}{L_2} \right) = -2 \log L_1 - (-2 \log L_2) \in \chi_1^2,$$

onde L_1 é a verosimilitude⁷ para o modelo baixo a hipótese nula en (3.6), igualmente L_2 é a verosimilitude para o modelo de análise da varianza con efectos aleatorios e “log” refírese ao logaritmo natural. O estatístico LR segue unha distribución chi cadrado cun único grao de liberdade, xa que o modelo máis complexo soamente incorpora un parámetro adicional con respecto ao máis sinxelo, que precisamente é a varianza entre grupos σ_u^2 . A demostración deste feito omítese neste traballo pero pode atoparse en [26, pp. 417–419].

Rexeitar a hipótese nula implica que hai evidencias estatisticamente significativas a favor da existencia de diferenzas entre os distintos grupos, en cuxo caso un modelo multinivel sería máis axeitado que un modelo que non tivese en conta os grupos. Agora ben, se cun determinado conxunto de datos non existen evidencias en contra da hipótese nula e entón non se pode rexeitar, isto non quere dicir que non se deban ter en conta os grupos á hora de axustar un modelo para

⁷Considérese un vector aleatorio X con función de densidade f_θ , sendo $\theta \in \mathbb{R}^q$ un vector de parámetros descoñecido. Entón, dada X_1, \dots, X_n unha mostra aleatoria simple de X , pódese estimar θ empregando a función de verosimilitude; que vén dada por $L(\theta) = \prod_{i=1}^n f(x_i(\theta))$ e de onde se obtén o estimador facendo $\hat{\theta} = \text{máx}_{\theta \in \mathbb{R}^q} L(\theta)$.

eses datos; é posible que as diferenzas entre os diferentes grupos se revelen soamente logo de engadir máis variables explicativas ao modelo e que permanezan agochadas ao considerar tan só as diferenzas entre os grupos relativas á media.

Exemplo 3.3. Realizarase o contraste sobre os efectos das distintas escolas para ver se efectivamente a escola á que asista un/unha estudante ten influencia sobre a nota que acade en Matemáticas. Para axustar o modelo dun só nivel para a media das notas en Matemáticas baixo a H_0 en (3.6) abonda con considerar un modelo de regresión para a media da escola dos xa vistos na Introducción mediante a función `lm` de .

```
vcmates <- lm(NotaMates ~ 1, data = mates7)
```

Tras o axuste deste modelo, obtense que a estimación da media global μ non é outra que a media dos valores almacenados na variable `NotaMates`, isto é, $\hat{\mu} = 16.823$. O valor do estatístico para o test de razón de verosimilitudes e o p-valor asociado ao contraste calcúlanse deseguido:

```
LR <- -2 * logLik(vcmates)[1] - (-2 * logLik(ranovamates))[1]
1 - pchisq(LR, df = 1)
```

Na Figura 3.5 represéntase a función de densidade χ_1^2 que segue o estatístico LR , cuxo valor vén dado pola liña vermella e, como se pode apreciar, deixa unha probabilidade moi baixa á súa dereita (1.96×10^{-9}). Así pois, existen evidencias estatisticamente significativas a favor da hipótese alternativa de (3.5) de que hai diferenzas entre as distintas escolas en canto á nota acadada polas/os súas/seus estudantes na materia de Matemáticas.

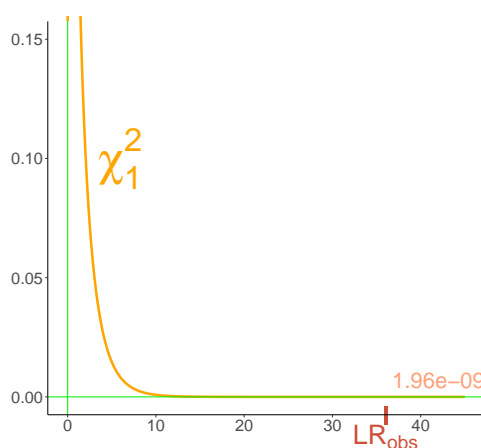


Figura 3.5: Función de densidade dunha chi-cadrado cun grao de liberdade en laranxa xunto co valor do estatístico do contraste sobre os efectos das escolas LR (de cor vermella) e mais o p-valor asociado a tal contraste.

Capítulo 4

Modelos mixtos con covariables relativas ao primeiro nivel

De xeito análogo ao que se fixo no Capítulo 1 introducindo os modelos *ANOVA* e *ANCOVA*, considerarase agora unha extensión do modelo de análise da varianza con efectos aleatorios ou *RANOVA* engadindo unha ou máis variables continuas que estén relacionadas coa variable resposta Y . Consideraranse neste capítulo modelos que inclúan soamente unha variable explicativa relativa ao primeiro nivel, o cal será suficiente para explicar con detalle a natureza destes. Os datos desta covariable denotaranse por X_{ij} , con $j = 1, \dots, J$ e $i = 1, \dots, n_j$; e o efecto desta pode ser ou ben fixo (modelo con intercepto aleatorio e pendente fixa) ou ben aleatorio (modelo con intercepto e pendente aleatorias). A continuación describiranse este tipo de modelos.

4.1. Modelo con intercepto aleatorio

Como se viu anteriormente, un modelo de análise da varianza con efectos aleatorios pode escribirse como

$$Y_{ij} = \mu + u_j + \varepsilon_{ij}, \text{ con } j = 1, \dots, J \text{ e } i = 1, \dots, n_j.$$

Pois ben, para ter en conta a información sobre os individuos que proporciona unha certa variable X e considerando fixo o efecto desta, basta con engadila ao modelo *RANOVA* multiplicada por un coeficiente fixo, do xeito:

$$Y_{ij} = \beta_{0j} + \beta_1 X_{ij} + \varepsilon_{ij}, \text{ con } j = 1, \dots, J \text{ e } i = 1, \dots, n_j; \quad (4.1)$$

onde $\beta_{0j} = \mu + u_{0j} \in N(\mu, \sigma_{u0}^2)$ é un intercepto aleatorio que varía entre os distintos grupos; mentres que a pendente β_1 é a mesma para todos os grupos. No eido dos modelos multinivel, a este modelo coñéceselle por **modelo con intercepto aleatorio** (e pendente fixa). De maneira

análoga ao presentado no Capítulo 3, μ representa a media global, os $u_{0j} \in N(0, \sigma_{u0}^2)$ son independentes e identicamente distribuídos, os $\varepsilon_{ij} \in N(0, \sigma_\varepsilon^2)$ son independentes, e tamén u_{0j} e ε_{ij} son variables aleatorias independentes entre si para cada $j = 1, \dots, J$ e $i = 1, \dots, n_j$. Destas propiedades séguese ademais que os interceptos aleatorios β_{0j} son independentes entre si e independentes dos erros de primeiro nivel.

Como interpretación xeométrica, a representación gráfica do axuste dun modelo multinivel con intercepto aleatorio sería bastante similar á mostrada na Figura 3.2 elaborada para o *RANOVA*, só que agora tanto a recta de regresión axustada para todos os datos sen ter en conta os grupos (a media global) como as rectas de regresión axustadas para os distintos grupos terán todas unha pendente β_1 ; como consecuencia de considerar o efecto fixo da covariable X . É dicir, as liñas da forma $\hat{Y}_{ij} = \hat{\mu} + \hat{u}_{0j} + \hat{\beta}_1 X_{ij}$, con $j = 1, \dots, J$; conformarán polo tanto un conxunto de rectas paralelas. Na Figura 4.1 ilústranse estes feitos, amosando a recta de regresión xeral xunto coas rectas estimadas para os colexios E3, E4, E5 e E7.

Recórdese que o modelo de análise da covarianza con efectos fixos ou *ANCOVA*, explicado na Sección 1.2, vén dado por

$$Y_{ij} = \mu + \tau_j + \gamma X_{ij} + \varepsilon_{ij}, \text{ con } j = 1, \dots, J \text{ e } i = 1, \dots, n_j;$$

e, polo tanto, fixado un grupo, non é máis ca unha recta de regresión con intercepto $(\mu + \tau_j)$ e pendente γ (a mesma para todos os grupos). Este modelo é bastante similar ao modelo con intercepto aleatorio que se está a considerar pero con interceptos fixos en lugar de aleatorios. Por tanto, aínda que se construíu o modelo con intercepto aleatorio como unha extensión natural do *RANOVA*, tamén se podería elaborar a partir do modelo de análise da covarianza con efectos fixos ou *ANCOVA*. Deste xeito, o modelo con intercepto aleatorio non é máis ca un modelo *ANCOVA* (sen interacción) con efectos aleatorios, é dicir, un *RANCOVA*¹ sen interacción.

O modelo con intercepto aleatorio e pendente fixa verifica a hipótese de homocedasticidade, xa que

$$\begin{aligned} \text{Var}(Y_{ij}|X = X_{ij}) &= \text{Var}(\mu + u_{0j} + \beta_1 X_{ij} + \varepsilon_{ij}|X = X_{ij}) = \text{Var}(u_{0j} + \varepsilon_{ij}) = \\ &= \text{Var}(u_{0j}) + \text{Var}(\varepsilon_{ij}) + 2\text{Cov}(u_{0j}, \varepsilon_{ij}) = \sigma_{u0}^2 + \sigma_\varepsilon^2, \end{aligned}$$

por ser as variables aleatorias u_{0j} e ε_{ij} independentes entre si. Ademais, a covarianza entre


¹Non é común o emprego desta denominación na literatura sobre os modelos multinivel, só se emprega neste caso particular para clarificar que o modelo con intercepto aleatorio non é máis ca un *ANCOVA* con efectos aleatorios (*Random effects ANCOVA*).

observacións de distintos grupos sempre é nula²:


$$\begin{aligned} Cov(Y_{ij}, Y_{i'j'} | X_{ij}, X_{i'j'}) &= Cov(\mu + u_{0j} + \beta_1 X_{ij} + \varepsilon_{ij}, \mu + u_{0j'} + \beta_1 X_{i'j'} + \varepsilon_{i'j'} | X_{ij}, X_{i'j'}) = \\ &= Cov(u_{0j} + \varepsilon_{ij}, u_{0j'} + \varepsilon_{i'j'}) = Cov(u_{0j}, u_{0j'}) + Cov(u_{0j}, \varepsilon_{i'j'}) + \\ &+ Cov(\varepsilon_{ij}, u_{0j'}) + Cov(\varepsilon_{ij}, \varepsilon_{i'j'}) = 0, \end{aligned}$$

por ser as variables aleatorias u_{0j} e ε_{ij} independentes separadamente. Por outra banda, entre observacións (i e i') dun mesmo grupo j resulta ser

$$\begin{aligned} Cov(Y_{ij}, Y_{i'j} | X_{ij}, X_{i'j}) &= Cov(\mu + u_{0j} + \beta_1 X_{ij} + \varepsilon_{ij}, \mu + u_{0j} + \beta_1 X_{i'j} + \varepsilon_{i'j} | X_{ij}, X_{i'j}) = \\ &= Cov(u_{0j} + \varepsilon_{ij}, u_{0j} + \varepsilon_{i'j}) = Cov(u_{0j}, u_{0j}) + Cov(u_{0j}, \varepsilon_{i'j}) + \\ &+ Cov(\varepsilon_{ij}, u_{0j}) + Cov(\varepsilon_{ij}, \varepsilon_{i'j}) = Var(u_{0j}) = \sigma_{u0}^2. \end{aligned}$$

Exemplo 4.1. Nas seguintes liñas, extenderase o modelo *RANOVA* construído previamente mediante a introdución dunha variable explicativa concernente ao nivel 1 do alumnado, o status socio-económico (que se vén denotando por **StSE**). Xa se viu na Sección 1.2 relativa á construción do modelo *ANCOVA* que o efecto do **StSE** nas notas acadadas por un/unha alumna/o era significativo. Na Figura 4.1 represéntase o axuste do modelo obtido coa seguinte sintaxe en :

```
matesmm1 <- lmer(NotaMates ~ StSE + (1 | Escola), REML = FALSE)
```

A pendente $\hat{\beta}_1 = 1.397$ é efectivamente a mesma para todas as escolas, mentres que o intercepto é diferente para cada colexio; o que ocasiona rectas paralelas. A media global estímase por $\hat{\mu} = 16.156$, que non é máis que o intercepto da recta negra asociada á escola “media” ($\hat{u}_{0j} = 0$), entendéndoo como a nota estimada cando **StSE** = 0. Así, para obter o intercepto asociado á escola E3, por exemplo, non hai máis que sumarlle a $\hat{\mu}$ a predición do efecto aleatorio para tal colexio \hat{u}_3 obtido mediante a función **ranef** de ; resultando $16.156 + 3.037 = 19.193$, valor que predí a recta descontinua azul cando **StSE** = 0.

Por outra banda, as estimacións da varianza dos coeficientes aleatorios resultan ser $\hat{\sigma}_{u0}^2 = 4.845$ e $\hat{\sigma}_{\varepsilon}^2 = 24.168$. Reparando nas disimilitudes deste modelo e do *RANOVA* advírtese que a adición do **StSE** reduciu a varianza a nivel de alumnado e mais a varianza total, o cal era esperado porque **StSE** é unha variable concernente ao nivel das/os estudantes. A varianza a nivel escolar, pola contra, incrementouse lixeiramente. Isto é debido a que o status socio-económico non se distribúe de forma moi regular entre as escolas. Aínda que en todos os centros hai alumnas e alumnos máis desfavorecidas/os e máis beneficiadas/os, non todos os colexios se ubican en lugares

²Denótase $Cov(Y_{ij}, Y_{i'j'} | X_{ij}, X_{i'j'})$ por comodidade o que realmente son covarianzas condicionais: $Cov(Y_{ij}, Y_{i'j'} | X = X_{ij}, X = X_{i'j'})$, e o mesmo acontece con todos os termos sucesivos.

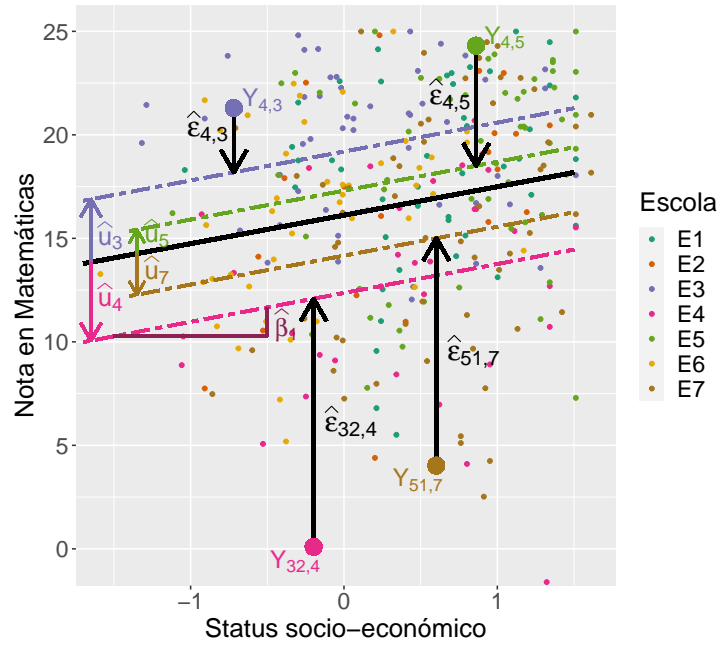


Figura 4.1: Ilustración do modelo mixto con intercepto aleatorio e pendente fixa para a nota de Matemáticas nas 7 escolas con participación unánime, xunto coa recta de regresión para a escola media, E3, E4, E5, E7 e mais algúns residuos concernentes a ambos niveis (alumnado e escolas).

coas mesmas características sociais ou económicas nin tampouco son frecuentados polo mesmo tipo de alumnado. Pode atoparse máis información arredor deste fenómeno que ocorre sobre as varianzas ao engadir variables explicativas de nivel 1 aos modelos nos apuntamentos do curso *LEMMA* impartido polo Centro de Modelos Multinivel da Universidade de Bristol [30]. A Figura 4.1 é un claro exemplo deste fenómeno, xa que os puntos de determinadas cores tenden a situarse máis cara a parte esquerda do gráfico mentres que outros se espallan máis cara a dereita, até os valores máis altos do *StSE*. Tal é o caso, por exemplo, das escolas E4 e E5. O alumnado do centro educativo E4 (de cor rosa) ten, en promedio, un status socio-económico claramente menor que o da escola E5 (de cor verde). De feito, a media do *StSE* para o colexio E4 resulta ser 0.367, mentres que para o centro E5 é 0.759. Outra fonte de variación entre escolas pode verse tamén na Figura 4.1, posto que as rectas correspondentes a centros educativos con maior intercepto (liñas da parte superior), tenden a ter máis alumnas/os con *StSE* maiores que 0.5; mentres que as escolas con interceptos máis pequenos tenden a ter máis alumnas/os cun menor status socio-económico.

Logo de ter en conta os efectos do *StSE*, a proporción total de varianza atribuíble ás diferenzas entre as distintas escolas sería:

$$VPC = \frac{\hat{\sigma}_{u0}^2}{\hat{\sigma}_{u0}^2 + \hat{\sigma}_{\epsilon}^2} = \frac{4.845}{4.845 + 24.168} = 0.167 = 16.7\%,$$

fronte ao 15.84% sen considerar o efecto do *StSE*. Tal incremento é debido ao acrescentamento

da varianza a nivel dos centros educativos.

Finalmente, nótese que no caso de ser a covariable X unha variable dicotómica, μ sería a media total de Y para os individuos con $x = 0$, $\mu + u_{0j}$ sería a media para os individuos con $x = 0$ no grupo j , e a “pendente” β_1 sería a diferenza na media que posúen os datos con $x = 1$ relativa á dos datos con $x = 0$ (en calquer grupo). Ilústrase a continuación esta situación.

Exemplo 4.2. Co propósito de ver que ocorre ao considerar unha variable explicativa dicotómica e de mellorar o modelo construído no Exemplo 4.1, engádeselle a este o efecto fixo asociado á variable **SocialMin**, que lémbrese indicaba se a/o estudante é membro dun grupo racial minoritario ou non. Codificarase o valor “Non” por 0 e o “Si” por 1 (cando a/o menor pertence a un grupo racial minoritario). Así, o modelo a considerar é o seguinte,

$$\text{NotaMates}_{ij} = \underbrace{\mu + \beta_1 \text{StSE}_{ij} + \beta_2 \text{SocialMin}_{ij}}_{\text{efectos fixos}} + \underbrace{u_j}_{\text{efectos aleatorios}} + \varepsilon_{ij}.$$

A continuación, amósase un resumo do modelo axustado.

```
matesmm2 <- lmer(NotaMates ~ StSE + SocialMin + (1 | Escola), REML = FALSE)
summary(matesmm2)

## Linear mixed model fit by maximum likelihood ['lmerMod']
## Formula: NotaMates ~ StSE + SocialMin + (1 | Escola)
##
##      AIC      BIC   logLik deviance df.resid
##  1861.4   1880.0   -925.7   1851.4     302
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -3.8713 -0.6260  0.1161  0.7342  2.1362
##
## Random effects:
##   Groups   Name      Variance Std.Dev.
##   Escola   (Intercept)  4.35     2.086
##   Residual                23.17     4.814
## Number of obs: 307, groups:  Escola, 7
##
## Fixed effects:
##              Estimate Std. Error t value
## (Intercept)  16.9501     0.8817  19.223
```

```
## StSE          0.9937      0.4509      2.204
## SocialMinSi   -2.7527      0.7455     -3.692
##
## Correlation of Fixed Effects:
##              (Intr) StSE
## StSE          -0.240
## SocialMinSi   -0.244  0.241
```

Para un/unha mesma/o estudante, acrescentar o **StSE** exactamente nun punto suporía un incremento na nota de $\hat{\beta}_1 = 0.99$ puntos; independentemente de se a/o alumna/o procede dun grupo racial minoritario ou non. De xeito semellante, para estudantes co mesmo **StSE**; o simple feito de pertencer a un grupo racial minoritario (**SocialMin** = 1) automaticamente augura unha mingua na nota acadada de $|\hat{\beta}_2| = 2.75$ puntos. Por outra banda, a interpretación do intercepto fixo $\hat{\mu} = 16.95$ debería ser a do prognóstico para un/unha alumno/a calquera cun status socio-económico igual a 0, non procedente dun grupo racial minoritario e sen considerar o efecto da escola á que acode. Ademais, todos estes coeficientes resultan ser significativos segundo o criterio de [27], xa que os t-valores resultan todos maiores que 1.96 en valor absoluto.

Na Figura 4.2 represéntanse de cor negra as rectas de regresión xerais para a escola media ($\hat{u}_{0j} = 0$), unha relativa aos datos sobre alumnado minoritario e outra diferente para o non minoritario. Posto que $\hat{\beta}_2 = -2.753 < 0$, a liña relativa ás/aos menores procedentes de grupos raciais minoritarios é a que se atopa por debaixo; e o mesmo ocorre para cada par de liñas de regresión axustadas relativo a un determinado centro, en particular, para os pares de liñas asociadas ás escolas E4 e E5; que son as dúas que a modo de exemplo se representan na gráfica.

4.2. Modelo con intercepto e pendente aleatorios

Extenderase agora o modelo con intercepto aleatorio visto previamente mediante a incorporación de efectos aleatorios na pendente asociada á variable explicativa X , o que consistirá en formular un modelo linear en cada grupo $j = 1, \dots, J$ e ademais pode representarse como

$$Y_{ij} = \beta_{0j} + \beta_{1j}X_{ij} + \varepsilon_{ij}, \text{ con } j = 1, \dots, J \text{ e } i = 1, \dots, n_j; \quad (4.2)$$

onde $\beta_{0j} = \gamma_{00} + u_{0j}$ e $\beta_{1j} = \gamma_{10} + u_{1j}$ son variables aleatorias independentes separadamente (e independentes dos erros), con distribución normal bivalente;

$$\begin{pmatrix} \beta_{0j} \\ \beta_{1j} \end{pmatrix} \in N \left(\begin{pmatrix} \gamma_{00} \\ \gamma_{10} \end{pmatrix}, \Sigma_u = \begin{pmatrix} \sigma_{u0}^2 & \sigma_{u01} \\ \sigma_{u01} & \sigma_{u1}^2 \end{pmatrix} \right),$$

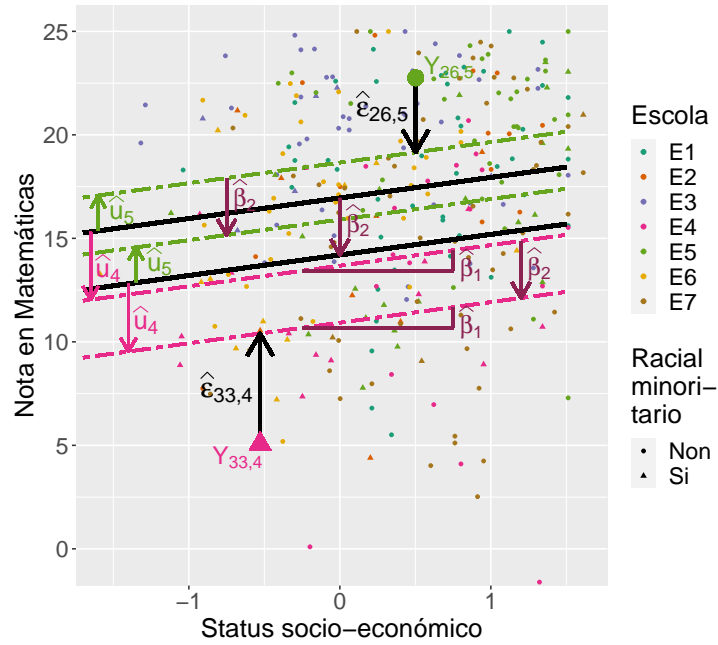


Figura 4.2: Ilustración do modelo mixto con intercepto aleatorio para a nota de Matemáticas nas 7 escolas considerando como variables explicativas *StSE* e *SocialMin*, xunto coas rectas de regresión para a escola media, E4, E5 e mais algúns residuos concernentes a ambos niveis.

onde γ_{00} e γ_{10} son o intercepto medio e a pendente media, respectivamente; σ_{u0}^2 é a varianza do intercepto, σ_{u1}^2 é a varianza da pendente e σ_{u01} é a covarianza entre intercepto e pendente. Seguiranse a supor as mesmas hipóteses sobre a independencia que ata agora; as variables aleatorias u_{0j} , u_{1j} e ε_{ij} son todas independentes e os efectos aleatorios do intercepto e da pendente son independentes dos erros de primeiro nivel. A diferenza reside en que agora hai variables non independentes entre si, posto que dentro de cada grupo pode existir correlación entre os efectos aleatorios do intercepto e da pendente.

Analogamente ao símil do modelo con intercepto aleatorio e un suposto *RANCOVA* sen interacción, poderíase aludir a este modelo como un *RANCOVA* con interacción, agora con efectos aleatorios tanto no intercepto como na pendente. Precisamente por variar o intercepto e a pendente entre os distintos grupos é polo que (4.2) se denomina **modelo con intercepto e pendente aleatorios**.

Ao igual que acontecía no Exemplo 4.1 coa adición dos efectos fixos dunha variable concernente ao nivel 1, a introdución de efectos aleatorios nunha variable a nivel de individuo reducirá a varianza nese nivel (σ_{ε}^2 , varianza dentro dos grupos); mentres que a varianza entre grupos (σ_{u0}^2 e σ_{u1}^2) pode manterse e mesmo aumentar.

Considerando que as fontes de variabilidade que interveñen no modelo se poden estruturar en

dous niveis, o modelo con intercepto e pendente aleatorios pode formularse de xeito xerárquico ou ben de forma combinada ou conxunta. Considerando $j = 1, \dots, J$ e $i = 1, \dots, n_j$ pódese escribir o modelo dos seguintes xeitos:

(i) Formulación xerárquica:

$$\text{Nivel 1 (individuos)} : Y_{ij} = \beta_{0j} + \beta_{1j}X_{ij} + \varepsilon_{ij}.$$

$$\text{Nivel 2 (grupos)} : \begin{cases} \beta_{0j} = \gamma_{00} + u_{0j}, \text{ con } u_{0j} \in N(0, \sigma_{u0}^2), \\ \beta_{1j} = \gamma_{10} + u_{1j}, \text{ con } u_{1j} \in N(0, \sigma_{u1}^2). \end{cases}$$

(ii) Formulación conxunta:
$$Y_{ij} = \underbrace{\gamma_{00} + \gamma_{10}X_{ij}}_{\text{efectos fixos}} + \underbrace{u_{1j}X_{ij} + u_{0j}}_{\text{efectos aleatorios}} + \varepsilon_{ij}.$$

Do mesmo xeito que no *RANOVA* os efectos aleatorios son desviacións da media global (do intercepto); neste contexto, os efectos aleatorios u_{0j} e u_{1j} poden interpretarse como desviacións dos parámetros (intercepto e pendente) no grupo j -ésimo con respecto ao intercepto medio e á pendente media, respectivamente. Ademais, o sumando $u_{1j}X_{ij}$ pode interpretarse como unha interacción entre o grupo e a variable explicativa (interacción entre niveis).

A diferenza do modelo con intercepto aleatorio visto en (4.1), que resultaba ser homocedástico, ao introducir efectos aleatorios na pendente asociada á variable explicativa X deixa de verificarse a hipótese de homocedasticidade, posto que

$$\begin{aligned} \text{Var}(Y_{ij}|X = X_{ij}) &= \text{Var}(\gamma_{00} + \gamma_{10}X_{ij} + u_{1j}X_{ij} + u_{0j} + \varepsilon_{ij}|X = X_{ij}) = \\ &= \text{Var}(u_{1j}X_{ij} + u_{0j} + \varepsilon_{ij}|X = X_{ij}) \stackrel{(a)}{=} \text{Var}(u_{1j}X_{ij}|X = X_{ij}) + \\ &+ \text{Var}(u_{0j}|X = X_{ij}) + \text{Var}(\varepsilon_{ij}|X = X_{ij}) + 2\text{Cov}(u_{1j}X_{ij}, u_{0j}|X = X_{ij}) + \\ &+ 2\text{Cov}(u_{1j}X_{ij}, \varepsilon_{ij}|X = X_{ij}) + 2\text{Cov}(u_{0j}, \varepsilon_{ij}|X = X_{ij}) = \\ &= X_{ij}^2 \text{Var}(u_{1j}|X = X_{ij}) + \text{Var}(u_{0j}|X = X_{ij}) + \text{Var}(\varepsilon_{ij}|X = X_{ij}) + \\ &+ 2X_{ij}\text{Cov}(u_{1j}, u_{0j}|X = X_{ij}) = \sigma_{u1}^2 X_{ij}^2 + \sigma_{u0}^2 + \sigma_{\varepsilon}^2 + 2\sigma_{u01}X_{ij} = \\ &= (\sigma_{u0}^2 + 2\sigma_{u01}X_{ij} + \sigma_{u1}^2 X_{ij}^2) + \sigma_{\varepsilon}^2, \end{aligned}$$

onde na igualdade (a) estase empregando a definición da varianza da suma dun certo número de variables aleatorias³; e o resto de igualdades son debidas ás suposicións figuradas sobre a independencia. Deste xeito, $\text{Var}(Y_{ij})$ depende do valor da covariable X e entón efectivamente o

³En particular, para tres variables aleatorias X , Y e Z ; abonda con razoer indutivamente que $\text{Var}(X+Y+Z) = \text{Var}((X+Y)+Z) = \text{Var}(X+Y) + \text{Var}(Z) + 2\text{Cov}(X+Y, Z) = \text{Var}(X) + \text{Var}(Y) + \text{Var}(Z) + 2\text{Cov}(X, Y) + 2\text{Cov}(X, Z) + 2\text{Cov}(Y, Z)$. En xeral, para certas variables X_1, \dots, X_n tense que

$$\text{Var}(X_1 + \dots + X_n) = \sum_{i=1}^n \text{Var}(X_i) + 2 \sum_{1 \leq i < j \leq n} \text{Cov}(X_i, X_j).$$

modelo non é homocedástico. De novo, a covarianza entre as observacións de distintos grupos é nula:

$$\begin{aligned}
Cov(Y_{ij}, Y_{i'j'} | X_{ij}, X_{i'j'}) &= Cov(\gamma_{00} + \gamma_{10}X_{ij} + u_{0j} + u_{1j}X_{ij} + \varepsilon_{ij}, \gamma_{00} + \gamma_{10}X_{i'j'} + u_{0j'} + \\
&\quad + u_{1j'}X_{i'j'} + \varepsilon_{i'j'} | X_{ij}, X_{i'j'}) = \\
&= Cov(u_{0j} + u_{1j}X_{ij} + \varepsilon_{ij}, u_{0j'} + u_{1j'}X_{i'j'} + \varepsilon_{i'j'} | X_{ij}, X_{i'j'}) = \\
&= Cov(u_{0j}, u_{0j'} | X_{ij}, X_{i'j'}) + Cov(u_{0j}, u_{1j'}X_{i'j'} | X_{ij}, X_{i'j'}) + \\
&\quad + Cov(u_{0j}, \varepsilon_{i'j'} | X_{ij}, X_{i'j'}) + Cov(u_{1j}X_{ij}, u_{0j'} | X_{ij}, X_{i'j'}) + \\
&\quad + Cov(u_{1j}X_{ij}, u_{1j'}X_{i'j'} | X_{ij}, X_{i'j'}) + Cov(u_{1j}X_{ij}, \varepsilon_{i'j'} | X_{ij}, X_{i'j'}) + \\
&\quad + Cov(\varepsilon_{ij}, u_{0j'} | X_{ij}, X_{i'j'}) + Cov(\varepsilon_{ij}, u_{1j'}X_{i'j'} | X_{ij}, X_{i'j'}) + \\
&\quad + Cov(\varepsilon_{ij}, \varepsilon_{i'j'} | X_{ij}, X_{i'j'}) = 0;
\end{aligned}$$


mentres que entre dúas observacións i e i' dun mesmo grupo j resulta ser

$$\begin{aligned}
Cov(Y_{ij}, Y_{i'j} | X_{ij}, X_{i'j}) &= Cov(\gamma_{00} + \gamma_{10}X_{ij} + u_{0j} + u_{1j}X_{ij} + \varepsilon_{ij}, \gamma_{00} + \gamma_{10}X_{i'j} + u_{0j} + \\
&\quad + u_{1j}X_{i'j} + \varepsilon_{i'j} | X_{ij}, X_{i'j}) = \\
&= Cov(u_{0j} + u_{1j}X_{ij} + \varepsilon_{ij}, u_{0j} + u_{1j}X_{i'j} + \varepsilon_{i'j} | X_{ij}, X_{i'j}) = \\
&= Cov(u_{0j}, u_{0j} | X_{ij}, X_{i'j}) + Cov(u_{0j}, u_{1j}X_{i'j} | X_{ij}, X_{i'j}) + \\
&\quad + Cov(u_{0j}, \varepsilon_{i'j} | X_{ij}, X_{i'j}) + Cov(u_{1j}X_{ij}, u_{0j} | X_{ij}, X_{i'j}) + \\
&\quad + Cov(u_{1j}X_{ij}, u_{1j}X_{i'j} | X_{ij}, X_{i'j}) + Cov(u_{1j}X_{ij}, \varepsilon_{i'j} | X_{ij}, X_{i'j}) + \\
&\quad + Cov(\varepsilon_{ij}, u_{0j} | X_{ij}, X_{i'j}) + Cov(\varepsilon_{ij}, u_{1j}X_{i'j} | X_{ij}, X_{i'j}) + \\
&\quad + Cov(\varepsilon_{ij}, \varepsilon_{i'j} | X_{ij}, X_{i'j}) = Var(u_{0j}) + X_{i'j}Cov(u_{0j}, u_{1j}) + \\
&\quad + X_{ij}Cov(u_{1j}, u_{0j}) + X_{ij}X_{i'j}Var(u_{1j}) = \\
&= \sigma_{u0}^2 + \sigma_{u01}(X_{ij} + X_{i'j}) + \sigma_{u1}^2X_{ij}X_{i'j}.
\end{aligned}$$

Nótese que na anterior ecuación, ao igual que se fixo na Sección 4.1 coas covarianzas do modelo con intercepto aleatorio, sempre que se escribe $Cov(Y_{ij}, Y_{i'j} | X_{ij}, X_{i'j})$, realmente estase a expresar $Cov(Y_{ij}, Y_{i'j} | X = X_{ij}, X = X_{i'j})$, isto é, trátase de covarianzas condicionais.

Exemplo 4.3. Seguindo coa base de datos que se vén empregando ao longo de todo o traballo, poderíase pensar que o status socio-económico dun menor (que se vén denotando por **StSE**) inflúe de xeito distinto nunhas escolas que noutras; en contraposición co que supón o modelo construído no Exemplo 4.2. Considerarase agora un modelo onde tanto o intercepto como a pendente son aleatorios, outorgándolle certa folgora á variable **StSE** para que a pendente asociada a esta sexa aleatoria. A variable concernente á pertenza a un grupo racial minoritario (que se vén denotando por **SocialMin**), pola contra, seguirase a considerar con efectos fixos. Así, o modelo a estudar é o seguinte,


$$\text{NotaMates}_{ij} = \underbrace{\gamma_{00} + \gamma_{10}\text{StSE}_{ij} + \beta_2\text{SocialMin}_{ij}}_{\text{efectos fixos}} + \underbrace{u_{0j} + u_{1j}\text{StSE}_{ij}}_{\text{efectos aleatorios}} + \varepsilon_{ij}. \quad (4.3)$$

Como xa se adiantaba nos capítulos previos, ao considerar pendentes aleatorias a complexidade do modelo increméntase notablemente, en particular no sentido de que ao contar soamente con 7 escolas non se dispón dos suficientes graos de liberdade como para permitirse ignorar os que se perden estimando os efectos fixos en primeiro lugar mediante o método de máxima verosimilitude. É por iso que tanto neste modelo como nos máis complexos que se consideren a partir deste momento se decidirá empregar o procedemento de máxima verosimilitude restrinxida (en  abondará con escribir `REML = TRUE` na sintaxe da función `lmer`) para estimar os parámetros, que lembrese a solución que daba para evitar a estimación simultánea que realizaba o método de máxima verosimilitude era restrinxir a 0 os efectos fixos nun primeiro momento para poder estimar así as compoñentes da varianza separadamente. Móstrase deseguido o axuste do modelo (4.3).

```
matesmm3 <- lmer(NotaMates ~ StSE + SocialMin + (1 + StSE | Escola),
  REML = TRUE)
summary(matesmm3)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: NotaMates ~ StSE + SocialMin + (1 + StSE | Escola)
##
## REML criterion at convergence: 1848.1
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -3.8890 -0.5973  0.1115  0.7313  2.1702
##
## Random effects:
##   Groups      Name                Variance Std.Dev. Corr
##   Escola  (Intercept)    5.66032  2.3791
##           StSE           0.06546  0.2559   -1.00
##   Residual                23.29780  4.8268
## Number of obs: 307, groups:  Escola, 7
##
## Fixed effects:
##              Estimate Std. Error t value
## (Intercept)  16.9218     0.9840  17.196
## StSE         1.0179     0.4620   2.203
## SocialMinSi  -2.7179     0.7524  -3.612
##
```

```
## Correlation of Fixed Effects:
##              (Intr) StSE
## StSE         -0.408
## SocialMinSi -0.222  0.237
## optimizer (nloptwrap) convergence code: 0 (OK)
## boundary (singular) fit: see help('isSingular')
```

Tendo en conta soamente os efectos fixos, estanse a axustar dúas rectas de regresión xerais para todas as escolas ou para a escola “media” ($\hat{u}_{0j} = \hat{u}_{1j} = 0$). Ao igual que no modelo con intercepto aleatorio, a liña para as/os menores procedentes dun grupo racial minoritario atoparase por debaixo. Isto é debido a que a recta para as/os alumnas/os con **SocialMin** = 0 terá un intercepto de $\hat{\gamma}_{00} = 16.922$, mentres que o alumnado estranxeiro sufrirá unha mingua na cualificación de $|\hat{\beta}_2| = 2.718$ puntos. Ademais, para esta escola “media” prodúcese un incremento de $\hat{\gamma}_{10} = 1.018$ puntos na nota acadada por cada incremento nunha unidade do status socio-económico, independentemente de se o alumnado procede dun grupo racial minoritario ou non. Asimesmo, o efecto do **StSE** para unha escola j estímase por $\hat{\gamma}_{10} + \hat{u}_{1j}$. Por exemplo, para a escola E4 a pendente asociada ao **StSE** estímase por $\hat{\gamma}_{10} + \hat{u}_{14} = 1.018 + 0.368$, onde a predición para o efecto aleatorio asociado á pendente da escola E4 se obtivo de novo mediante a función **ranef** de ; e de ser $\hat{u}_{14} > 0$ séguese que as dúas rectas axustadas para esta escola terán unha pendente relativamente máis pronunciada que a media. Ademais, a varianza entre escolas para esas pendentes estímase por $\hat{\sigma}_{u1}^2 = 0.065$; mentres que a varianza $\hat{\sigma}_{u0}^2 = 5.66$ do intercepto interprétase como a varianza entre escolas cando **StSE** = 0.

Xa que agora se considera a pendente aleatoria, na saída do **summary** tamén se manifesta a estimación para o coeficiente de correlación entre o intercepto e a pendente en cada grupo, $\hat{\rho}_{u01} = -1$. Agora ben, a partir desta é sinxelo obter a estimación da covarianza, $\hat{\sigma}_{u01} = -0.609$. O signo indica que escolas con intercepto grande (sobre a nota media lograda con **StSE** = 0, tanto para alumnado minoritario como non) tenden a ter pendentes menos pronunciadas e viceversa.

Para ilustrar este modelo con intercepto e pendente aleatorios, represéntanse na Figura 4.3 a recta de regresión xeral para a escola “media” (en cor negra) e mais as rectas axustadas asociadas ás escolas E3 e E4 (en azul e en rosa, respectivamente) para o caso das/os estudantes provintes do grupo racial maioritario (**SocialMin** = 0). Ademais, para a escola E4 tamén se representa a recta axustada no caso **SocialMin** = 1; mediante unha liña tamén rosa e punteada. Por considerar o efecto proporcionado pola variable **SocialMin** fixo, claramente as dúas liñas rosas teñen a mesma pendente, mentres que a correspondente ao caso da minoría racial atópase verticalmente desprazada nunha magnitude de $\hat{\beta}_2 = -2.718$; analogamente ás simbolizadas na Figura 4.2.

Finalmente, queda claro que a puntuación media é menor para alumnas/os procedentes dun

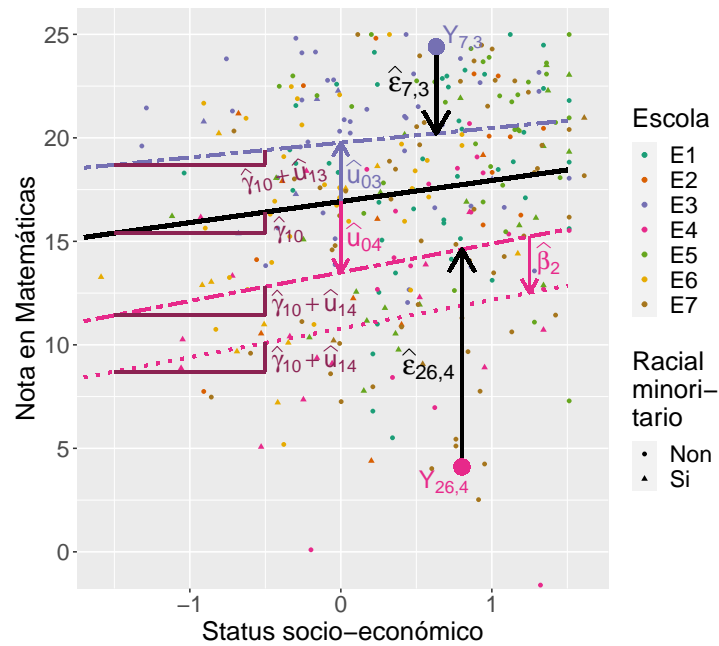


Figura 4.3: Ilustración do modelo mixto con intercepto e pendente aleatorios para a nota de Matemáticas nas 7 escolas considerando pendente fixa asociada á variable categórica **SocialMin**, xunto coa recta de regresión para a escola “media” cando **SocialMin** = 0 e para algunhas outras escolas, algúns parámetros e mais algúns residuos concernentes a ambos niveis (alumnado e escolas).

grupo racial minoritario; pero non é tan flagrante que esta diferenza deba ser a mesma para todas as escolas. Poderíase pensar que tamén o efecto da variable **SocialMin** debería ser aleatorio. Tras efectuar un test de razón de verosimilitudes desíste desta idea, posto que o p-valor que arroxa é de 0.805. Por tanto, conclúese que, considerando só o **StSE** como variable explicativa, pertencer a un grupo racial minoritario causa un efecto semellante sobre a nota de Matemáticas en todas as escolas.

Capítulo 5

Modelos mixtos con covariables relativas ao segundo nivel

Ao igual que no Capítulo 4 se introduciron variables explicativas definidas no primeiro nivel (individuos), aos modelos mixtos tamén se lles poden engadir variables relativas a niveis superiores. As variables de segundo nivel denomínanse **variables contextuais** e os seus efectos sobre os individuos do nivel 1 son os **efectos contextuais**. Tal e como ben explica Goldstein en [13], as variables contextuais poden obterse de varios xeitos diferentes:

- poden observarse directamente datos relativos ao nivel 2, como por exemplo, a variable **AmbDiscrim** (magnitude que mide o ambiente discriminatorio presente en cada escola),
- ou ben pódense resumir os datos de nivel 1 para construír uns relativos ao segundo nivel, como por exemplo a variable **MStSE**, que, para un colexio j , toma o valor da media dos status socio-económico de todas/os as/os súas/seus estudantes.

Á súa vez, estes poden extraerse dunha fonte externa (por exemplo, se o **Tamaño** das escolas se consultase en bases de datos dos EEUU), ou ben da mesma fonte da que se extraeron os datos de nivel 1 (por exemplo, se o **Tamaño** das escolas se obtivese no propio estudo presentado en [4]).

Nas seguintes liñas empregarase a nomenclatura de [10], isto é, as variables contextuais diranse **variables globais** se son específicas do segundo nivel e soamente fan referencia a unha característica propia do grupo, sen ningunha alusión ao nivel 1; mentres que se foron obtidas mediante resumos de variables de primeiro nivel diranse **variables composicionais**. Ademais, as variables de segundo nivel denotaranse por W_j^1 ; $j = 1, \dots, J$.

¹As variables asociadas ao segundo nivel non teñen subíndice i porque, por definición, o seu valor é o mesmo para todos os individuos dun mesmo grupo.

A variable explicativa de nivel 2 pode engadirse a un modelo multinivel exactamente do mesmo xeito que unha variable explicativa de nivel 1. Así, se W_j é a variable contextual e X_{ij} é a variable asociada ao nivel 1; o modelo con intercepto aleatorio (4.1) convértese en

$$Y_{ij} = \beta_{0j} + \beta_1 X_{ij} + \beta_2 W_j + \varepsilon_{ij}, \text{ con } j = 1, \dots, J \text{ e } i = 1, \dots, n_j; \quad (5.1)$$

onde de novo $\beta_{0j} = \gamma_{00} + u_{0j} \in N(\gamma_{00}, \sigma_{u0}^2)$ e $\varepsilon_{ij} \in N(0, \sigma_\varepsilon^2)$, e ademais estas variables aleatorias son independentes tanto por separado como entre si mesmas, para cada $j = 1, \dots, J$ e $i = 1, \dots, n_j$.

5.1. Variables contextuais composicionais


No caso particular de que a variable contextual sexa a media dunha variable X do primeiro nivel, entón a variable composicional non é outra que $W_j = \bar{X}_{\bullet j}$ para cada $j = 1, \dots, J$. Así, (5.1) reescríbese do seguinte xeito,

$$Y_{ij} = \beta_{0j} + \beta_1 X_{ij} + \beta_2 \bar{X}_{\bullet j} + \varepsilon_{ij}, \text{ con } j = 1, \dots, J \text{ e } i = 1, \dots, n_j; \quad (5.2)$$

onde β_1 é o efecto de dentro dos grupos de X , $\beta_1 + \beta_2$ é o efecto entre grupos e β_2 é o efecto contextual, é dicir, o efecto das medias grupais de X ($\bar{X}_{\bullet j}$) sobre Y que non se manifestaba considerando só os efectos máis elementais. Co obxectivo de que o efecto entre grupos $\beta_1 + \beta_2$ sexa un coeficiente máis do modelo, a prol de que mediante a implementación informática se estime directamente, o modelo (5.2) pode expresarse equivalentemente como


$$\begin{aligned} Y_{ij} &= \beta_{0j}^* + \beta_1^* (X_{ij} - \bar{X}_{\bullet j}) + \beta_2^* \bar{X}_{\bullet j} + \varepsilon_{ij} = \\ &= \underbrace{\gamma_{00} + \gamma_{10}(X_{ij} - \bar{X}_{\bullet j}) + \delta \bar{X}_{\bullet j}}_{\text{efectos fixos}} + \underbrace{u_{0j}}_{\text{efectos aleatorios}} + \varepsilon_{ij}; \end{aligned} \quad (5.3)$$

para cada $j = 1, \dots, J$ e $i = 1, \dots, n_j$; sen máis que facer $\beta_{0j}^* = \beta_{0j} = \gamma_{00} + u_{0j}$, $\beta_1^* = \beta_1 = \gamma_{10}$ e $\beta_2^* = \beta_1 + \beta_2$. O efecto contextual é frecuente denotalo por $\delta = \beta_1 + \beta_2$. Isto tamén se pode pensar considerando efectos aleatorios na pendente, sen máis que ter en conta o termo $\beta_{1j}^* = \beta_{1j} = \gamma_{10} + u_{1j}$ en lugar de β_1^* . Realmente, o que se fai é considerar como variables explicativas as diferenzas entre as observacións e as medias dos seus grupos ($X_{ij} - \bar{X}_{\bullet j}$), no que na literatura concernente aos modelos multinivel se soe denominar un **modelo de Cronbach**, posto que a idea de centrar a variable de primeiro nivel antes de introducila no modelo co propósito de distinguir o efecto entre grupos e o efecto dentro dos grupos foi inicialmente proposta polo experto en Psicoloxía da Educación homónimo ao modelo; segundo afirman Hox et al. en [18].

Xa que tanto o modelo presentado en (5.2) como o modelo de Cronbach (5.3) son equivalentes, ao longo deste TFG farase uso do primeiro, que é o que permite unha sintaxe máis sinxela e eficiente da función `lmer` de . En consecuencia, a estimación do efecto entre grupos ($\hat{\beta}_1 + \hat{\beta}_2$) haberá que calculala por separado.

Exemplo 5.1. Sobre a mesma base de datos coa que se viña traballando (`mates7`), no Exemplo 4.3 considerábase o modelo que explicaba a nota en Matemáticas en función do status socio-económico (variable que se vén denotando por `StSE`) con efectos aleatorios, e mais do efecto fixo concernente á pertenza a un grupo racial minoritario (variable categórica dada por `SocialMin`). Agora engadiráselle ao mesmo a variable composicional `MStSE`, que non é máis que a media dos status socio-económico das/os estudantes para cada escola e que polo tanto verifica $MStSE_j = \overline{StSE}_{\bullet j}$ para cada $j = 1, \dots, J$. Xa que o seu valor é o mesmo para todas/os as/os menores dun mesmo centro educativo, a variable `MStSE` é unha variable contextual, e por medir o status socio-económico promedio dos colexios séguese que é ademais unha variable composicional². O modelo construído virá dado pola expresión

$$NotaMates_{ij} = \underbrace{\gamma_{00} + \gamma_{10}StSE_{ij} + \beta_2SocialMin_{ij} + \beta_3MStSE_j}_{\text{efectos fixos}} + \underbrace{u_{0j} + u_{1j}StSE_{ij}}_{\text{efectos aleatorios}} + \varepsilon_{ij}; \quad (5.4)$$

de novo para todo $j = 1, \dots, J$ e $i = 1, \dots, n_j$. Así, a única diferenza coa expresión do modelo previo con intercepto e pendente aleatorios (4.3) estudado no Exemplo 4.3 é a adición do termo $\beta_3 MStSE_j$, co cal a diferenza coa Figura 4.3 do modelo (4.3) consistirá puramente nos interceptos das distintas rectas. Transcribindo esta fórmula na sintaxe requerida pola función `lmer` de , obtense o seguinte axuste do modelo (5.4):

```
matesmm4 <- lmer(NotaMates ~ StSE + SocialMin + MStSE + (1 + StSE | Escuela),
  REML = TRUE)
summary(matesmm4)

## Linear mixed model fit by REML ['lmerMod']
## Formula: NotaMates ~ StSE + SocialMin + MStSE + (1 + StSE | Escuela)
##
## REML criterion at convergence: 1843.9
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -3.8761 -0.6031  0.1107  0.7389  2.1732
##
## Random effects:
```

²Alguns autores/as non concordan en que as variables composicionais sexan un caso particular das contextuais, e tratan estas dúas situacións conxuntamente sen facer ningún tipo de distinción (como [28]). Neste traballo, pola contra, deféndese esta clasificación; ao igual que deciden facelo moitas/os outras/os autoras/es salientábeis na literatura dos modelos multinivel (como [18], [10] ou [21]). Á hora de analizar modelos mixtos non é importante que clasificación se esté a empregar, é soamente unha cuestión conceptual; esclarece a que nivel pertence unha certa variable.

```
## Groups      Name      Variance Std.Dev. Corr
## Escola      (Intercept) 6.69176 2.5868
##              StSE      0.05674 0.2382 -1.00
## Residual              23.30565 4.8276
## Number of obs: 307, groups: Escola, 7
##
## Fixed effects:
##              Estimate Std. Error t value
## (Intercept) 17.0893      1.6662 10.256
## StSE        1.0275      0.4670 2.200
## SocialMinSi -2.7128      0.7537 -3.599
## MStSE       -0.4411      3.3888 -0.130
##
## Correlation of Fixed Effects:
##              (Intr) StSE   SclMnS
## StSE         -0.124
## SocialMinSi -0.102 0.241
## MStSE       -0.773 -0.150 -0.037
## optimizer (nloptwrap) convergence code: 0 (OK)
## boundary (singular) fit: see help('isSingular')
```

O coeficiente fixo asociado á variable **MStSE** estímase por $\hat{\beta}_3 = -0.441$, co cal un neno ou nena, procedente dun grupo racial minoritario ou non e cun status socio-económico fixado de antemán; ao incrementarse nun punto o promedio dos status socio-económico das/os demais alumnas/os da súa escola, espérase que experimente unha diminución na súa nota de Matemáticas de case medio punto. En definitiva, se un membro do alumnado se queda por detrás dos seus compañeiros e compañeiras; ben socialmente ou ben economicamente, non só ficará inamovible no eido da Educación mentres o resto perfecciona as súas notas, senón que incluso sufrirá unha mingua nos seus resultados; en particular, na materia de Matemáticas.

Na Figura 5.1 ilústrase a natureza do modelo (5.4). Como xa se adiantaba, é moi semellante á Figura 4.3; a diferenza reside enteiramente nos interceptos asociados ás distintas escolas. Mentres que no Exemplo 4.3 as estimacións dos efectos aleatorios do intercepto \hat{u}_{0j} con $j = 1, \dots, J$ se interpretaban como a desviación do intercepto para a escola j -ésima con respecto ao intercepto medio (o da recta de cor negra, entendéndoo como o intercepto habitual cando **StSE** = 0); neste contexto é preciso ter en conta a estimación do efecto fixo asociado á variable contextual composicional **MStSE** á hora de estudar os interceptos dos distintos colexios, posto que para obter o intercepto dunha escola j tamén haberá que sumarlle agora a \hat{u}_{0j} o termo $\hat{\beta}_3 \text{MStSE}_j = \hat{\beta}_3 \overline{\text{StSE}}_{\bullet j}$.

³Para obter p-valores asociados aos tests de razón de verosimilitudes máis precisos pode empregarse o denominado procedemento *bootstrap* paramétrico, sobre o cal se pode atopar máis información en [27, pp. 294–301].

senso rexeitar a hipótese nula, isto é, o promedio dos status socio-económico do alumnado dunha escola non inflúe de xeito significativo na nota dun/dunha estudante calquera. Consecuentemente, a estimación das varianzas tamén aumenta con respecto ás obtidas no modelo (4.3) axustado no Exemplo 4.3 (excepto a concernente á pendente aleatoria asociada á variable `StSE`, $\hat{\sigma}_{u1} = 0.057$).

5.2. Variables contextuais globais

O seguinte paso natural no incremento da complexidade do modelo (5.1) é permitir que a pendente asociada a X varíe entre os diferentes grupos relativos ao segundo nivel; co cal o modelo con intercepto e pendente aleatorios tórnase agora en


$$Y_{ij} = \beta_{0j} + \beta_{1j}X_{ij} + \beta_2W_j + \varepsilon_{ij}, \text{ con } j = 1, \dots, J \text{ e } i = 1, \dots, n_j; \quad (5.5)$$

onde se seguirán a supor as mesmas hipóteses sobre a independencia que até agora, é dicir, de novo $\varepsilon_{ij} \in N(0, \sigma_\varepsilon^2)$ son independentes e ademais $\beta_{0j} = \gamma_{00} + u_{0j} \in N(\gamma_{00}, \sigma_{u0}^2)$ e $\beta_{1j} = \gamma_{10} + u_{1j} \in N(\gamma_{10}, \sigma_{u1}^2)$, con $j = 1, \dots, J$, son variables aleatorias independentes separadamente (e independentes dos erros). Así, dentro de cada grupo pode existir correlación entre os efectos aleatorios do intercepto e da pendente, sendo σ_{u01} a súa covarianza. De aquí en diante, o coeficiente fixo asociado á variable global W escribírase tamén $\beta_2 = \gamma_{01}$. Nótese que o modelo (5.5) pode escribirse como:

$$Y_{ij} = \underbrace{\gamma_{00} + \gamma_{10}X_{ij} + \gamma_{01}W_j}_{\text{efectos fixos}} + \underbrace{u_{0j} + u_{1j}X_{ij}}_{\text{efectos aleatorios}} + \varepsilon_{ij}.$$

Exemplo 5.2. Partindo do modelo (4.3) axustado na Sección 4.2, xa que o modelo axustado previamente e que lémbrese incluía como covariable a variable contextual composicional `MStSE` non resultou ser significativo; engadiránselle agora variables contextuais de tipo global a prol de construír un novo modelo que axuste as notas en Matemáticas de xeito máis eficiente. Na base de datos `mates7` dispónse de 5 variables globais; aínda que as variables `Sector` (carácter público ou privado da escola) e `Particip` (proporción de estudantes da escola que participan no estudo) non resultan útiles, posto que as 7 escolas que se están a ter en consideración son todas católicas e tiveron unha participación unánime no estudo [4]. Pódense empregar para axustar o novo modelo, polo tanto, as variables `Tamaño`, `AmbDiscrim` e `MaioriaMin`; que lémbrese reflectían o número de estudantes, unha medida do ambiente discriminatorio e se a cantidade das/os matriculadas/os membros dun grupo racial minoritario supera o 40 % do total da escola, respectivamente.

Nótese que a variable `MaioriaMin`, tal e como se describe, ben podería tratarse dunha variable composicional do mesmo estilo que a do Exemplo 5.1. Isto é debido a que `MaioriaMin` podería construírse para as sete escolas da base de datos `mates7` simplemente calculando proporcións sobre a variable de primeiro nivel `SocialMin` (que indicaba se un/unha estudante era membro

dun grupo racial minoritario). Pola contra, isto só pode facerse no caso particular das escolas E1, E2, ... , E7 debido a que tiveron unha participación unánime no estudo presentado en [4]; mais con todos os outros centros educativos presentes na base de datos orixinal isto non é posible, posto que non se ten información acerca de todas/os as/os estudantes; indicio de que tal variable se extraeu nun primeiro momento dunha fonte externa ao estudo académico descrito en [4]. Polo tanto, neste traballo considerárase que **MaioriaMin** é unha variable contextual global. De novo salientase que á hora de axustar un modelo multinivel, a clasificación que se empregue non ten importancia algunha, polo que se pola contra **MaioriaMin** se considerase como unha variable composicional; as conclusións que se acadarían non variarían no máis mínimo coas obtidas a continuación. Deseguido resumíranse as conclusións obtidas grazas ao entorno estatístico . O código empregado pode atoparse no Anexo A.8.2.

Por unha banda, considerando o efecto fixo do **Tamaño** ao engadilo ao modelo (4.3) obtense que a estimación do coeficiente asociado a este resulta ser case nulo. Ademais, tras a realización dun test de razón de verosimilitudes compróbase que non resulta ser significativo. Conclúese entón que o tamaño dunha escola non inflúe de xeito significativo na nota acadada por un/unha estudante na materia de Matemáticas. Unha explicación a dito resultado podería ser a dada por [14] en relación ao tamaño das clases. Ao contrario do que se soe pensar, que as clases sexan máis numerosas non sempre implica un menor rendemento académico. Isto ocorre tan só cando o número de estudantes con malas notas nesa mesma clase supera un certo umbral; e no caso particular dos 7 centros de **mates7**, as porcentaxes de alumnas/os con notas menores que 5 puntos (por exemplo) son, respectivamente:

```
## Escola
##   E1   E2   E3   E4   E5   E6   E7
## 0.00 4.76 0.00 8.82 0.00 0.00 6.78
```

Ningunha porcentaxe supera o umbral do 10 % (por exemplo), isto podería ser a xustificación de que o tamaño das escolas non resulte significativo.

Por outra banda, engadindo agora o efecto da variable global **AmbDiscrim** ao modelo (4.3), tal e como era de esperar a estimación do coeficiente resulta negativa; isto é, o aumento do ambiente discriminatorio nunha certa escola ten efectos perniciosos sobre as notas das/os súas/seus estudantes. Pola contra, novamente este coeficiente tampouco resulta significativo.

A variable que resta por engadir ao modelo (4.3) é **MaioriaMin**, variable dicotómica que se codifica por 1 para as escolas con máis do 40 % do alumnado membro dun grupo racial minoritario, e por 0 no caso contrario. No Exemplo 4.2 viuse que o simple feito de pertencer a un grupo racial minoritario auguraba unha mingua importante na nota de Matemáticas. Pola contra, ao axustar un novo modelo tendo en conta a variable **MaioriaMin**, para as escolas con máis do

40 % do alumnado procedente dun grupo racial minoritario augúrase un incremento do intercepto de $\hat{\gamma}_{01} = 1.175$ puntos na nota de Matemáticas (entendéndoo cando $\text{StSE} = 0$). Isto non goza de senso ningún e pode deberse á estrutura dos datos; posto que só a escola E2 (a de menor tamaño) verifica que $\text{MaioriaMin} = 1$, e estes datos poderían resultar insuficientes para extraer conclusións axeitadas. De feito, engadir a variable MaioriaMin resulta máis significativo que a anexión das dúas anteriores; aínda que non o suficiente, pois o test de razón de verosimilitudes entre o modelo considerando MaioriaMin e o previo (4.3) fornece un p-valor de 0.094; maior que o nivel de significación $\alpha = 0.001$ considerado neste traballo; e ademais lémbrese que segundo [27] ou [9], o p-valor de 0.094 foi subestimado e o real aínda podería ser aínda máis grande.

5.3. Interacción entre niveis

Afondando aínda máis no terreo dos modelos multinivel, poderíase permitir que o efecto dunha covariable dependa do valor doutra variable explicativa. No caso particular de que unha variable estea asociada ao nivel 1 e a outra ao nivel 2 isto denomínase **interacción entre niveis**. Deste xeito, partindo do previo modelo (5.5) con intercepto e pendente aleatorios no que se introduce información sobre algunha característica do grupo a través dunha variable de segundo nivel; e considerando interacción entre niveis, estase a construír un modelo cuxas formulacións xerárquica e combinada son, para todo $i = 1, \dots, n_j$ e para todo $j = 1, \dots, J$:

(i) Formulación xerárquica:

$$\begin{aligned} \text{Nivel 1 : } Y_{ij} &= \beta_{0j} + \beta_{1j}X_{ij} + \varepsilon_{ij}. \\ \text{Nivel 2 : } \begin{cases} \beta_{0j} = \gamma_{00} + \gamma_{01}W_j + u_{0j}, & \text{con } u_{0j} \in N(0, \sigma_{u0}^2), \\ \beta_{1j} = \gamma_{10} + \gamma_{11}W_j + u_{1j}, & \text{con } u_{1j} \in N(0, \sigma_{u1}^2). \end{cases} \end{aligned}$$

(ii) Formulación conxunta:
$$Y_{ij} = \underbrace{\gamma_{00} + \gamma_{10}X_{ij} + \gamma_{01}W_j + \gamma_{11}W_jX_{ij}}_{\text{efectos fixos}} + \underbrace{u_{0j} + u_{1j}X_{ij}}_{\text{efectos aleatorios}} + \varepsilon_{ij}.$$

De novo, suponse que os erros son independentes dos efectos aleatorios do intercepto e pendente, pero estes últimos poden estar correlados. Séguense imponendo tamén as hipóteses de normalidade sobre os erros de ambos niveis. Nótese que a parte aleatoria coincide coa do modelo con intercepto e pendente aleatorios (4.2), sen considerar a variable de segundo nivel W . A principal característica deste modelo e asemade a diferenza co modelo sen interacción reside na inserción do termo W_jX_{ij} , que se pode entender como a interacción entre variables de ambos niveis; sendo γ_{11} o coeficiente de tal interacción.

Finalmente, no caso de ser a covariable W unha variable dicotómica, a interpretación dos coeficientes é semellante á feita no Exemplo 4.2 para SocialMin .

Capítulo 6


Conclusións

Ao longo deste traballo viuse que os modelos de regresión clásicos, como o modelo de regresión linear ou os modelos de análise da varianza e covarianza non son suficientes cando se está a traballar con datos aniñados xerarquicamente, os cales son moi frecuentes no eido da Educación, a Medicina ou as Ciencias Medioambientais. Cando os individuos forman grupos, é obvio pensar que os individuos clasificados nun mesmo grupo tenderán a ter un comportamento máis semellante que uns individuos calesquera de grupos diferentes e polo tanto con menos información en común. Ademais, habitualmente o interesante non son os grupos presentes no conxunto de datos, senón que se pretende aplicar técnicas da Inferencia Estatística sobre unha poboación máis grande de grupos. É aquí onde xorden os denominados modelos mixtos, modelos multinivel ou modelos de efectos aleatorios. En particular, neste traballo tratáronse os modelos mixtos lineares con resposta continua e púxose de manifesto a súa utilidade para estudar bases de datos cunha estrutura xerárquica de dous niveis, onde os individuos se atopan no primeiro nivel e están aniñados en grupos no segundo nivel, mediante a incorporación de efectos aleatorios.

Cabe salientar que os modelos mixtos lineares vistos neste traballo poderían ampliarse a formas máis complexas, como os modelos mixtos lineares xeralizados (denotados habitualmente como *GLMM*), que xorden cando a variable resposta Y posee certas restricións, como por exemplo, que se trate dunha variable de recuento. Estes modelos son a continuación natural dos modelos mixtos lineares, e toda a información concernente aos modelos *GLMM* poden consultarse na extensa obra de Demidenko [6].


Outra posible extensión dos modelos mixtos lineares sería asumir que en cada grupo non se ten unha recta, senón un modelo polinómico, por exemplo. Deste xeito, daríase máis flexibilidade para modelar o comportamento da variable resposta en cada grupo de interese.

Ao longo deste traballo ilustrouse a utilidade dos modelos mixtos lineares con resposta continua no eido da Educación mediante a base de datos *mates7* creada para esta ocasión e mais

a ferramenta estatística . Lémbrese que o código desenvolto ao longo deste traballo está dispoñible no repositorio *Modelos_Mixtos_con_R* de *GitHub* e é de carácter enteiramente reproducible. Isto serviu para estudar que características e situacións inflúen na nota de Matemáticas dun/dunha alumno/a calquera, asemade se construíu un modelo mixto para tratar de predecir as notas das/os mesmas/os. Obtívose que a escola á que acode un/unha alumno/a inflúe enormemente na nota acadada na materia de Matemáticas por este/a, así como a situación socioeconómica da súa familia. Ademais, estudouse como inflúe a integridade social das/os estudantes na nota dende un punto de vista racial; obtendo que as/os alumnas/os procedentes dun grupo racial minoritario acadan en promedio unha nota significativamente menor que as/os outras/os estudantes en xeral. Por outra banda, tamén se tocaron aspectos concernentes á política no eido da Educación. Ao contrario do que se soe crer, tanto o tamaño das escolas como o número de alumnas/os por clase non inflúen na nota das/os estudantes. Isto só ocorre cando a porcentaxe de menores de baixo rendemento na mesma clase ou na mesma escola supera certo umbral. É por isto que os diferentes gobernos do Mundo non deberían invertir os seus bens en diminuír o tamaño dos centros educativos ou o ratio de alumnas/os por clase, senón en mellorar a calidade da Educación que se ofrece nas/os mesmas/os.

Anexo A

Código de R

Neste Anexo reproducirase parte do código de  empregado ao longo de todo o traballo. A súa integridade pode consultarse no repositorio *Modelos_Mixtos_con_R* de *GitHub*, tanto o relativo á construción dos diversos modelos que se estudaron como o concernente a todas as figuras ilustradas, todas programadas empregando o paquete `ggplot2`. A sintaxe empregada á hora de redactar o código é a defendida por Hadley Wickham en [31]. Tamén merecen especial recoñecemento [32] e [5]; posto que moitas das seguintes liñas de código puideron elaborarse grazas ás ideas proporcionadas nestas obras. Por último, o código dispónse seguindo o formato que emprega o paquete `knitr`, o cal tamén se empregou para xerar este documento e sobre o cal se pode atopar máis información na obra orixinal do autor do paquete Yihui [33]. Todos estes libros son de código aberto e atópanse dispoñibles en *GitHub*.

A.1. Creación da base de datos mates

```
pkgs <- c('lme4', 'nlme', 'lmtest', 'ggplot2', 'car', 'dplyr', 'RColorBrewer',
  'viridis', 'gtools', 'patchwork', 'knitr', 'ggthemes', 'lattice', 'latex2exp')
install.packages(setdiff(pkgs, installed.packages()[,"Package"]),
  dependencies = TRUE)

library(lme4); library(nlme); library(lmtest); library(ggplot2); library(car)
library(dplyr); library(RColorBrewer); library(viridis); library(gtools)
library(patchwork); library(knitr); library(ggthemes); library(lattice)
library(latex2exp)

mates <- merge(MathAchieve, MathAchSchool, by = "School")
```

```

mates <- mates[, -12]
names(mates) <- c("Escola", "SocialMin", "Sexo", "StSE", "NotaMates",
  "MStSE", "Tamaño", "Sector", "Particip", "AmbDiscrim", "MaioriaMin")
levels(mates$Sector) <- c("Publica", "Catolica")
levels(mates$SocialMin) <- c("Non", "Si")
levels(mates$Sexo) <- c("Home", "Muller")
mates$Escola <- factor(mates$Escola, levels = sort(levels(mates$Escola)))
levels(mates$Escola) <- c(1:160)
attach(mates)

### Validación da normalidade por grupos de escola
table(Escola)
length(levels(Escola)) #160 escolas
pval <- rep(0, length(levels(Escola)))
for (i in 1:length(levels(Escola))){
  aux <- NotaMates[Escola == levels(Escola)[i]]
  pval[i] <- shapiro.test(aux)$p.value
}
sum(pval < 0.001) #só 5 escolas

#Validación da homoxeneidade de varianzas
leveneTest(lm(NotaMates ~ as.vector(Escola)), center = mean)

```

A.2. Introducción

```

set.seed(123)
ind <- sample(1:dim(mates)[1], 40) #40 alumnas/os aleatorios
mls0 <- lm(NotaMates[ind] ~ StSE[ind]) #modelo linear simple
eps <- residuals(mls0)

```

A.2.1. Figura 1

```

ggplot(data.frame(cbind(StSE[ind], NotaMates[ind])), aes(x = StSE[ind],
  y = NotaMates[ind])) +

```

```

geom_point() +
xlab("Status socio-económico de 40 nenas/os") +
ylab("Nota en Matemáticas de 40 nenas/os") +
geom_smooth(method='lm', se = FALSE, col = "darkcyan") +
geom_segment(x = StSE[ind][12], y = NotaMates[ind][12], xend = StSE[ind][12],
  yend = NotaMates[ind][12] - eps[12], color = "tomato3", arrow = arrow(),
  size = 1) +
geom_point(aes(x = StSE[ind][12], y = NotaMates[ind][12]), col = "purple4",
  size = 4) +
annotate("text", x = StSE[ind][12] + .125, y = NotaMates[ind][12] - eps[12]/2 +
  1/3, parse = TRUE, label = expression(widehat(epsilon)[i]), size = 10,
  col = "tomato3", cex = 2) +
annotate("text", x = StSE[ind][12] - .15, y = NotaMates[ind][12], parse = TRUE,
  label = expression(Y[i]), size = 9, col = "purple4", cex = 2) +
annotate("text", x = 0.1, y = mls0$coefficients[[1]] - 2, parse = TRUE,
  label = expression(paste(widehat(beta)[0], + widehat(beta)[1], "x")),
  size = 8, col = "darkcyan") +
theme(
  axis.title.x = element_text(size = 20),
  axis.text.x = element_text(size = 18),
  axis.title.y = element_text(size = 20),
  axis.text.y = element_text(size = 18))

```

A.3. Creación da base de datos mates7

```

escolas_particip100 <- which(MathAchSchool$PRACAD == 1)
alum_escolas_particip100 <- which(Escola %in% escolas_particip100)
length(alum_escolas_particip100) #307 alumnos de 7 escolas

### Validación da normalidade por grupos de escola
pval <- rep(0, length(escolas_particip100))
alpha <- 0.001
for (i in 1:length(escolas_particip100)){
  aux <- NotaMates[Escola == escolas_particip100[i]]
  pval[i] <- shapiro.test(aux)$p.value
}

```

```
sum(pval < alpha)  #resultan significativamente normais

mates7 <- mates[alum_escolas_particip100, ] #evitamos nesgos de selección
dim(mates7) # 307 alumnas/os e 11 variables
mates7$Escola <- factor(mates7$Escola, levels = escolas_particip100)
levels(mates7$Escola) <- paste("E", 1:7, sep = "") #Nome das 7 escolas
attach(mates7)
```

A.4. ANOVA

```
anova_mates7 <- lm(NotaMates ~ Escola - 1)
summary(anova_mates7)
mu_local <- numeric(7)
for (i in 1:length(levels(mates7$Escola))){
  mu_local[i] <- mean(NotaMates[Escola == levels(mates7$Escola)[i]])
} #vector de medias locais
```

A.4.1. Figura 1.1

```
ggplot(mates7, aes(x = Escola, y = NotaMates, color = Escola)) +
  geom_boxplot() +
  labs(x = "Escolas", y = "Notas en Matemáticas", color = "Escolas") +
  scale_color_brewer(palette = "Dark2") +
  geom_segment(x = 0.625, y = mu_local[1], xend = 1.375, yend = mu_local[1],
    linetype = "dotted", col = "firebrick4") +
  geom_segment(x = 1.625, y = mu_local[2], xend = 2.375, yend = mu_local[2],
    linetype = "dotted", col = "firebrick4") +
  geom_segment(x = 2.625, y = mu_local[3], xend = 3.375, yend = mu_local[3],
    linetype = "dotted", col = "firebrick4") +
  geom_segment(x = 3.625, y = mu_local[4], xend = 4.375, yend = mu_local[4],
    linetype = "dotted", col = "firebrick4") +
  geom_segment(x = 4.625, y = mu_local[5], xend = 5.375, yend = mu_local[5],
    linetype = "dotted", col = "firebrick4") +
  geom_segment(x = 5.625, y = mu_local[6], xend = 6.375, yend = mu_local[6],
```

```

  linetype = "dotted", col = "firebrick4") +
geom_segment(x = 6.625, y = mu_local[7], xend = 7.375, yend = mu_local[7],
  linetype = "dotted", col = "firebrick4") +
geom_hline(yintercept = mean(NotaMates), linetype = "dashed", col = "navy") +
theme(
  legend.position = "none",
  axis.title.x = element_text(size = 18),
  axis.text.x = element_text(size = 16),
  axis.title.y = element_text(size = 18),
  axis.text.y = element_text(size = 16))

```

A.4.2. Función Bonferroni_taboa

```

Bonferroni_taboa <- function(varresposta, varfactor, alpha_sen_CB = .05,
  ndixitos = NULL){
  if (!is.numeric(varresposta)){
    stop("varresposta debe ser da clase numeric, proba con
      as.numeric(varresposta).")
  }
  if (!is.factor(varfactor)){
    stop("varfactor debe ser da clase factor, proba con factor(varfactor).")
  }
  if (!is.numeric(alpha_sen_CB)){
    stop("0 nivel de significación sen correccion de Bonferroni alpha_sen_CB
      ten que ser un número.")
  }
  if (!is.numeric(ndixitos) & !is.null(ndixitos)){
    stop("0 número de díxitos \"ndixitos\" ten que ser un número natural.")
  }
  if (is.numeric(ndixitos)){
    if (ndixitos < 0){
      stop("0 número de díxitos \"ndixitos\" ten que ser un número natural.")
    }
  }
  n = length(varresposta)
  J = length(levels(varfactor))

```

```

nj = table(varfactor)
desvtip = sqrt(deviance(lm(varresposta ~ varfactor)) / (n-J))
comb <- gtools::combinations(J, 2, 1:J) #num_comb = J*(J-1)/2
alpha_con_CB <- alpha_sen_CB / (2 * nrow(comb))
ct = qt(1 - alpha_con_CB, n - J)
Bonferroni = data.frame(dif = rep(0, nrow(comb)), inf = rep(0, nrow(comb)),
  sup = rep(0, nrow(comb)))
for (k in 1:nrow(comb)){
  rownames(Bonferroni)[k] <- paste(comb[k, 1], comb[k, 2], sep = "-")
}
mu_local <- numeric(7)
for (i in 1:length(levels(varfactor))){
  mu_local[i] <- mean(varresposta[varfactor == levels(varfactor)[i]])
}
for (k in 1:nrow(comb)){
  Bonferroni$dif[k] = mu_local[comb[k, 2]] - mu_local[comb[k, 1]]
  Bonferroni$inf[k] = mu_local[comb[k, 2]] - mu_local[comb[k, 1]] -
    ct * desvtip * sqrt(1/nj[comb[k, 2]] + 1/nj[comb[k, 1]])
  Bonferroni$sup[k] = mu_local[comb[k, 2]] - mu_local[comb[k, 1]] +
    ct * desvtip * sqrt(1/nj[comb[k, 2]] + 1/nj[comb[k, 1]])
}
cat("##### Intervalos de confianza simultáneos", "\n", " # con corrección de
  Bonferroni: #####", "\n")
if (!is.null(ndixitos)){
  print(round(Bonferroni, ndixitos))
}
else{print(Bonferroni)}
cat("\n", "##### Grupos cuxas medias difiren: #####")
difiren <- numeric(nrow(comb))
nondifiren <- numeric(nrow(comb))
for (k in 1:nrow(comb)){
  if(sign(Bonferroni[k, 2]) == sign(Bonferroni[k, 3])){
    cat("\n", "- Os grupos", rownames(Bonferroni)[k], "difiren nas súas
      medias")
    difiren[k] <- rownames(Bonferroni)[k]
  }
  else{

```

```

    nondifiren[k] <- rownames(Bonferroni)[k]
  }
}
cat("\n\n", "## Nivel de significación empregado\n    coa corrección de
    Bonferroni: ", alpha_sen_CB, "/", nrow(comb), "=", format(alpha_con_CB,
    digits = ndixitos + 1, scientific = TRUE))
difiren_ch <- difiren[nondifiren == 0]
nondifiren_ch <- nondifiren[difiren == 0]
saida <- list("intervsim" = Bonferroni, "difiren" = difiren_ch, "nondifiren" =
    nondifiren_ch, "alpha_CB" = alpha_con_CB)
return(invisible(saida))
}

```

A.4.3. Test F e contrastes pareados

```

anova(lm(NotaMates ~ Escola)) #táboa ANOVA
pf(10.475, df1 = 6, df2 = 300, lower.tail = F) #p-valor test F
source("Bonferroni_taboa.R") #cargado da función
Bonferroni_taboa(NotaMates, Escola, alpha_sen_CB = 0.001, ndixitos = 3)

```

A.5. ANCOVA

```

ancova1 <- lm(NotaMates ~ StSE + Escola) #ANCOVA sen interacción
ancova2 <- lm(NotaMates ~ StSE * Escola) #ANCOVA con interacción
anova(ancova1, ancova2) #test F, interacción non significativa
#Contraste sobre o efecto dos grupos e de StSE:
mod_StSE <- lm(NotaMates ~ StSE) #sen grupos
mod_g <- lm(NotaMates ~ Escola) #sen StSE
anova(ancova1, mod_StSE) #efecto das escolas significativo
anova(ancova1, mod_g) #efecto do StSE significativo

```

A.5.1. Figura 1.2

```

coefs_anc1 <- ancova1$coefficients #estimaciones ANCOVA sen interacción
dark_2 <- brewer.pal(n = 7, name = "Dark2") #paleta de cores

rang <- data.frame(i = numeric(7), f = numeric(7))
for (k in 1:length(levels(Escola))){
  rang[k,] <- range(StSE[Escola == levels(Escola)[k]])
} #Función para os rangos das rectas
recta <- function(x = 0, grupo = 1){
  if (!grupo %in% 1:7){
    stop("Tal grupo non existe.")
  }
  if (grupo == 1){
    as.numeric(coefs_anc1[1]) + as.numeric(coefs_anc1[2]) * x
  }
  else{
    as.numeric(coefs_anc1[1]) + as.numeric(coefs_anc1[grupo + 1]) +
      as.numeric(coefs_ancova1[2]) * x
  }
} #Función para as ecuacións das rectas

ancova_si <- ggplot(mates7, aes(x = StSE, y = NotaMates, color = Escola)) +
  geom_point() +
  labs(x = "Status socio-económico", y = "Notas en Matemáticas",
       color = "Escola") +
  coord_cartesian(ylim = c(-.5, 25)) +
  scale_color_brewer(palette = "Dark2") +
  geom_segment(x = rang[1,1], xend = rang[1,2], y = recta(rang[1,1]), yend =
    recta(rang[1,2]), col = dark_2[1], lwd = 1.25) +
  geom_segment(x = rang[2,1], xend = rang[2,2], y = recta(rang[2,1], 2), yend =
    recta(rang[2,2], 2), col = dark_2[2], lwd = 1.25) +
  geom_segment(x = rang[3,1], xend = rang[3,2], y = recta(rang[3,1], 3), yend =
    recta(rang[3,2], 3), col = dark_2[3], lwd = 1.25) +
  geom_segment(x = rang[4,1], xend = rang[4,2], y = recta(rang[4,1], 4), yend =
    recta(rang[4,2], 4), col = dark_2[4], lwd = 1.25) +
  geom_segment(x = rang[5,1], xend = rang[5,2], y = recta(rang[5,1], 5), yend =
    recta(rang[5,2], 5), col = dark_2[5], lwd = 1.25) +
  geom_segment(x = rang[6,1], xend = rang[6,2], y = recta(rang[6,1], 6), yend =

```



```

    recta(rang[6,2], 6), col = dark_2[6], lwd = 1.25) +
geom_segment(x = rang[7,1], xend = rang[7,2], y = recta(rang[7,1], 7), yend =
    recta(rang[7,2], 7), col = dark_2[7], lwd = 1.25) +
theme(
  legend.position = "none",
  axis.title.x = element_text(size = 16),
  axis.text.x = element_text(size = 14),
  axis.title.y = element_text(size = 16),
  axis.text.y = element_text(size = 14))

ancova_ci <- ggplot(mates7, aes(x = StSE, y = NotaMates, color = Escola)) +
  geom_point() +
  labs(x = "Status socio-económico", color = "Escola") +
  coord_cartesian(ylim = c(-.5, 25) ) +
  scale_color_brewer(palette = "Dark2") +
  geom_smooth(method = "lm", se = FALSE, lwd = 1.25) +
  theme(
    legend.title = element_text(size = 16),
    legend.text = element_text(size = 14),
    axis.title.x = element_text(size = 16),
    axis.text.x = element_text(size = 14),
    axis.title.y = element_blank(),
    axis.text.y = element_text(size = 14))

ancova_si + ancova_ci #sintaxe válida grazas ao paquete de R patchwork

```

A.6. RANOVA

```

ranovamates <- lmer(NotaMates ~ (1 | Escola), data = mates7, REML = FALSE)
summary(ranovamates)
sigma2_eps_ran = summary(ranovamates)$sigma^2
sigma2_u_ran = as.data.frame(summary(ranovamates)$varcor)[1,4]
VT_ran = sigma2_u_ran + sigma2_eps_ran
VPC_ran = sigma2_u_ran / VT_ran

```

A.6.1. VARCOMP e contraste sobre os efectos das escolas

```
vcmates <- lm(NotaMates ~ 1, data = mates7)
summary(vcmates)
#Contraste sobre o efecto da escola
mu_t <- vcmates$coef[[1]] #estimación media global (a da mostra)
LR <- -2 * logLik(vcmates)[1] - (-2 * logLik(ranovamates))[1]
1 - pchisq(LR, df = 1) #p-valor
```

A.6.2. Figura 3.5

```
x <- seq(0, 45, length = 1000)
y <- dchisq(x, df = 1)
df <- data.frame(x,y)
ggplot(df, aes(x = x, y = y)) +
  geom_line(color = "orange", size = 1.25) +
  xlab(" ") +
  coord_cartesian(ylim = c(0, .15), clip = "off") +
  geom_hline(yintercept = 0, col= "green2") +
  geom_vline(xintercept = 0, col= "green2") +
  geom_segment(x = LR, y = -0.004, xend = LR, yend = -.0115, col= "tomato3",
    linetype = "solid", size = 1.5) +
  annotate("text", x = 6, y = .1, parse = TRUE, label = expression(chi[1]^2),
    col = "orange", size = 17) +
  annotate("text", x = LR, y = -0.017, parse = TRUE, label = expression(LR[obs]),
    col = "tomato3", size = 10) +
  annotate("text", x = LR + 6, y = .007, parse = TRUE, label =
    expression(1.96e-09), col = "lightsalmon", size = 8) +
  theme_classic() +
  theme(
    axis.title.x = element_text(size = 10),
    axis.text.x = element_text(size = 17),
    axis.title.y = element_blank(),
    axis.text.y = element_text(size = 17))
```

A.6.3. Conxunto de datos u0df

```

u0 <- ranef(ranovamates, postVar = TRUE)
u0se <- sqrt(attr(u0[[1]], "postVar")[1, , ])
escolaidentif <- rownames(u0[[1]])
u0df <- cbind(escolaidentif, u0[[1]], u0se)
colnames(u0df) <- c("escolaidentif", "u0", "u0se")
u0df <- u0df[order(u0df$u0), ]
u0df <- cbind(u0df, c(1:dim(u0df)[1]))
colnames(u0df)[4] <- "u0pos"
for (k in 1:nrow(u0df)){
  numero <- as.numeric(strsplit(u0df$escolaidentif[k], split = "")[[1]][2])
  u0df$escolaidentif[k] <- numero
}
u0df <- u0df[order(u0df$escolaidentif), ] #reordenación
estloc <- summary(ranovamates)$coef[1] + u0df$u0
u0df$estloc <- estloc
u0df$escolaidentif <- paste("E", 1:7, sep = "")
naive_res <- mu_local - summary(ranovamates)$coef[1]
u0df$naive_res <- naive_res
shrink <- numeric(7)
for (i in 1:length(levels(mates7$Escola))){
  shrink[i] <- u0[[1]][i, 1] / naive_res[i]
}
u0df$shrinkage <- shrink
kable(u0df, format = "pipe", digits = 3, row.names = FALSE)

```

A.6.4. Figura 3.3: Gráfico de eiruga

```

ggplot(u0df, aes(x = u0pos, y = u0)) +
  geom_segment(x = u0df$u0pos, y = u0df$u0 - 1.96 * u0df$u0se, xend = u0df$u0pos,
    yend = u0df$u0 + 1.96 * u0df$u0se, size = 1, col = "darkcyan") +
  geom_point(size = 3, col = "tomato3") +
  xlab("") +
  ylab("Residuos a nivel de Escola") +
  scale_x_continuous(breaks = 1:7) +
  geom_hline(yintercept = 0, size = 1) +
  annotate("text", x = u0df$u0pos + 0.3, y = u0df$u0, parse = TRUE, label =

```

```

paste("E", 1:7, sep = ""), col = "tomato3", size = 7) +
theme(
  axis.title.x = element_text(size = 19),
  axis.text.x = element_text(size = 17),
  axis.title.y = element_text(size = 19),
  axis.text.y = element_text(size = 17))

```

A.6.5. Figura 3.2

O código de  relativo á Figura 3.4 concernente ao modelo dun só nivel para a media é análogo ao disposto deseguido.

```

mates7$Estudante <- 1:nrow(mates7)
attach(mates7)
eps_ran <- residuals(ranovamates)

ggplot(mates7, aes(x = Estudante, y = NotaMates, color = Escola)) +
  geom_point(size = 1.25) +
  coord_cartesian(xlim = c(0, 350), ylim = c(1, 25)) +
  xlab("Indentificador de estudante") +
  ylab("Nota en Matemáticas") +
  scale_color_brewer(palette = "Dark2") +
  #Media global
  geom_segment(x = -7, y = summary(ranovamates)$coef[1], xend = 317,
    yend = summary(ranovamates)$coef[1], size = 1.35, col = 1) +
  #Medias locais
  geom_segment(x = -7, y = estloc[3], xend = 294, yend = estloc[3], lwd = 1,
    linetype = "twodash", color = dark_2[3]) +
  geom_segment(x = -7, y = estloc[4], xend = 294, yend = estloc[4], lwd = 1,
    linetype = "twodash", color = dark_2[4]) +
  geom_segment(x = 0, y = estloc[5], xend = 294, yend = estloc[5], lwd = 1,
    linetype = "twodash", color = dark_2[5]) +
  geom_segment(x = 0, y = estloc[7], xend = 294, yend = estloc[7], lwd = 1,
    linetype = "twodash", color = dark_2[7]) +
  #Frechas
  geom_segment(x = Estudante[72], y = NotaMates[72], xend = Estudante[72],
    yend = NotaMates[72] - eps_ran[72], color = 1, arrow = arrow(),

```

```

    size = 1.15) +
geom_segment(x = Estudiante[140], y = NotaMates[140], xend = Estudiante[140],
  yend = NotaMates[140] - eps_ran[140], color = 1, arrow = arrow(),
  size = 1.15) +
geom_segment(x = Estudiante[185], y = NotaMates[185], xend = Estudiante[185],
  yend = NotaMates[185] - eps_ran[185], color = 1, arrow = arrow(),
  size = 1.15) +
geom_segment(x = Estudiante[264], y = NotaMates[264], xend = Estudiante[264],
  yend = NotaMates[264] - eps_ran[264], color = 1, arrow = arrow(),
  size = 1.15) +
#Erros
annotate("text", x = Estudiante[72] + 20, y = (NotaMates[72] + estloc[3])/2,
  parse = TRUE, label = expression(widehat(epsilon)[7][",")[3]),
  col = 1, size = 7) +
annotate("text", x = Estudiante[140] - 23, y = (NotaMates[140] + estloc[4])/2,
  parse = TRUE, label = expression(widehat(epsilon)[26][",")[4]),
  col = 1, size = 7) +
annotate("text", x = Estudiante[185] + 24.5, y = (NotaMates[185] + estloc[5])/2
  - .1, parse = TRUE, label = expression(widehat(epsilon)[37][",")[5]),
  col = 1, size = 7) +
annotate("text", x = Estudiante[264] - 23, y = (NotaMates[264] + estloc[7])/2
  - .5, parse = TRUE, label = expression(widehat(epsilon)[16][",")[7]),
  col = 1, size = 7) +
#Observaciones
geom_point(aes(x = Estudiante[72], y = NotaMates[72]), col = dark_2[3],
  size = 5) +
geom_point(aes(x = Estudiante[140], y = NotaMates[140]), col = dark_2[4],
  size = 5) +
geom_point(aes(x = Estudiante[185], y = NotaMates[185]), col = dark_2[5],
  size = 5) +
geom_point(aes(x = Estudiante[264], y = NotaMates[264]), col = dark_2[7],
  size = 5) +
annotate("text", x = Estudiante[72] - 18, y = NotaMates[72] - .55, parse =
  TRUE, label = expression(Y[7][",")[3]), col = dark_2[3], size = 6) +
annotate("text", x = Estudiante[140] - 12, y = NotaMates[140] - .65, parse =
  TRUE, label = expression(Y[26][",")[4]), col = dark_2[4], size = 6) +
annotate("text", x = Estudiante[185] - 12, y = NotaMates[185] - .65, parse =

```

```

  TRUE, label = expression(Y[37][","][5]), col = dark_2[5], size = 6) +
  annotate("text", x = Estudante[264] - 12, y = NotaMates[264] - .65, parse =
    TRUE, label = expression(Y[16][","][7]), col = dark_2[7], size = 6) +
  #Media global símbolo
  annotate("text", x = 329, y = summary(ranovamates)$coef[1], parse = TRUE,
    label = expression(widehat(mu)), size = 7) +
  #Medias locais símbolos
  annotate("text", x = 330, y = estloc[3] + .25, parse = TRUE, label =
    expression(paste(widehat(mu) + widehat(u)[3], "=", widehat(mu)[3])),
    size = 6, col = dark_2[3]) +
  annotate("text", x = 330, y = estloc[4], parse = TRUE, label =
    expression(paste(widehat(mu) + widehat(u)[4], "=", widehat(mu)[4])),
    size = 6, col = dark_2[4]) +
  annotate("text", x = 330, y = estloc[5] - .1, parse = TRUE, label =
    expression(paste(widehat(mu) + widehat(u)[5], "=", widehat(mu)[5])),
    size = 6, col = dark_2[5]) +
  annotate("text", x = 330, y = estloc[7], parse = TRUE, label =
    expression(paste(widehat(mu) + widehat(u)[7], "=", widehat(mu)[7])),
    size = 6, col = dark_2[7]) +
  #Resíduos a nivel de escola  $u^{\text{gorro}}_{\{0j\}}$ 
  geom_segment(x = -7, y = summary(ranovamates)$coef[1], xend = -7, yend =
    estloc[3], color = dark_2[3], arrow = arrow(length = unit(0.2, "inches")),
    linetype = "solid", size = .9) +
  geom_segment(x = -7, y = summary(ranovamates)$coef[1], xend = -7, yend =
    estloc[4], color = dark_2[4], arrow = arrow(length = unit(0.2, "inches")),
    linetype = "solid", size = .9) +
  geom_segment(x = 20, y = summary(ranovamates)$coef[1], xend = 20, yend =
    estloc[5], color = dark_2[5], arrow = arrow(length = unit(0.15, "inches")),
    linetype = "solid", size = .75) +
  geom_segment(x = 20, y = summary(ranovamates)$coef[1], xend = 20, yend =
    estloc[7], color = dark_2[7], arrow = arrow(length = unit(0.15, "inches")),
    linetype = "solid", size = .75) +
  #Símbolos resíduos a nivel de escola  $u^{\text{gorro}}_{\{0j\}}$ 
  annotate("text", x = 5, y = (summary(ranovamates)$coef[1] + estloc[3]) / 2
    - .3, parse = TRUE, label = expression(widehat(u)[3]), col = dark_2[3],
    size = 6.5) +
  annotate("text", x = 5, y = (summary(ranovamates)$coef[1] + estloc[4]) / 2,

```

```

    parse = TRUE, label = expression(widehat(u)[4]), col = dark_2[4],
    size = 6.5) +
  annotate("text", x = 36, y = (summary(ranovamates)$coef[1] + estloc[5]) / 2,
    parse = TRUE, label = expression(widehat(u)[5]), col = dark_2[5],
    size = 6.5) +
  annotate("text", x = 36, y = (summary(ranovamates)$coef[1] + estloc[7]) / 2,
    parse = TRUE, label = expression(widehat(u)[7]), col = dark_2[7],
    size = 6.5) +
  theme(
    legend.title = element_text(size = 18),
    legend.text = element_text(size = 16),
    axis.title.x = element_blank(),
    axis.text.x = element_blank(),
    axis.title.y = element_text(size = 18),
    axis.text.y = element_text(size = 16))

```

A.7. Modelos mixtos con covariables relativas ao nivel 1

A.7.1. Modelo con intercepto aleatorio e pendente fixa

```

matesmm1 <- lmer(NotaMates ~ StSE + (1 | Escola), data = mates7, REML = FALSE)
summary(matesmm1)
coefsmm1 <- summary(matesmm1)$coef
sigma2_eps_mm1 <- summary(matesmm1)$sigma^2
sigma2_u_mm1 <- as.data.frame(summary(matesmm1)$varcor)[1,4]
VT_mm1 <- sigma2_u_mm1 + sigma2_eps_mm1
VPC_mm1 <- sigma2_u_mm1 / VT_mm1
#Cálculo dos efectos aleatorios e dos erros (nivel 1)
u0_mm1 <- ranef(matesmm1, postVar = TRUE)
uj_mm1 <- u0_mm1[[1]][,1]
u0se_mm1 <- sqrt(attr(u0_mm1[[1]], "postVar")[1, , ])
eps_mm1 <- residuals(matesmm1)

```

A.7.2. Modelo con intercepto aleatorio e pendente fixa con SocialMin

```

matesmm2 <- lmer(NotaMates ~ StSE + SocialMin + (1 | Escola), data = mates7,
  REML = FALSE)
summary(matesmm2)
coefsmm2 <- summary(matesmm2)$coef
sigma2_eps_mm2 <- summary(matesmm2)$sigma^2
sigma2_u0_mm2 <- as.data.frame(summary(matesmm2)$varcor)[1,4]
VT_mm2 <- sigma2_u0_mm2 + sigma2_eps_mm2
VPC_mm2 <- (sigma2_u0_mm2) / VT_mm2
#Cálculo dos efectos aleatorios e dos erros (nivel 1)
u0_mm2 <- ranef(matesmm2, postVar = TRUE)
uj_mm2 <- u0_mm2[[1]][,1]
u0se_mm2 <- sqrt(attr(u0_mm2[[1]], "postVar")[1, , ])
eps_mm2 <- residuals(matesmm2)

```


A.7.3. Modelo con intercepto e pendiente aleatorios e mais variable categórica

```

matesmm3 <- lmer(NotaMates ~ StSE + SocialMin + (1 + StSE | Escola),
  REML = TRUE)
summary(matesmm3)
coefsmm3 <- summary(matesmm3)$coef
sigma2_eps_mm3 <- summary(matesmm3)$sigma^2
sigma2_u0_mm3 <- as.data.frame(summary(matesmm3)$varcor)[1,4]
sigma2_u1_mm3 <- as.data.frame(summary(matesmm3)$varcor)[2,4]
sigma_u01_mm3 <- as.data.frame(summary(matesmm3)$varcor)[3,4]
VT_mm3 <- sigma2_u0_mm3 + sigma2_u1_mm3 + sigma2_eps_mm3
VPC_mm3 <- (sigma2_u0_mm3 + sigma2_u1_mm3) / VT_mm3
#Cálculo dos efectos aleatorios e dos erros (nivel 1)
u0_mm3 <- ranef(matesmm3, postVar = TRUE)
u0j_mm3 <- u0_mm3[[1]][,1]
u1j_mm3 <- u0_mm3[[1]][,2]
eps_mm3 <- residuals(matesmm3)

```

A.7.4. Figura 4.3

O código de  relativo á Figura 4.1 concernente ao modelo mixto con intercepto aleatorio e pendente fixa `matesmm1` (construído en A.7.1), e mais o código relativo á Figura 4.2 concernente

ao modelo mixto `matesmm2` (construído en A.7.2) son análogos ao código disposto deseguido.

```
rectamm3 <- function(x = 0, escola = NULL, sm = 0){
  if (!sm %in% 0:1){
    stop("Só hai dúas posibilidades para sm, non (0) ou si (1).")
  }
  if (!is.null(escola)){
    if (!escola %in% 1:7){
      stop("Tal escola non existe, só do 1 ao 7.")
    }
  }

  if (is.null(escola)){
    if (sm == 0){
      coefsmm3[1] + coefsmm3[2] * x
    }
    else{
      coefsmm3[1] + coefsmm3[3] + coefsmm3[2] * x
    }
  }
  else{
    if (sm == 0){
      coefsmm3[1] + u0j_mm3[escola] + (coefsmm3[2] + u1j_mm3[escola]) * x
    }
    else{
      coefsmm3[1] + coefsmm3[3] + u0j_mm3[escola] +
        (coefsmm3[2] + u1j_mm3[escola]) * x
    }
  }
}

ggplot(mates7, aes(x = StSE, y = NotaMates, color = Escola, shape = SocialMin)) +
  geom_point(size = 1.25) +
  coord_cartesian(xlim = c(range(StSE)[1], range(StSE)[2]), ylim = c(-.5, 25) ) +
  xlab("Status socio-económico") +
  ylab("Nota en Matemáticas") +
  labs(shape = "Racial\nminority-\ntario") +
  scale_color_brewer(palette = "Dark2") +
```

```

scale_shape_manual(values = c(16, 17)) +
#Media global
geom_segment(x = -1.7, xend = 1.5, y = rectamm3(-1.7), yend = rectamm3(1.5),
  size = 1.35, col = 1) +
#Media local 4
geom_segment(x = -1.7, xend = 1.50, y = rectamm3(-1.7, escola = 3),
  yend = rectamm3(1.50, escola = 3), lwd = 1, linetype = "twodash",
  color = dark_2[3]) +
geom_segment(x = -1.7, xend = 1.50, y = rectamm3(-1.7, escola = 4),
  yend = rectamm3(1.50, escola = 4), lwd = 1, linetype = "twodash",
  color = dark_2[4]) +
geom_segment(x = -1.7, xend = 1.50, y = rectamm3(-1.7, escola = 4, 1),
  yend = rectamm3(1.50, escola = 4, 1), lwd = 1, linetype = "dotted",
  color = dark_2[4]) +
#Frechas erros nivel 1
geom_segment(x = StSE[72], xend = StSE[72], y = NotaMates[72], yend =
  NotaMates[72] - eps_mm3[72], color = 1, arrow = arrow(), size = 1.15) +
geom_segment(x = StSE[140], xend = StSE[140], y = NotaMates[140], yend =
  NotaMates[140] - eps_mm3[140], color = 1, arrow = arrow(), size = 1.15) +
#Alumnos
geom_point( aes(x = StSE[72], y = NotaMates[72] ), col = dark_2[3], size = 5,
  pch = 16) +
geom_point( aes(x = StSE[140], y = NotaMates[140] ), col = dark_2[4], size = 5,
  pch = 16) +
#Símbolos erros nivel 1
annotate("text", x = StSE[72] - .18, y = (NotaMates[72] + (rectamm3(
  StSE[72], escola = 3)) ) / 2 + .35, parse = TRUE, label =
  expression(widehat(epsilon)[7][","][3]), col = 1, size = 7) +
annotate("text", x = StSE[140] - .22, y = (NotaMates[140] + (rectamm3(
  StSE[140], escola = 4)) ) / 2 - .35, parse = TRUE, label =
  expression(widehat(epsilon)[26][","][4]), col = 1, size = 7) +
#Observaciones
annotate("text", x = StSE[72] + .175, y = NotaMates[72] + .425, parse = TRUE,
  label = expression(Y[7][","][3]), col = dark_2[3], size = 6) +
annotate("text", x = StSE[140] - .12, y = NotaMates[140] - .775, parse = TRUE,
  label = expression(Y[26][","][4]), col = dark_2[4], size = 6) +
#Residuos a nivel de escola de intercepto u^gorro_{0j}, para SocialMin = 0

```

```

geom_segment(x = 0, xend = 0, y = rectamm3(), yend = rectamm3(0, 3), color =
  dark_2[3], arrow = arrow(length = unit(0.15, "inches")), linetype = "solid",
  size = .9) +
geom_segment(x = 0, xend = 0, y = rectamm3(), yend = rectamm3(0, 4), color =
  dark_2[4], arrow = arrow(length = unit(0.15, "inches")), linetype = "solid",
  size = .9) +
annotate("text", x = +.155, y = (rectamm3() + rectamm3(0, 3)) / 2 - .1,
  parse = TRUE, label = expression(widehat(u)[0][3]), col = dark_2[3],
  size = 6.5) +
annotate("text", x = +.155, y = (rectamm3() + rectamm3(0, 4)) / 2, parse =
  TRUE, label = expression(widehat(u)[0][4]), col = dark_2[4], size = 6.5) +
#Pendente xeral gamma_{10}
geom_segment(x = -1.5, xend = -.5, y = rectamm3(-1.5), yend = rectamm3(-1.5),
  linetype = "solid", size = 1, col = "violetred4") +
geom_segment(x = -.5, xend = -.5, y = rectamm3(-1.5), yend = rectamm3(-.5),
  linetype = "solid", size = 1, col = "violetred4") +
annotate("text", x = -.345, y = (rectamm3(-1.5) + rectamm3(-.5)) / 2 - .35,
  parse = TRUE, label = expression(widehat(gamma)[1][0]), col = "violetred4",
  size = 6) +
#Pendente escola E4 gamma_{10} + ugorro_{14} con SocialMin = 0
geom_segment(x = -1.5, xend = -.5, y = rectamm3(-1.5, 4), yend = rectamm3(
  -1.5, 4), linetype = "solid", size = 1, col = "violetred4") +
geom_segment(x = -.5, xend = -.5, y = rectamm3(-1.5, 4), yend = rectamm3(
  -.5, 4), linetype = "solid", size = 1, col = "violetred4") +
annotate("text", x = -.175, y = (rectamm3(-1.5, 4) + rectamm3(-.5, 4)) / 2
  - .1, parse = TRUE, label = expression(widehat(gamma)[1][0] +
  widehat(u)[1][4]), col = "violetred4", size = 6) +
#Pendente escola E4 gamma_{10} + ugorro_{14} con SocialMin = 1
geom_segment(x = -1.5, xend = -.5, y = rectamm3(-1.5, 4, 1), yend = rectamm3(
  -1.5, 4, 1), linetype = "solid", size = 1, col = "violetred4") +
geom_segment(x = -.5, xend = -.5, y = rectamm3(-1.5, 4, 1), yend = rectamm3(
  -.5, 4, 1), linetype = "solid", size = 1, col = "violetred4") +
annotate("text", x = -.175, y = (rectamm3(-1.5, 4, 1) + rectamm3(-.5, 4, 1))
  / 2 - .1, parse = TRUE, label = expression(widehat(gamma)[1][0] +
  widehat(u)[1][4]), col = "violetred4", size = 6) +
#Pendente escola E3 gamma_{10} + ugorro_{13}
geom_segment(x = -1.5, xend = -.5, y = rectamm3(-1.5, 3), yend = rectamm3(

```

```

-1.5, 3), linetype = "solid", size = 1, col = "violetred4") +
geom_segment(x = -.5, xend = -.5, y = rectamm3(-1.5, 3), yend = rectamm3(
-.5, 3), linetype = "solid", size = 1, col = "violetred4") +
annotate("text", x = -.375, y = (rectamm3(-1.5, 3) + rectamm3(-.5, 3)) / 2
-.85, parse = TRUE, label = expression(widehat(gamma)[1][0] +
widehat(u)[1][3]), col = "violetred4", size = 6, angle = 10) +
#Pendente associada a SocialMin
geom_segment(x = 1.25, xend = 1.25, y = rectamm3(1.25, 4), yend = rectamm3(
1.25, 4, 1), color = dark_2[4], arrow = arrow(length = unit(0.15, "inches")),
linetype = "solid", size = .9) +
annotate("text", x = 1.375, y = (rectamm3(1.35, 4) + rectamm3(1.35, 4, 1)) / 2
+.15, parse = TRUE, label = expression(widehat(beta)[2]), col = dark_2[4],
size = 6) +
theme(
  legend.title = element_text(size = 18),
  legend.text = element_text(size = 16),
  axis.title.x = element_text(size = 18),
  axis.text.x = element_text(size = 16),
  axis.title.y = element_text(size = 18),
  axis.text.y = element_text(size = 16))

```

A.8. Modelos mixtos con covariables relativas ao nivel 2

A.8.1. Variable contextual composicional

```

matesmm4 <- lmer(NotaMates ~ StSE + SocialMin + MStSE + (1 + StSE | Escola),
  REML = TRUE)
summary(matesmm4)
coefsmm4 <- summary(matesmm4)$coef
sigma2_eps_mm4 <- summary(matesmm4)$sigma^2
sigma2_u0_mm4 <- as.data.frame(summary(matesmm4)$varcor)[1,4]
sigma2_u1_mm4 <- as.data.frame(summary(matesmm4)$varcor)[2,4]
sigma_u01_mm4 <- as.data.frame(summary(matesmm4)$varcor)[3,4]
#Cálculo dos efectos aleatorios e dos erros (nivel 1)
u0_mm4 <- ranef(matesmm4, postVar = TRUE)
u0j_mm4 <- u0_mm4[[1]][,1]

```

```

u1j_mm4 <- u0_mm4[[1]][,2]
eps_mm4 <- residuals(matesmm4)

# Contraste sobre efecto fixo de MStSE
LR_mm4 <- 2 * logLik(matesmm4)[1] - 2 * logLik(matesmm3)[1]
pval_mm4_MStSE <- 1 - pchisq(LR_mm4, df = 1) # p-valor

```

A.8.2. Variable contextual global

```

#-----#
# Engadiremos Tamaño ao modelo matesmm3
matesmm5 <- lmer(NotaMates ~ StSE + SocialMin + Tamaño + (1 + StSE | Escola),
  REML = FALSE)
summary(matesmm5)
lrtest(matesmm3, matesmm5)

#Porcentaxe de estudantes en cada escola cunha nota menor que 5
notas_menor5 <- numeric(7)
for (i in 1:length(levels(Escola))){
  notas_menor5[i] <- sum(NotaMates[Escola == levels(Escola)[i]] < 5)
}
round(notas_menor5 / table(Escola) * 100, 2)
#-----#
# Engadiremos AmbDiscrim ao modelo matesmm3
matesmm6 <- lmer(NotaMates ~ StSE + SocialMin + AmbDiscrim + (1 + StSE | Escola),
  REML = FALSE)
summary(matesmm6)
lrtest(matesmm3, matesmm6)
#-----#
# Engadiremos MaioriaMin ao modelo matesmm3
matesmm7 <- lmer(NotaMates ~ StSE + SocialMin + MaioriaMin + (1 + StSE | Escola),
  REML = FALSE)
summary(matesmm7)
lrtest(matesmm3, matesmm7)

```


Bibliografía

- [1] Akaike, H.: *A new look at the statistical identification model*. IEEE Transactions on Automatic Control, **19**, 716 (1974)
- [2] Akaike, H.: *Likelihood and the Bayes procedure*, in “*Bayesian Statistics*”, ed. by JM Bernardo, MH DeGroot, DV Lindley and AFM Smith (1980)
- [3] Bates, D., Maechler, M., Bolker, B., Walker, S.: *Lme4: Linear mixed-effects models using “Eigen” and S4* (2021). URL <https://cran.r-project.org/web/packages/lme4/lme4.pdf>
- [4] Bryk, A.S., Raudenbush, S.W.: *Hierarchical linear models: Applications and data analysis methods*. Sage Publications, Inc (1992)
- [5] Chang, W.: *R graphics cookbook: practical recipes for visualizing data*. O’Reilly Media (2018). URL <https://r-graphics.org/>
- [6] Demidenko, E.: *Mixed models: theory and applications with R*. John Wiley & Sons (2013)
- [7] DiPrete, T.A., Forristal, J.D.: *Multilevel models: methods and substance*. Annual review of sociology, **20**(1), 331–357 (1994)
- [8] Faraway, J.J.: *Linear models with R*. Chapman and Hall/CRC (2004)
- [9] Faraway, J.J.: *Extending the linear model with R: generalized linear, mixed effects and nonparametric regression models*. Chapman and Hall/CRC (2016)
- [10] Fox, J.: *Applied regression analysis and generalized linear models*. Sage Publications (2015)
- [11] Gbur, E.E., Stroup, W.W., McCarter, K.S., Durham, S., Young, L.J., Christman, M., West, M., Kramer, M.: *Analysis of generalized linear mixed models in the agricultural and natural resources sciences*, vol. 156. John Wiley & Sons (2020)
- [12] Goldstein, H.: *Restricted unbiased iterative generalized least-squares estimation*. Biometrika, **76**(3), 622–623 (1989)

-
- [13] Goldstein, H.: Multilevel statistical models. John Wiley & Sons (2011)
- [14] Goldstein, H., Sc, B.: Models for reality: New approaches to the understanding of educational processes. University of London, Institute of Education (1998)
- [15] Grilli, L., Rampichini, C.: *A handful of critical choices in multilevel modelling*. BEIO, Boletín de Estadística e Investigación Operativa, **34**(1), 7–24 (2018)
- [16] Hedges, L.V., Hedberg, E.C.: *Intraclass correlation values for planning group-randomized trials in education*. Educational Evaluation and Policy Analysis, **29**(1), 60–87 (2007)
- [17] Holm, S.: *A simple sequentially rejective multiple test procedure*. Scandinavian Journal of Statistics, **6**, 65–70 (1979)
- [18] Hox, J.J., Moerbeek, M., Van de Schoot, R.: Multilevel analysis: Techniques and applications. Routledge (2017)
- [19] Ibarrola, R.V., Pérez, A.G.: Principios de inferencia estadística. UNED, Universidad Nacional de Educación a Distancia (2006)
- [20] Kutner, M.H., Nachtsheim, C.J., Neter, J., Li, W., et al.: Applied linear statistical models. McGraw-Hill New York (2005)
- [21] Leyland, A.H., Groenewegen, P.P.: Multilevel modelling for public health and health services research: health in context. Springer Nature (2020)
- [22] McNeish, D.: *Small sample methods for multilevel modeling: A colloquial elucidation of REML and the Kenward-Roger correction*. Multivariate Behavioral Research, **52**(5), 661–670 (2017)
- [23] Petrov, V.V., Ernesto, M.P.: Teoría de la probabilidad. 519.2 PET. Dirac (2008)
- [24] Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., Team, R.C.: *Linear and nonlinear mixed effects models* (2007). URL <https://cran.r-project.org/web/packages/nlme/nlme.pdf>
- [25] Rabe-Hesketh, S., Skrondal, A.: Multilevel and longitudinal modeling using Stata. STATA Press (2008)
- [26] Rao, C.R.: Linear statistical inference and its applications, vol. 2. Wiley New York (1973)
- [27] Roback, P., Legler, J.: Beyond multiple linear regression: applied generalized linear models and multilevel models in R. Chapman and Hall/CRC (2021). URL <https://bookdown.org/roback/bookdown-BeyondMLR/>
- [28] Scott, M.A., Simonoff, J.S., Marx, B.D.: The SAGE handbook of multilevel modeling. Sage (2013)

-
- [29] Snijders, T.A., Bosker, R.J.: Multilevel analysis: An introduction to basic and advanced multilevel modeling. Sage (2011)
- [30] Steele, F.: *Module 5: Introduction to Multilevel modelling concepts*. LEMMA (Learning Environment for Multilevel Methodology and Applications), Centre for Multilevel Modelling, University of Bristol (2008)
- [31] Wickham, H.: Advanced R. CRC Press (2019). URL <https://adv-r.hadley.nz/>
- [32] Wickham, H., Grolemund, G.: R for data science: import, tidy, transform, visualize, and model data. O'Reilly Media, Inc. (2016). URL <https://r4ds.had.co.nz/>
- [33] Xie, Y.: Dynamic Documents with R and knitr. Chapman and Hall/CRC (2017). URL <https://yihui.org/knitr/>
- [34] Zuur, A.F., Ieno, E.N., Walker, N.J., Saveliev, A.A., Smith, G.M., et al.: Mixed effects models and extensions in ecology with R, vol. 574. Springer (2009)