aula modelagem adicionais e Introducao ao pacote lme4

Prof. Caio Azevedo

Mais sobre modelagem

- Voltaremos a alguns dos exemplos vistos até o momento, em particular, aqueles vistos em aula sobre modelos de dois níveis.
- Como discutido, podemos considerar diversos modelos hierárquicos para um mesmo conjunto de dados e diversos aspectos devem ser levados em consideração, para tal escolha (veja).
- Num primeiro momento, embora não seja essa abordagem mais apropriada, trabalharemos apenas com estimação e comparação de modelos, como visto aqui. Posteriormente, discutiremos a verificação da qualidade do ajuste.

Mais sobre modelagem

- Uma abordagem mais apropriada é como aquela descrita aqui (nos slides 17 e 18), a qual discutiremos (e usaremos) ao longo do curso.
- Mesmo que o modelo escolhido (como veremos adiante) não se ajuste bem aos dados, continuaremos com as análises, sob as devidas ressalvas.

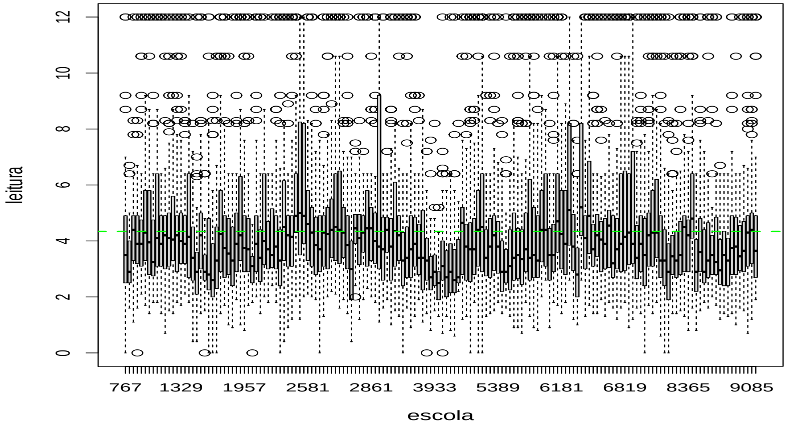
Exemplo 3: Desempenho cognitivo (Achievement data)

- Relativo ao "Indiana's Prime Time" que consite em um mecanismo de financiamento projetado para reduzir o tamanho das turmas (classes) e/ou reduzir o tamanho da razão aluno/professor razão nas primeiras séries do ensino fundamental (Estados Unidos).
- Várias variáveis foram medidas: conhecimento (desempenho) em várias áreas, idade, sexo, turma, escola entre outras.
- Refere-se a 10.903 estudantes (Multilevel Modeling using R, embora no banco de dados haja apenas 10.320 estudantes) distribuídos em 160 escolas com quantidades bem variadas (de 11 a 143) de alunos em cada uma delas.

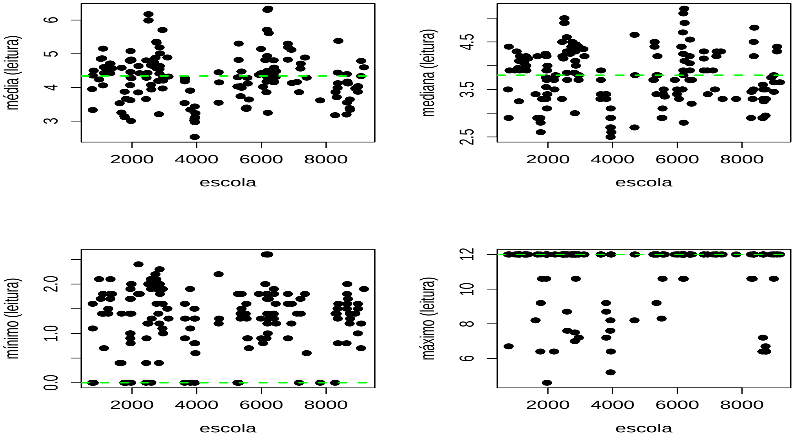
Exemplo 3: Desempenho cognitivo (Achievement data)

- Objetivo (neste exemplo): identificar, ao longo das escolas, como o desempenho num teste de leitura (geread) varia e é afetado pelo desempenho num teste de vocabulário (gevocab), ao longo das escolas (schools). Quanto maior os valores dos desempenhos (escores) maior o conhecimento na área em questão. Doravante nomearemos as três variáveis anteriores de: leitura, vocabulário e escola, respectivamente.
- Vamos explorar, descritivamente, os dados. Lembre que não, necessariamente, as análises descritivas servirão para verificar (antecipadamente) as hipóteses do modelo a ser usado (temos uma covariável quantitativa).

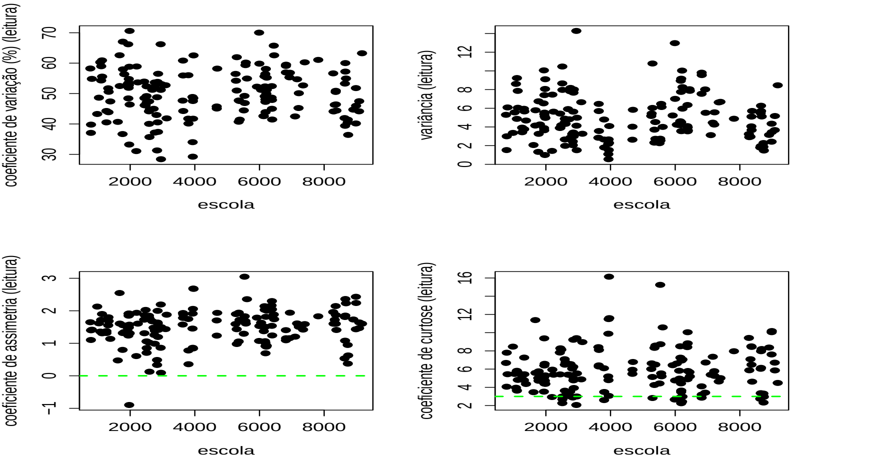
Box plot da leitura por escola



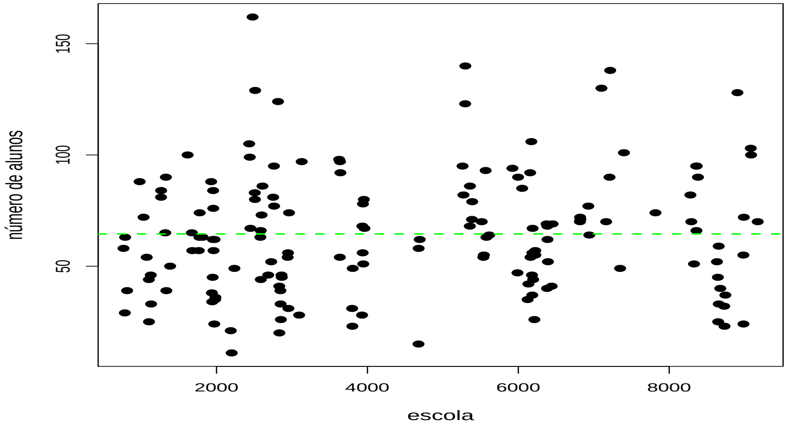
Medidas resumo da leitura: parte 1



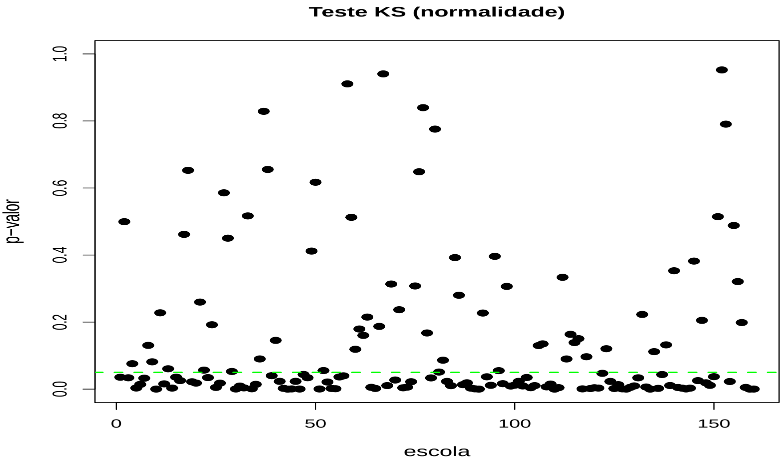
Medidas resumo da leitura: parte 2



Número de alunos por escola



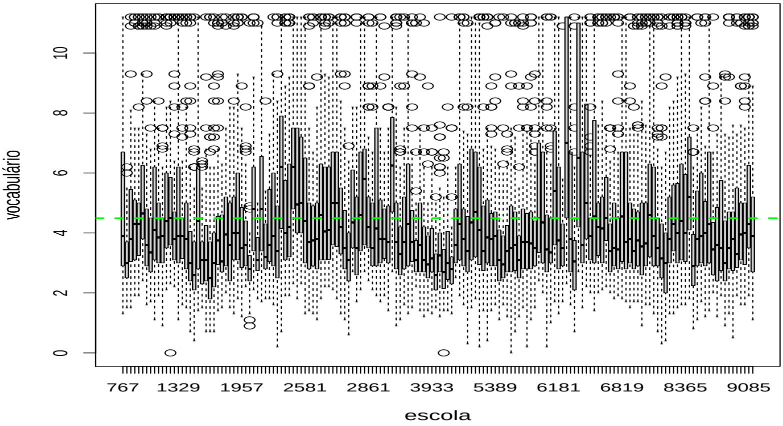
Teste de KS (leitura) por escola



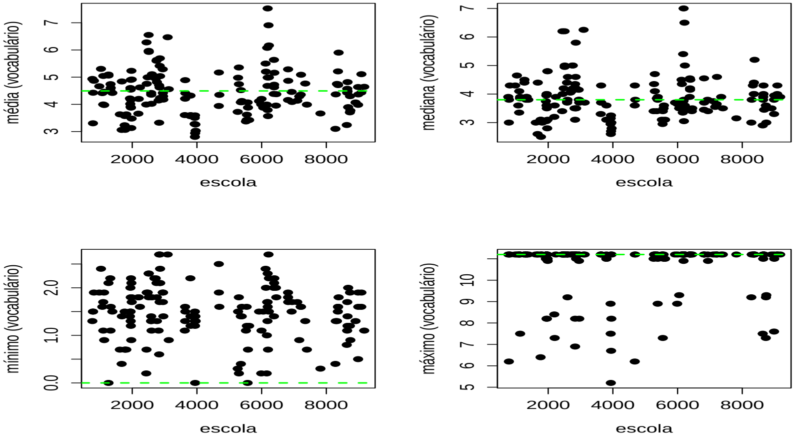
Teste de homocedasticidade e CCI

- $CCI_1 = 0{,}072$, $CCI_2 = 0{,}072$, $CCI_3 = 0{,}073$ (percentuais pequenos). Contudo, note que temos que considerar a covariável "vocabulário".
- Teste de Levene (homocedasticidades): 3,14 (p-valor $< 0{,}0001$).
- Note que os testes (normalidade e homecedasticidade) bem como a estimativa da CCI, são compatíveis com o modelo sem covariáveis. Na presença de covariáveis, tais resultados não são (em geral) apropriados, haja vista que nos modelos de regressão representa-se distribuições condicionais.

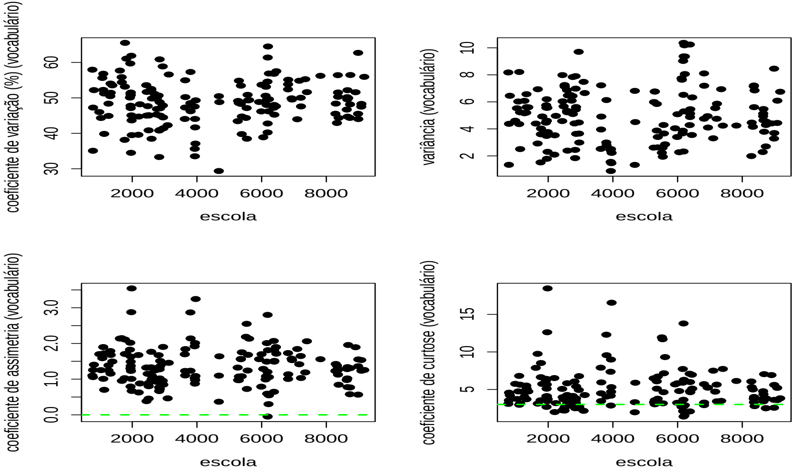
Box plot do vocabulário por escola



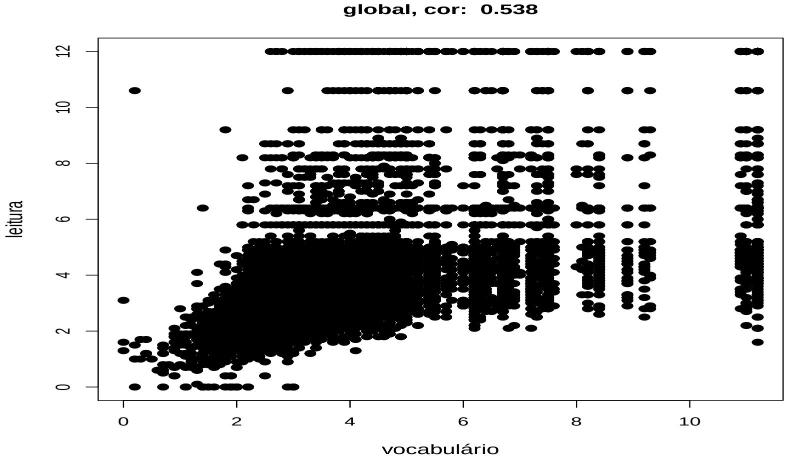
Medidas resumo do vocabulário: parte 1



Medidas resumo do vocabulário: parte 2



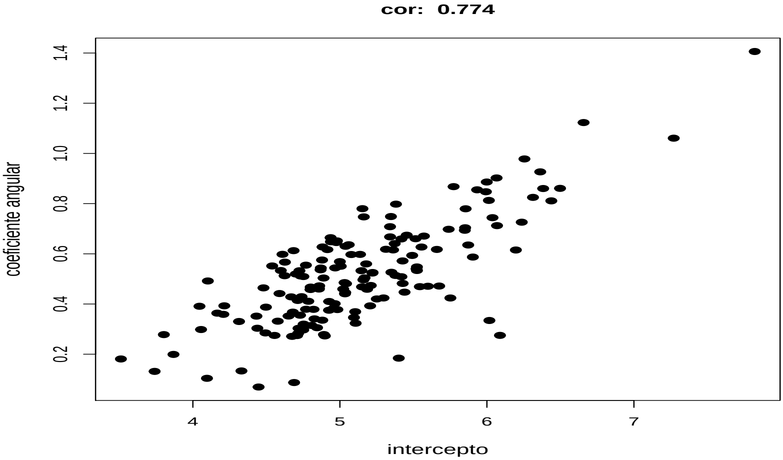
Dispersão (todas as escolas) vocabulário × leitura



Dispersão (por escola) vocabulário $\times$ leitura

Gráficos de dispersão

Dispersão entre os coef. de regressão entre as escolas



Modelo com dois níveis e uma covariável no nível 1

- Neste caso, o modelo (doravante, modelo 1) é dado por:

$$
\begin{aligned}
Y_{ji} &= \beta_{0j} + \beta_{1j}(x_{ji} - 6) + \xi_{ji}, j = 1, 2, ..., 160; \\
& \quad i = 1, 2, ..., n_j \,(\text{nível 1 - alunos}) \\
\beta_{0j} &= \gamma_{00} + u_{0j} \,(\text{nível 2 - escolas}) \\
\beta_{1j} &= \gamma_{10} + u_{1j} \,(\text{nível 2 - escolas})
\end{aligned}
$$

- Erros e efeitos aleatórios: $\xi_{ji} \overset{iid}{\sim} N(0, \sigma^2)$,

$\boldsymbol{u}_j = (u_{0j}, u_{1j})' \overset{iid}{\sim} N_2(0, \boldsymbol{\Psi})$, $\xi_{ji} \perp \boldsymbol{u}_j, \forall i, j$ e $(\gamma_{00}, \gamma_{10})'$ são não aleatórios, $\boldsymbol{\Psi} = \begin{bmatrix} \psi_{00} & \psi_{01} \\ \psi_{01} & \psi_{11} \end{bmatrix}$

Modelo 1

- Y_{ji} : é o escore em leitura do indivíduo i a escola j.
- $x^*_{ji} = x_{ji} - 6$: é o escore no teste de vocabulário do indivíduo i a escola j, menos o valor 6.
- (momentos condicionais) $\mathcal{E}(Y_{ji}|\boldsymbol{u}_j) = \gamma_{00} + u_{0j} + (\gamma_{10} + u_{1j})x^*_{ji}$ - valor esperado da leitura para alunos pertecentes a escola j e com valor do escore no teste de vocabulário igual x_{ji}. $\mathcal{V}(Y_{ji}|\boldsymbol{u}_j) = \sigma^2$.
- (momentos marginais) $\mathcal{E}(Y_{ji}) = \gamma_{00} + \gamma_{10}x_{ji}$ - valor esperado da leitura para alunos pertecentes a escola j e com valor do vocabulário igual x_{ji}. $\mathcal{V}(Y_{ji}) = \psi_{00} + \psi_{11}x^{*2}_{ji} + \psi_{01}x^*_{ji} + \sigma^2$.
- $Cov(Y_{ji}, Y_{ji'}) = \psi_{00} + \psi_{11}x^*_{ji}x^*_{ji'} + \psi_{01}x^*_{ji} + \psi_{01}x^*_{ji'}$, $i \neq i'$

Modelo 1

- $Corre(Y_{ji}, Y_{ji'}) = \frac{\psi_{00}+\psi_{11}x_{ji}^{*}x_{ji'}^{*}+\psi_{01}x_{ji}^{*}+\psi_{01}x_{ji'}^{*}}{\sqrt{\psi_{00}+\psi_{11}x_{ji}^{*2}+\psi_{01}x_{ji}^{*}+\sigma^{2}}\sqrt{\psi_{00}+\psi_{11}x_{ji'}^{*2}+\psi_{01}x_{ji'}^{*}+\sigma^{2}}}$
- γ_{00} : leitura esperada para indivíduos com vocabulário igual a 6.
- γ_{10} : incremento na leitura (positivo ou negativo) para o aumento em uma unidade no vocabulário.

Modelo 1

- $u_{0j} = \mathcal{E}(Y_{ji}|\boldsymbol{u}_j, x_{ji} = 6) - \mathcal{E}(Y_{ji}|x_{ji} = 6)$ é a diferença entre o valor esperado para a escola j e a esperança global, para alunos com $x_{ji} = 6$. Outra interpretação: é o quanto o intercepto da escola j se difere do intercepto comum a todas as escolas.
- $u_{1j} = \frac{1}{x_{ji}^*}\left(\mathcal{E}(Y_{ji}|u_{1j}) - \mathcal{E}(Y_{ji})\right), \forall x_{ji}^* \neq 0$, é a diferença entre o valor esperado da leitura para a escola j, em relação a distribuição de u_{0j} e a esperança global da leitura (ponderado pelo valor x_{ji}^*). Outra interpretação: é o quanto o coeficiente angular da escola j se difere do coeficiente angular comum a todas as escolas.

Casos pariticulares

- Coeficientes independentes/não correlacionados (modelo 2): $\psi_{01} = 0$
- Somente interceptos aleatórios (modelo 3): $\psi_{11} = 0$
- Somente coeficientes angulares aleatórios (modelo 4): $\psi_{00} = 0$
- Sem covariável (modelo 5): $\psi_{11} = 0$ e $\gamma_{10} = 0$

Pacote "lme4"

- Pacote que permite ajustar modelos mistos/hierárquicos:
  - Modelos hierárquicos normais/normais (erro associado a resposta/efeitos aleatórios) lineares (o que estamos vendo até agora). Função: "lmer".
  - Modelos hierárquicos generalizados (veremos mais adiante, como em). Função: "glmer".
  - Modelos hierárquicos normais/normais (erro associado a resposta/efeitos aleatórios) não-lineares. Função: "nlmer".

Pacote “lme4”

- Sintaxes básicas muito parecidas com as das funções “lm”, “glm” e “nls”.
- Exemplo: lmer(respostas$\sim$efeitos fixos + (.|efeitos aleatórios), data = “dataframe”). A parte (.) refere-se a escolha da estrutura (intercepto, coeficiente angular etc).
- Recomenda-se o estudo de Bates et al (2015), Galecki and Burzykowski (2012) e manual do lmer.
- Nesse pacote temos opções de ajuste (estimação) frequentista (predição Bayesiana empírica dos efeitos aleatórios), cálculo de resíduos e testes de hipótese.

Exemplos genéricos de sintaxe

Fórmula	Alternativa	Significado
(1 | g)	1 + (1 | g)	Interceptos aleatórios com média fixa (Int.)
0 + offset(0) + (1 | g)	-1 + offset(0) + (1 | g)	Interceptos aleatórios com média a priori
(1 | g1/g2)	(1 | g1)+(1 | g1:g2)	Interceptos variando entre g_1 e entre g_1 dentro de g_2
(1 | g1) + (1 | g2)	1 + (1 | g1) + (1 | g2)	Interceptos variando entre g_1 e g_2
x + (x | g)	1 + x + (1 + x | g)	Intercepto e coeficiente angular correlacionados
x + (x || g)	1 + x + (1 | g) + (0 + x | g)	Intercepto e coeficiente angular não-correlacionados

Sintaxe para o ajuste dos cinco modelos

- M5: lmer(geread~1+(1|schoolF),data=achieve)
- M4: lmer(geread~1+I(gevocab-auxregC)+(-1+I(gevocab-auxregC)|schoolF),data=achieve)
- M3: lmer(geread~1+I(gevocab-auxregC)+(1|schoolF),data=achieve)
- M2: lmer(geread~1+I(gevocab-auxregC)+(1|schoolF)+(0+I(gevocab-auxregC)|schoolF),data=achieve)
- M1: lmer(geread~1+I(gevocab-auxregC)+(1+I(gevocab-auxregC)|schoolF),data=achieve)

Estatísticas de Compração de Modelo (link)

Estatística	modelo				
	1	2	3	4	5
AIC	43004,85	43068,99	43145,20	43206,70	46274,31
BIC	43048,30	43105,20	43174,17	43235,67	46296,03
AICC	42992,86	43059,00	43137,20	43198,71	46268,31
HQIC	43019,54	43081,23	43154,99	43216,49	46281,65
CAIC	43054,30	43110,20	43178,17	43239,67	46299,03
SABIC	43029,23	43089,31	43161,46	43222,96	46286,50

Estimativas: efeitos fixos e componentes de variância

Efeitos fixos

Parâmetro	Estimativa	EP	IC(95%)	Estat. z	p-valor
γ_{00}	5,13	0,05	[5,04 ; 5,22]	113,10	$<$ 0,0001
γ_{10}	0,52	0,01	[0,49 ; 0,55]	36,10	$<$ 0,0001

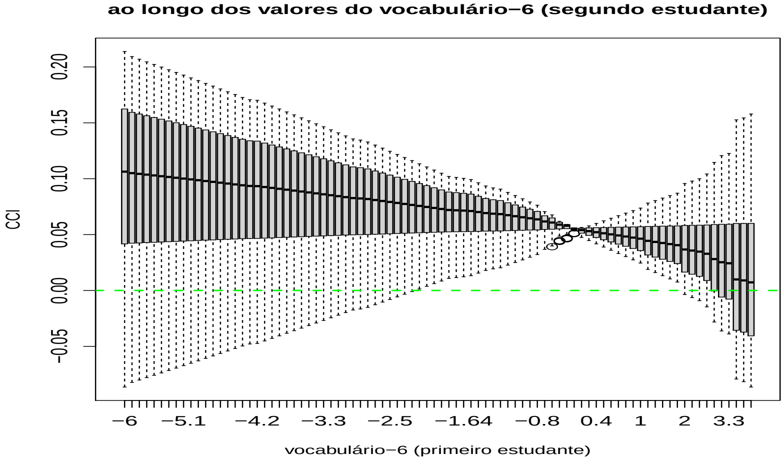
Componentes de variância

ψ_{00}	ψ_{11}	ψ_{01}	ρ_{01}	σ^2
0,21657	0,01930	0,05240	0,81044	3,66594

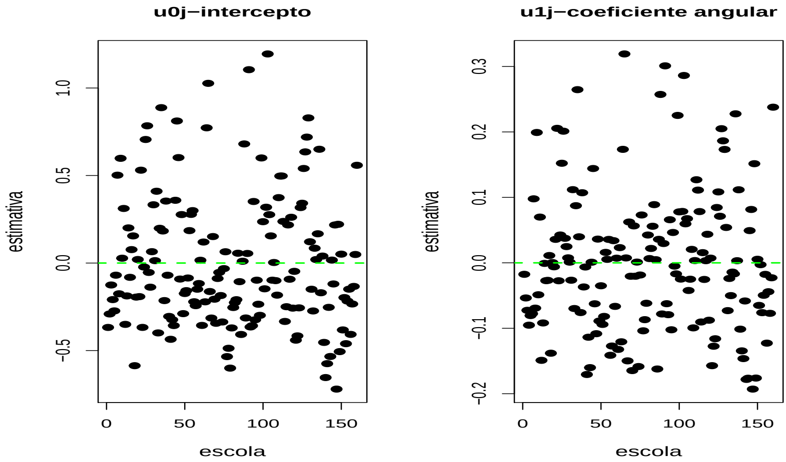
CCI via modelo: dispersão

link

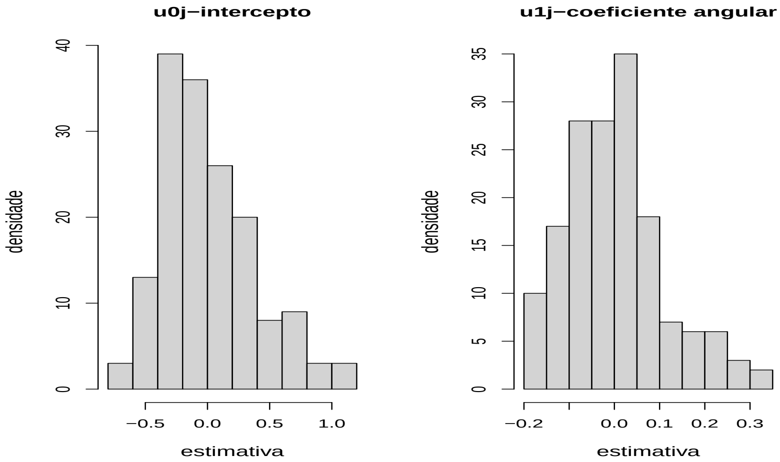
CCI via modelo: histograma



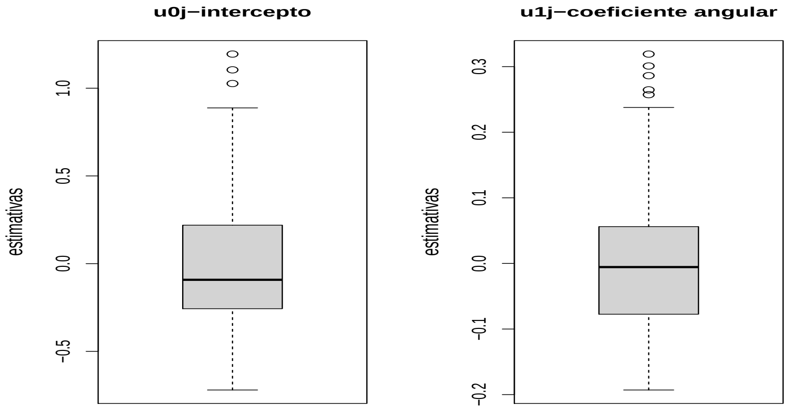
Estimativas dos efeitos aleatórios



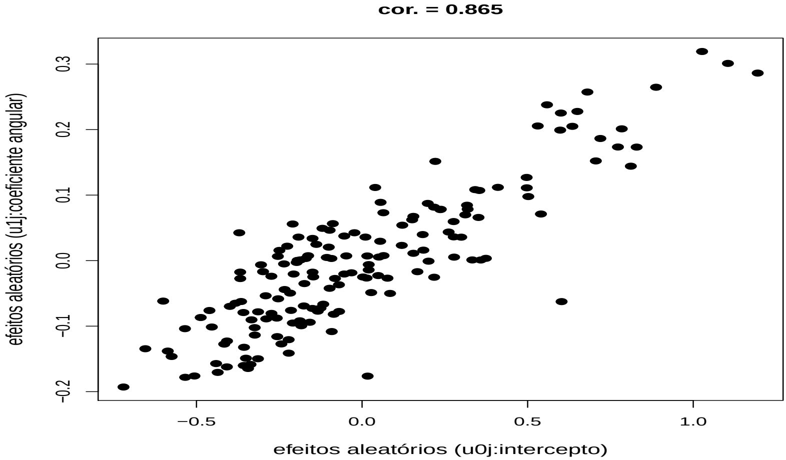
Histograma das estimativas dos efeitos aleatórios



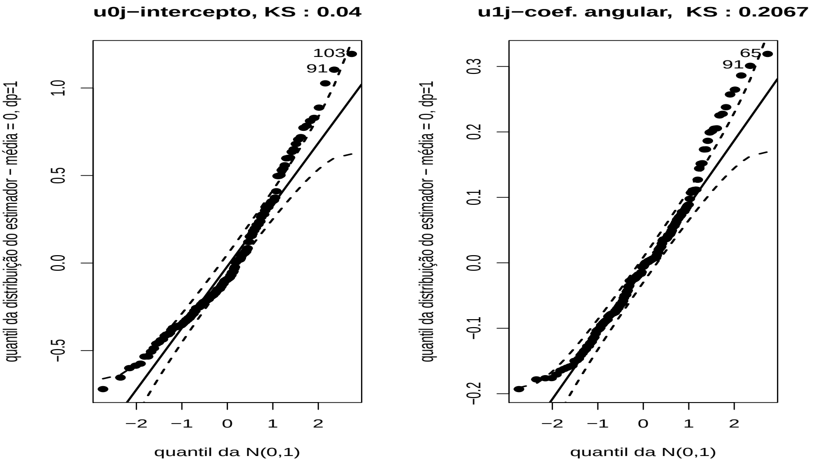
Boxplot das estimativas dos efeitos aleatórios



Dispersão das estimativas dos efeitos aleatórios



QQ plot das estimativas dos efeitos aleatórios



Comentários

- O melhor modelo foi o M1, que considera interceptos e coeficientes angulares aleatórios (faz-se mister realizar análises de daignóstico/residual).
- Os efeitos fixos foram significativos. O incremento na nota em leitura é da ordem de [0,49;0,55] pontos, para o aumento em uma unidade no escore em vocabulário. Por outro lado, o escore esperado em leitura, para um escore observado de 6 em vocabulário, é da ordem de [5,04;5,22].
- A variabilidade devida aos efeitos aelatórios é pequena mas, pelo menos em parte, sigificativa, em relação a variabilidade total.
- Correlação signficativa entre os efeitos aleatórios.