CAPÍTULO 16

Modelos Multinível de Regressão para Dados em Painel

Devemos expandir o círculo do nosso amor até que ele englobe todo o nosso bairro; do bairro, por sua vez, deve desdobrar-se para toda a cidade; da cidade para o estado e assim sucessivamente, até que o objeto do nosso amor inclua todo o universo.
Mahatma Gandhi

Ao final deste capítulo, você terá condições de:

- Estabelecer as circunstâncias a partir das quais os modelos de regressão multinível podem ser utilizados.
- Entender como funcionam as estruturas aninhadas de dados agrupados e de dados com medidas repetidas, e saber definir diversos tipos de constructos a partir dos quais os modelos multinível podem ser utilizados.
- Propor modelos em que seja possível identificar os efeitos fixos e os efeitos aleatórios sobre a variável dependente.
- Estimar parâmetros de modelos hierárquicos lineares de dois níveis com dados agrupados e de três níveis com medidas repetidas, e saber interpretá-los.
- Compreender a decomposição de variância dos efeitos aleatórios em caráter multinível.
- Calcular e interpretar as correlações intraclasse de cada nível da análise.
- Saber diferenciar um modelo multinível de um modelo tradicional de regressão.
- Elaborar testes de razão de verossimilhança para comparar estimações de diferentes modelos multinível.
- Estimar modelos de regressão multinível no Stata Statistical Software® e no IBM SPSS Statistics Software® e interpretar seus resultados.

16.1. INTRODUÇÃO

Os **modelos multinível de regressão para dados em painel** têm adquirido importância considerável em diversas áreas do conhecimento, e a publicação de trabalhos que fazem uso de estimações relacionadas a esses modelos tem sido cada vez mais frequente, muito em função da determinação de constructos de pesquisa que consideram a existência de **estruturas aninhadas de dados**, em que determinadas variáveis apresentam variação entre unidades distintas que representam grupos, porém não entre observações pertencentes a um mesmo grupo. O próprio desenvolvimento computacional e o investimento que determinadas empresas fabricantes de softwares de análise de dados têm feito na capacidade de processamento para estimação de modelagens multinível também oferecem suporte a pesquisadores cada vez mais interessados nesse tipo de abordagem.

Imagine que um grupo de pesquisadores tenha interesse em estudar como o desempenho de firmas, medido, por exemplo, por determinado indicador de rentabilidade, comporta-se em relação a determinadas características de operação das empresas (porte, investimento, entre outras) e com relação às características do setor em que cada firma atua (participação no PIB, incentivos fiscais e de legislação, entre outras). Como as características dos setores não variam entre firmas provenientes do mesmo setor, caracteriza-se uma **estrutura**

de dados agrupados em dois níveis, com firmas (nível 1) aninhadas em empresas (nível 2). A estimação de um modelo multinível pode propiciar ao pesquisador uma possibilidade de verificar se existem características de firmas que explicam eventuais diferenças de desempenho entre companhias provenientes do mesmo setor, bem como se existem características dos setores que explicam eventuais diferenças no desempenho de firmas provenientes de setores distintos.

Imagine ainda que este estudo seja ampliado para que se investigue a evolução temporal do desempenho dessas firmas. Ao contrário dos modelos longitudinais de regressão para dados em painel (Capítulo 15), em que as variáveis sofrem alterações entre observações e ao longo do tempo, imagine que o banco de dados seja estruturado apenas com variáveis de firmas (estrutura de governança, linhas de produção, entre outras) e de setores (incidência tributária, legislação, entre outras) que não se alteram durante o período analisado. Desta forma, caracteriza-se uma **estrutura de dados com medidas repetidas em três níveis**, com períodos de tempo (nível 1) aninhados em firmas (níveis 2), e estas em setores (nível 3) e, a partir da qual, podem ser estimados modelos com o intuito de se investigar se existe variabilidade no desempenho, ao longo do tempo, entre firmas de um mesmo setor e entre aquelas provenientes de setores distintos e, em caso afirmativo, se existem características de firmas e de setores que explicam essa variabilidade.

Em tese, o pesquisador pode definir um constructo com uma quantidade maior de níveis de análise, mesmo que a interpretação dos parâmetros do modelo não seja algo trivial. Por exemplo, imagine o estudo do desempenho escolar, ao longo do tempo, de estudantes aninhados em escolas, estas em distritos municipais, estes em municípios e estes em estados da federação. Nesse caso, estaríamos trabalhando com seis níveis de análise (evolução temporal, estudantes, escolas, distritos municipais, municípios e estados).

A principal vantagem dos modelos multinível sobre modelos tradicionais de regressão estimados, por exemplo, por MQO (Capítulo 12), refere-se à possibilidade de que seja levado em consideração o aninhamento natural dos dados. Em outras palavras, **os modelos multinível permitem que sejam identificadas e analisadas as heterogeneidades individuais e entre grupos a que pertencem estes indivíduos, tornando possível a especificação de componentes aleatórios em cada nível da análise**. Por exemplo, se empresas estiverem aninhadas em setores, é possível que se defina um componente aleatório no nível de firma e outro no nível de setor, ao contrário do que permitiria um modelo tradicional de regressão, em que o efeito do setor sobre o desempenho das firmas seria considerado de maneira homogênea. Nesse sentido, os modelos multinível também podem ser chamados de **modelos de coeficientes aleatórios**.

Neste capítulo, estudaremos os modelos multinível com o intuito de investigar comportamentos de variáveis dependentes métricas e, a partir dos quais, serão gerados resíduos normalmente distribuídos, porém não independentes e sem variância constante. Assim, nosso foco será nos modelos multinível lineares, conhecidos também por **modelos lineares mistos** (em inglês, ***linear mixed models* - *LMM***) ou **modelos hierárquicos lineares** (em inglês, ***hierarchical linear models* - *HLM***). Essa é a razão para que modelos multinível aplicados a dados aninhados em dois níveis sejam também denominados **HLM2**, e que modelos aplicados a dados aninhados em três níveis sejam conhecidos por **HLM3**.

De acordo com West, Welch e Gałecki (2015), a denominação modelos lineares mistos vem do fato de que esses modelos apresentam **especificação linear** e as variáveis explicativas envolvem um **misto de efeitos fixos e aleatórios**, ou seja, podem ser inseridas tanto em componentes de efeitos fixos, quanto em componentes de efeitos aleatórios. Enquanto os parâmetros estimados de **efeitos fixos** indicam a relação entre as variáveis explicativas e a variável dependente métrica, os componentes de **efeitos aleatórios** podem ser representados pela combinação de variáveis explicativas e termos aleatórios não observados.

No apêndice deste capítulo, faremos uma breve apresentação de modelos multinível não lineares, com aplicações em Stata de exemplos de modelos dos tipos logístico, Poisson e binomial negativo.

Seguindo a lógica do capítulo anterior, elaboraremos todas as modelagens neste capítulo em Stata. Além disso, acreditamos que a elaboração de estimações em SPSS também possa propiciar ao pesquisador a oportunidade de comparação do manuseio dos softwares, dos procedimentos e rotinas para estimação dos modelos e das lógicas com que são apresentados os *outputs*, permitindo que se decida qual software utilizar, em função das características de cada um e da própria acessibilidade para uso.

Neste capítulo, portanto, trataremos dos modelos multinível de regressão para dados em painel, com os seguintes objetivos: (1) introduzir os conceitos sobre estruturas aninhadas de dados; (2) definir o tipo de modelo

a ser estimado em função das características dos dados; (3) estimar parâmetros por meio de diversos métodos em Stata e SPSS; (4) interpretar os resultados obtidos por meio dos vários tipos de estimações existentes para os modelos multinível; e (5) definir a estimação mais adequada para efeitos de diagnóstico e previsão em nos casos estudados. Inicialmente, serão introduzidos os principais conceitos inerentes a cada modelagem. Na sequência, serão apresentados os procedimentos para a elaboração dos modelos propriamente ditos em Stata e SPSS.

16.2. ESTRUTURAS ANINHADAS DE DADOS

Os modelos multinível de regressão permitem que se investigue o comportamento de determinada variável dependente Y, que representa o fenômeno de interesse, com base no comportamento de variáveis explicativas, nas quais alterações podem ocorrer, para dados agrupados, entre observações e entre grupos a que pertencem essas observações, e, para dados com medidas repetidas, também ao longo do tempo. Em outras palavras, **devem existir variáveis que apresentam dados que se alteram entre indivíduos que representam determinado nível, porém permanecem inalteradas para certos grupos de indivíduos, sendo que esses grupos representam um nível superior**.

Imagine inicialmente uma base com dados referentes a n indivíduos, sendo cada indivíduo $i = 1, ..., n$ pertencente a um dos $j = 1, ..., J$ grupos, sendo obviamente $n > J$. Assim, esse banco de dados pode apresentar determinadas variáveis explicativas $X_1, ..., X_Q$ referentes a cada indivíduo i, e outras variáveis explicativas $W_1, ..., W_S$ referentes a cada grupo j, porém invariantes para os indivíduos de determinado grupo. A Tabela 16.1 apresenta o modelo geral de uma base com **estrutura aninhada de dados agrupados em dois níveis** (indivíduo e grupo).

Tabela 16.1 Modelo geral de uma base com estrutura aninhada de dados agrupados em dois níveis.

Observação (Indivíduo i) Nível 1	**Grupo j Nível 2**	Y_{ij}	X_{1ij}	X_{2ij}	...	X_{Qij}	W_{1j}	W_{2j}	...	W_{Sj}
1	1	Y_{11}	X_{111}	X_{211}		X_{Q11}	W_{11}	W_{21}		W_{S1}
2	1	Y_{21}	X_{121}	X_{221}		X_{Q21}	W_{11}	W_{21}		W_{S1}
⋮	⋮	⋮	⋮	⋮		⋮	⋮	⋮		⋮
n_1	1	Y_{n_11}	X_{1n_11}	X_{2n_11}		X_{Qn_11}	W_{11}	W_{21}		W_{S1}
$n_1 + 1$	2	$Y_{n_1+1,2}$	$X_{1n_1+1,2}$	$X_{2n_1+1,2}$		$X_{Qn_1+1,2}$	W_{12}	W_{22}		W_{S2}
$n_1 + 2$	2	$Y_{n_1+2,2}$	$X_{1n_1+2,2}$	$X_{2n_1+2,2}$		$X_{Qn_1+2,2}$	W_{12}	W_{22}		W_{S2}
⋮	⋮	⋮	⋮	⋮	...	⋮	⋮	⋮	...	⋮
n_2	2	Y_{n_22}	X_{1n_22}	X_{2n_22}		X_{Qn_22}	W_{12}	W_{22}		W_{S2}
⋮	⋮	⋮	⋮	⋮		⋮	⋮	⋮		⋮
$n_{J-1} + 1$	J	$Y_{n_{J-1}+1,J}$	$X_{1n_{J-1}+1,J}$	$X_{2n_{J-1}+1,J}$		$X_{Qn_{J-1}+1,J}$	W_{1J}	W_{2J}		W_{SJ}
$n_{J-1} + 2$	J	$Y_{n_{J-1}+2,J}$	$X_{1n_{J-1}+2,J}$	$X_{2n_{J-1}+2,J}$		$X_{Qn_{J-1}+2,J}$	W_{1J}	W_{2J}		W_{SJ}
⋮	⋮	⋮	⋮	⋮		⋮	⋮	⋮		⋮
n	J	Y_{nJ}	X_{1nJ}	X_{2nJ}		Z_{QnJ}	W_{1J}	W_{2J}		W_{SJ}

Com base na Tabela 16.1, podemos verificar que $X_1, ..., X_Q$ são variáveis de nível 1 (dados alteram-se entre indivíduos) e $W_1, ..., W_S$ são variáveis de nível 2 (dados alteram-se entre grupos, porém não para os indivíduos de cada grupo). Além disso, as quantidades de indivíduos nos grupos 1, 2, ..., J são iguais, respectivamente, a n_1, $n_2 - n_1$, ..., $n - n_{J-1}$. A Figura 16.1 permite que visualizemos o aninhamento existente entre as unidades do nível 1 (indivíduos) e as unidades do nível 2 (grupos), o que caracteriza a existência de dados agrupados.

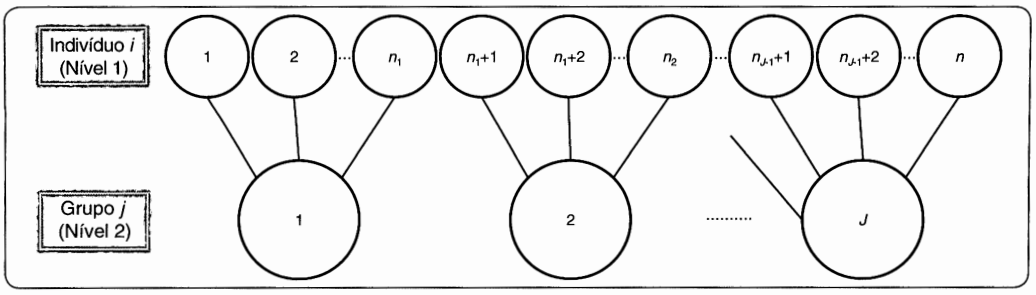


Figura 16.1 Estrutura aninhada de dados agrupados em dois níveis.

Caso $n_1 = n_2 - n_1 = ... = n - n_{J-1}$, teremos uma **estrutura equilibrada** de dados aninhados.

Imagine ainda outra base com dados em que, além do aninhamento apresentado para dados agrupados, há a evolução temporal, ou seja, dados com medidas repetidas. Logo, além dos indivíduos, que passarão a pertencer ao nível 2 e, portanto, serão nomeados de $j = 1, ..., J$, aninhados nos $k = 1, ..., K$ grupos (agora pertencentes ao nível 3), teremos também $t = 1, ..., T_j$ períodos em que cada indivíduo j é monitorado. Logo, este novo banco de dados pode apresentar as mesmas variáveis explicativas $X_1, ..., X_Q$ referentes a cada indivíduo j, porém agora invariantes para cada indivíduo j nos períodos de monitoramento. Além disso, pode também apresentar as mesmas variáveis explicativas $W_1, ..., W_S$ referentes a cada grupo k, porém também invariantes ao longo do tempo para cada grupo k. A Tabela 16.2 oferece a lógica com que se apresenta uma base com **estrutura aninhada de dados com medidas repetidas em três níveis** (tempo, indivíduo e grupo).

Tabela 16.2 Modelo geral de uma base com estrutura aninhada de dados com medidas repetidas em três níveis.

Período t (Medida Repetida) Nível 1	**Observação (Indivíduo j) Nível 2**	**Grupo k Nível 3**	Y_{ijk}	X_{1jk}	X_{2jk}	...	X_{Qjk}	W_{1k}	W_{2k}	...	W_{Sk}
1	1	1	Y_{111}	X_{111}	X_{211}		X_{Q11}	W_{11}	W_{21}		W_{S1}
2	1	1	Y_{211}	X_{111}	X_{211}		X_{Q11}	W_{11}	W_{21}		W_{S1}
⋮	⋮		⋮	⋮	⋮		⋮				
T_1	1		$Y_{T_1 11}$	X_{111}	X_{211}		X_{Q11}				
$T_1 + 1$	2	⋮	$Y_{T_1+1,21}$	X_{121}	X_{221}		X_{Q21}	⋮	⋮		⋮
$T_1 + 2$	2		$Y_{T_1+2,21}$	X_{121}	X_{221}		X_{Q21}				
⋮	⋮		⋮	⋮	⋮	...	⋮			...	
T_2	2	1	$Y_{T_2 21}$	X_{121}	X_{221}		X_{Q21}	W_{11}	W_{21}		W_{S1}
⋮	⋮	⋮	⋮	⋮	⋮		⋮	⋮	⋮		⋮
$T_{J-1} + 1$	J	K	$Y_{T_{J-1}+1,JK}$	X_{1JK}	X_{2JK}		X_{QJK}	W_{1K}	W_{2K}		W_{SK}
$T_{J-1} + 2$	J	K	$Y_{T_{J-1}+2,JK}$	X_{1JK}	X_{2JK}		X_{QJK}	W_{1K}	W_{2K}		W_{SK}
⋮	⋮	⋮	⋮	⋮	⋮		⋮	⋮	⋮		⋮
T_J	J	K	$Y_{T_J JK}$	X_{1JK}	X_{2JK}		X_{QJK}	W_{1K}	W_{2K}		W_{SK}

Com base na estrutura da Tabela 16.2, podemos verificar agora que a variável correspondente ao período de tempo é uma variável explicativa de nível 1, visto que os dados alteram-se em cada linha da base, e que $X_1, ..., X_Q$ passam a ser variáveis de nível 2 (dados alteram-se entre indivíduos, porém não para um mesmo indivíduo ao longo do tempo) e $W_1, ..., W_S$ passam a ser variáveis de nível 3 (dados alteram-se entre grupos, porém não para um mesmo grupo ao longo do tempo). Além disso, as quantidades de períodos em que os indivíduos $1, 2, ..., J$ são monitorados são iguais, respectivamente, a T_1, $T_2 - T_1, ..., T_J - T_{J-1}$. A Figura 16.2, de maneira análoga ao exposto para o caso com dois níveis, permite que visualizemos o aninhamento existente entre as unidades do nível 1 (variação temporal), as unidades do nível 2 (indivíduos) e as unidades do nível 3 (grupos), o que acaba por caracterizar uma estrutura de dados com medidas repetidas.

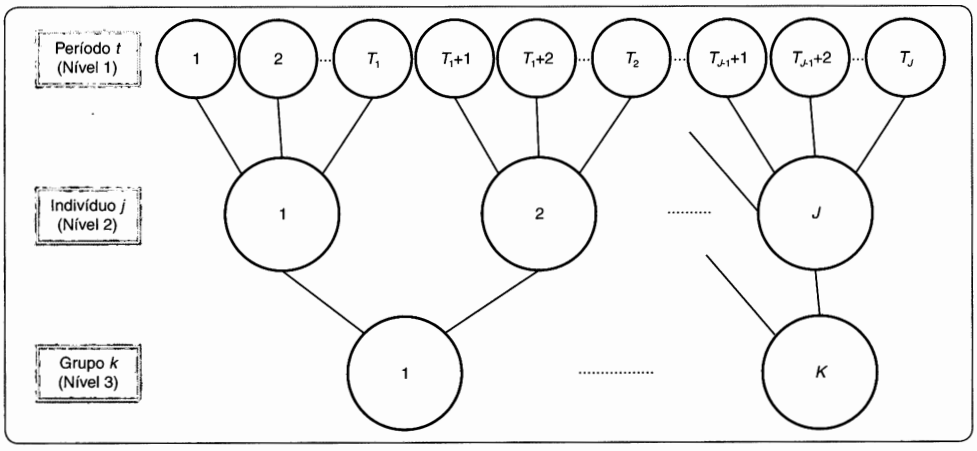


Figura 16.2 Estrutura aninhada de dados com medidas repetidas em três níveis.

Caso $T_1 = T_2 - T_1 = ... = T_J - T_{J-1}$, teremos um **painel balanceado**.

Podemos verificar, pelas Tabelas 16.1 e 16.2, bem como nas correspondentes Figuras 16.1 e 16.2, que as estruturas de dados apresentam **aninhamento absoluto**, ou seja, determinado indivíduo encontra-se aninhado a apenas um grupo, este a apenas outro grupo e assim sucessivamente. Entretanto, podem existir estruturas de dados em **aninhamento com classificação cruzada**, em que determinadas observações de um grupo podem fazer parte de um grupo em nível superior, com as demais fazendo parte de outro grupo em nível superior. Por exemplo, imagine o estudo do desempenho de firmas aninhadas em setores e em países. Podem existir, por exemplo, firmas atuantes em mineração e provenientes do Brasil, e outras atuantes em aviação e também provenientes do Brasil. Entretanto, caso haja na base, por exemplo, firmas mineradoras provenientes da Austrália, passa a ser caracterizado o aninhamento com classificação cruzada, fazendo-se necessária a estimação de **modelos hierárquicos com classificação cruzada** (em inglês, ***hierarchical cross-classifed models*** - ***HCM***). Estes modelos não são objeto da presente edição do livro, porém um pesquisador mais interessado poderá estudá-los em profundidade em Raudenbush e Bryk (2002), Raudenbush *et al.* (2004) e Rabe-Hesketh e Skrondal (2012a, 2012b).

Enquanto nas seções 16.4.1 e 16.5.1 estimaremos modelos hierárquicos lineares de dois níveis com dados agrupados (HLM2) em Stata e SPSS, respectivamente, as seções 16.4.2 e 16.5.2 são destinadas à estimação de modelagens hierárquicas lineares de três níveis com medidas repetidas (HLM3) nos mesmos softwares. Antes disso, porém, é necessário que sejam apresentadas e discutidas, na próxima seção, as formulações algébricas de cada um destes modelos.

16.3. MODELOS HIERÁRQUICOS LINEARES

Nesta seção, apresentaremos as formulações algébricas e as especificações dos modelos hierárquicos lineares de dois níveis com dados agrupados (seção 16.3.1) e dos modelos hierárquicos lineares de três níveis com medidas repetidas (seção 16.3.2).

16.3.1. Modelos hierárquicos lineares de dois níveis com dados agrupados (HLM2)

A fim de compreendermos como é definida a expressão geral de um modelo hierárquico linear com dados agrupados em dois níveis, precisamos usar um modelo de regressão linear múltipla, cuja especificação, baseada na expressão (12.1), é apresentada a seguir:

$$Y_i = b_0 + b_1.X_{1i} + b_2.X_{2i} + ... + b_Q.X_{Qi} + r_i \tag{16.1}$$

em que Y representa o fenômeno em estudo (variável dependente), b_0 representa o intercepto, b_1, b_2, ..., b_Q são os coeficientes de cada variável, X_1, ..., X_Q são variáveis explicativas (métricas ou *dummies*) e r representa os termos de erro. Os subscritos i representam cada uma das observações da amostra em análise ($i = 1, 2, ..., n$, em que n é o tamanho da amostra). Note que alguns termos apresentam nomenclatura diferente daquela proposta no Capítulo 12 (por exemplo, os termos de erro), já que outro nível de análise será considerado para a definição da modelagem hierárquica.

O modelo representado pela expressão (16.1) apresenta observações consideradas homogêneas, ou seja, não provenientes de grupos distintos que poderiam, por alguma razão, influenciar diferentemente o comportamento da variável *Y*. Entretanto, poderíamos pensar em dois grupos de observações, a partir dos quais seriam estimados dois modelos diferentes, conforme segue:

$$Y_{i1} = b_{01} + b_{11}.X_{1i1} + b_{21}.X_{2i1} + ... + b_{Q1}.X_{Qi1} + r_{i1} \tag{16.2}$$

$$Y_{i2} = b_{02} + b_{12}.X_{1i2} + b_{22}.X_{2i2} + ... + b_{Q2}.X_{Qi2} + r_{i2} \tag{16.3}$$

em que os coeficientes b_{01} e b_{02} representam, respectivamente, os valores médios esperados de *Y* para as observações dos grupos 1 e 2, quando todas as variáveis explicativas forem iguais a zero, e $b_{11}, b_{21}, ..., b_{Q1}$ e $b_{12}, b_{22}, ..., b_{Q2}$ são, respectivamente, os coeficientes das variáveis $X_1, ..., X_Q$ no modelo de cada grupo (1 e 2). Além disso, r_1 e r_2 representam os termos específicos de erro em cada modelo.

Portanto, para $j = 1, ..., J$ grupos, podemos escrever a expressão geral de um modelo de regressão para dados agrupados, considerado um **modelo de primeiro nível**, da seguinte forma:

$$\begin{aligned} Y_{ij} &= b_{0j} + b_{1j}.X_{1ij} + b_{2j}.X_{2ij} + ... + b_{Qj}.X_{Qij} + r_{ij} \\ &= b_{0j} + \sum_{q=1}^{Q} b_{qj}.X_{qij} + r_{ij} \end{aligned} \tag{16.4}$$

Com objetivos didáticos e para fins de elaboração de um gráfico ilustrativo, podemos escrever a expressão dos valores esperados de *Y*, ou seja, $\hat{Y}$, para cada observação *i* pertencente a cada grupo *j*, quando houver apenas uma variável explicativa *X* no modelo proposto, da seguinte forma:

Grupo 1:
$$\hat{Y}_{i1} = \beta_{01} + \beta_{11}.X_{i1} \tag{16.5}$$

Grupo 2:
$$\hat{Y}_{i2} = \beta_{02} + \beta_{12}.X_{i2} \tag{16.6}$$

$$\vdots$$

Grupo *J*:
$$\hat{Y}_{iJ} = \beta_{0J} + \beta_{1J}.X_{iJ} \tag{16.7}$$

em que os parâmetros β são as estimações dos coeficientes *b*, seguindo o padrão adotado no livro.

O gráfico da Figura 16.3 apresenta, de maneira conceitual, a plotagem das expressões (16.5) a (16.7) e, por meio dele, verificamos que os modelos individuais que representam as observações de cada grupo podem apresentar interceptos e inclinações diferentes, fato que pode ocorrer em função de determinadas características dos próprios grupos.

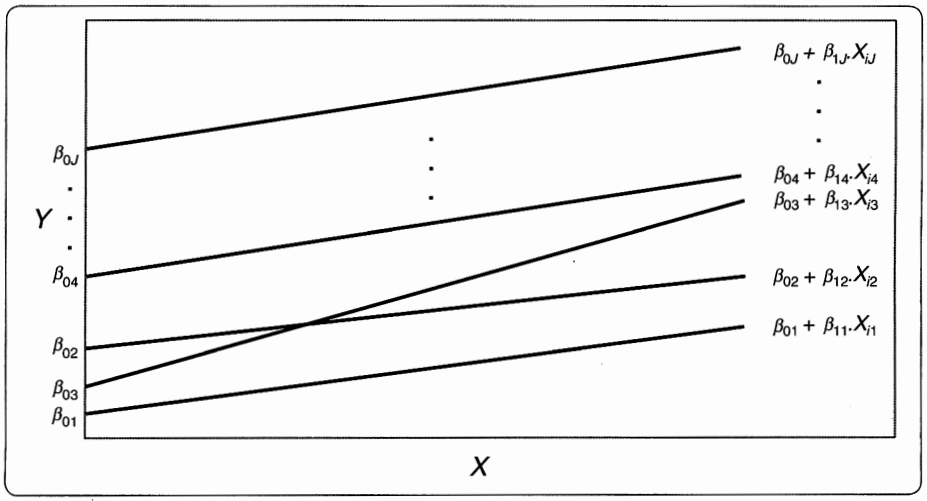


Figura 16.3 Modelos individuais que representam as observações de cada um dos *J* grupos.

Logo, devem existir características de grupos (**segundo nível**), invariantes para as observações pertencentes a cada grupo (conforme explicita a Tabela 16.1), que podem explicar as diferenças nos interceptos e nas inclinações dos modelos que representam esses grupos. Neste sentido, com base no seguinte modelo de regressão com uma variável explicativa X e com observações aninhadas em $j = 1, ..., J$ grupos:

$$Y_{ij} = b_{0j} + b_{1j}.X_{ij} + r_{ij} \tag{16.8}$$

podemos escrever, da seguinte forma, as expressões dos interceptos b_{0j} e das inclinações b_{1j} em função de determinada variável explicativa W, que representa uma característica dos j grupos:

Interceptos:

Grupo 1: $$b_{01} = \gamma_{00} + \gamma_{01}.W_1 + u_{01} \tag{16.9}$$

Grupo 2: $$b_{02} = \gamma_{00} + \gamma_{01}.W_2 + u_{02} \tag{16.10}$$

$$\vdots$$

Grupo *J*: $$b_{0J} = \gamma_{00} + \gamma_{01}.W_J + u_{0J} \tag{16.11}$$

ou, de maneira geral:

$$b_{0j} = \gamma_{00} + \gamma_{01}.W_j + u_{0j} \tag{16.12}$$

em que γ_{00} representa o valor esperado da variável dependente para determinada observação i pertencente a um grupo j quando $X = W = 0$ (intercepto geral), e γ_{01} representa a alteração no valor esperado da variável dependente para determinada observação i pertencente a um grupo j quando houver uma alteração unitária na característica W do grupo j, *ceteris paribus*. Além disso, u_{0j} representa os termos de erro que indicam a **existência de aleatoriedade nos interceptos** que pode ser gerada pela presença de observações provenientes de grupos distintos na base de dados.

Inclinações:

Grupo 1: $$b_{11} = \gamma_{10} + \gamma_{11}.W_1 + u_{11} \tag{16.13}$$

Grupo 2: $$b_{12} = \gamma_{10} + \gamma_{11}.W_2 + u_{12} \tag{16.14}$$

$$\vdots$$

Grupo *J*: $$b_{1J} = \gamma_{10} + \gamma_{11}.W_J + u_{1J} \tag{16.15}$$

ou, de maneira geral:

$$b_{1j} = \gamma_{10} + \gamma_{11}.W_j + u_{1j} \tag{16.16}$$

em que γ_{10} representa a alteração no valor esperado da variável dependente para determinada observação i pertencente a um grupo j quando houver uma alteração unitária na característica X do indivíduo i, *ceteris paribus* (mudança na inclinação em razão de X), e γ_{11} representa a alteração no valor esperado da variável dependente para determinada observação i pertencente a um grupo j quando houver uma alteração unitária no produto $W.X$, também *ceteris paribus* (mudança na inclinação em razão de $W.X$). Além disso, u_{1j} representa os termos de erro que indicam a **existência de aleatoriedade nas inclinações** dos modelos referentes aos grupos, que também pode ser gerada pela presença de observações provenientes de grupos distintos na base de dados.

Combinando as expressões (16.8), (16.12) e (16.16), chegamos à seguinte expressão:

$$Y_{ij} = \underbrace{\left(\gamma_{00} + \gamma_{01}.W_j + u_{0j}\right)}_{\substack{\textbf{intercepto com} \\ \textbf{efeitos aleatórios}}} + \underbrace{\left(\gamma_{10} + \gamma_{11}.W_j + u_{1j}\right)}_{\substack{\textbf{inclinação com} \\ \textbf{efeitos aleatórios}}}.X_{ij} + r_{ij} \tag{16.17}$$

que facilita a visualização de que o intercepto e a inclinação podem sofrer influência de termos aleatórios decorrentes da existência de observações pertencentes a grupos distintos.

Em essência, a modelagem multinível representa, portanto, um conjunto de técnicas que, além de estimarem os parâmetros do modelo proposto, permitem que sejam **estimados os componentes de variância dos termos de erro** (por exemplo, no modelo da expressão (16.17), u_{0j}, u_{1j} e r_{ij}), **bem como as respectivas significâncias estatísticas**, a fim de que se verifique, de fato, se ocorrem aleatoriedades nos interceptos e nas inclinações oriundas da presença de níveis superiores na análise. **Caso não se verifique a significância estatística das variâncias dos termos de erro u_{0j} e u_{1j}** no modelo da expressão (16.17), ou seja, se ambas forem estatisticamente iguais a zero, **passa a ser adequada a estimação de um modelo de regressão linear por meio de métodos tradicionais, como o MQO**, visto que não se comprova a existência de aleatoriedades nos interceptos e nas inclinações.

Podemos assumir que os efeitos aleatórios u_{0j} e u_{1j} apresentam distribuição normal multivariada, possuem médias iguais a zero e variâncias iguais, respectivamente, a τ_{00} e τ_{11}. Além disso, os termos de erro r_{ij} apresentam distribuição normal, com média igual a zero e variância igual a σ^2. Logo, podemos definir as seguintes matrizes de variância-covariância dos termos de erro:

$$var[\mathbf{u}] = var\begin{bmatrix} u_{0j} \\ u_{1j} \end{bmatrix} = \mathbf{G} = \begin{bmatrix} \tau_{00} & \sigma_{01} \\ \sigma_{01} & \tau_{11} \end{bmatrix} \tag{16.18}$$

$$var[\mathbf{r}] = var\begin{bmatrix} r_{1j} \\ \vdots \\ r_{nj} \end{bmatrix} = \sigma^2.\mathbf{I}_n = \begin{bmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \sigma^2 \end{bmatrix} \tag{16.19}$$

Essas matrizes serão utilizadas na apresentação, logo em seguida, dos métodos de estimação dos parâmetros de um modelo multinível.

Fazendo uso da expressão (15.19), podemos, portanto, definir a relação entre as variâncias destes termos de erro, conhecida por **correlação intraclasse**, conforme segue:

$$rho = \frac{\tau_{00} + \tau_{11}}{\tau_{00} + \tau_{11} + \sigma^2} \tag{16.20}$$

Essa correlação intraclasse mede a proporção de variância total que é devida aos níveis 1 e 2. Caso seja igual a zero, não ocorre variância dos indivíduos entre os grupos do nível 2. Entretanto, se for consideravelmente diferente de zero pela presença de ao menos um termo de erro significante decorrente da presença do nível 2 na análise, procedimentos tradicionais de estimação dos parâmetros do modelo, como mínimos quadrados ordinários, não são adequados. No limite, o fato de ser igual a 1, ou seja, $\sigma^2 = 0$, indica que não existem diferenças entre os indivíduos, isto é, todos são idênticos, o que é muito pouco provável de acontecer. Essa correlação é também chamada de **correlação intraclasse de nível 2**.

Na seção 16.4.1 faremos uso de **testes de razão de verossimilhança** com o intuito de verificar se $\tau_{00} = \tau_{11} = 0$, o que favoreceria a estimação de um modelo tradicional de regressão, ou ao menos se $\tau_{11} = 0$, o que permitiria que o pesquisador optasse por um **modelo com interceptos aleatórios** ($\tau_{00} \neq 0$) em vez de um **modelo com inclinações aleatórias** ($\tau_{11} \neq 0$).

Podemos rearranjar a expressão (16.17), para separar o componente de efeitos fixos, no qual são estimados os parâmetros do modelo, do componente de efeitos aleatórios, a partir do qual são estimadas as variâncias dos termos de erro. Assim, temos que:

$$Y_{ij} = \underbrace{\gamma_{00} + \gamma_{10}.X_{ij} + \gamma_{01}.W_j + \gamma_{11}.W_j.X_{ij}}_{\textbf{Efeitos Fixos}} + \underbrace{u_{0j} + u_{1j}.X_{ij} + r_{ij}}_{\textbf{Efeitos Aleatórios}} \tag{16.21}$$

que permite que o pesquisador visualize mais facilmente que o componente de efeitos aleatórios também pode influenciar o comportamento da variável dependente. Podemos notar, inclusive, que uma variável explicativa pode fazer parte deste componente aleatório. Estimando um modelo multinível como este, verificaremos que, enquanto os efeitos fixos referem-se à relação entre o comportamento de determinadas características e o comportamento de Y, os efeitos aleatórios permitem que se analisem eventuais distorções no comportamento de Y entre as unidades do segundo nível de análise.

De maneira geral, e partindo-se da expressão (16.4), podemos definir, da seguinte maneira, um modelo com dois níveis de análise, em que o primeiro nível oferece as variáveis explicativas $X_1, ..., X_Q$ referentes a cada indivíduo i, e o segundo nível, as variáveis explicativas $W_1, ..., W_S$ referentes a cada grupo j:

Nível 1:
$$Y_{ij} = b_{0j} + \sum_{q=1}^{Q} b_{qj}.X_{qij} + r_{ij} \tag{16.22}$$

Nível 2:
$$b_{qj} = \gamma_{q0} + \sum_{s=1}^{S_q} \gamma_{qs}.W_{sj} + u_{qj} \tag{16.23}$$

em que $q = 0, 1, ..., Q$ e $s = 1, ..., S_q$.

Em relação à estimação do modelo, enquanto os parâmetros dos efeitos fixos são estimados tradicionalmente, em softwares como o Stata e SPSS, por **máxima verossimilhança** (ou, em inglês, ***maximum likelihood estimation - MLE***), assim como realizado ao longo dos capítulos anteriores, os componentes de variância dos termos de erro podem ser estimados tanto por máxima verossimilhança, quanto por **máxima verossimilhança restrita** (ou, em inglês, ***restricted estimation of maximum likelihood - REML***).

As estimações dos parâmetros por *MLE* ou por *REML* são computacionalmente intensas, razão pela qual não as elaboraremos algebricamente neste capítulo, como fizemos em capítulos anteriores, na aplicação de exemplos práticos. Entretanto, ambas exigem a otimização de determinada função-objetivo, que geralmente parte de valores iniciais dos parâmetros e usa uma sequência de iterações para encontrar os parâmetros que maximizam a função de verossimilhança previamente definida.

A fim de introduzirmos especificamente os conceitos pertinentes ao método *REML*, vamos imaginar, por exemplo, um modelo de regressão apenas com a constante, sendo Y_i ($i = 1, ..., n$) uma variável dependente com distribuição normal, média μ e variância σ^2_Y. Enquanto a estimação por máxima verossimilhança de σ^2_Y é obtida considerando os n termos $Y_i - \mu$, a estimação de σ^2_Y por *REML* é obtida a partir dos $(n - 1)$ primeiros termos de $Y_i - \overline{Y}_i$, cuja distribuição independente de μ. Em outras palavras, a elaboração de um método de máxima verossimilhança a esta última distribuição gera uma estimação não viesada de σ^2_Y, por esta ser a própria variância amostral obtida pela divisão dos elementos por $(n - 1)$. Esta é a razão da estimação por máxima verossimilhança restrita também ser conhecida como estimação por **máxima verossimilhança reduzida**.

Para apresentarmos as expressões das funções de verossimilhança e de verossimilhança restrita a partir das quais, por maximização, os parâmetros de um modelo multinível podem ser estimados, vamos escrever, em notação matricial, a expressão geral de um modelo multinível com efeitos fixos e aleatórios da seguinte forma:

$$\mathbf{Y} = \mathbf{A}.\boldsymbol{\gamma} + \mathbf{B}.\mathbf{u} + \mathbf{r} \tag{16.24}$$

em que $\mathbf{Y}$ é um vetor $n \times 1$ que representa a variável dependente, $\mathbf{A}$ é uma matriz $n \times (q + s + q \cdot s + 1)$ com dados de todas as variáveis a serem inseridas no componente de efeitos fixos do modelo, $\boldsymbol{\gamma}$ é um vetor $(q + s + q \cdot s + 1) \times 1$ com todos os parâmetros de efeitos fixos estimados, $\mathbf{B}$ é a matriz $n \times (q + 1)$ com dados de todas as variáveis a serem inseridas no componentes de efeitos aleatórios $\mathbf{u}$, sendo $\mathbf{u}$ um vetor de termos aleatórios de erro com dimensões $(q + 1) \times 1$ e com matriz de variância-covariância $\mathbf{G}$. Além disso, $\mathbf{r}$ é um vetor $n \times 1$ de termos de erro com média zero e matriz de variância $\sigma^2 \cdot \mathbf{I}_n$. Com base nas expressões (16.18) e (16.19), podemos definir que:

$$var\begin{bmatrix}\mathbf{u}\\ \mathbf{r}\end{bmatrix} = \begin{bmatrix}\mathbf{G} & \mathbf{0}\\ \mathbf{0} & \sigma^2.\mathbf{I}_n\end{bmatrix} \tag{16.25}$$

e, neste sentido, a matriz de variância-covariância $n \times n$ de $\mathbf{Y}$, dada por $\mathbf{V}$, pode ser obtida da seguinte forma:

$$\mathbf{V}=\mathbf{B}.\mathbf{G}.\mathbf{B}'+\sigma^2.\mathbf{I}_n \tag{16.26}$$

A partir dessa matriz, conforme demonstram Searle, Casella e McCulloch (2006), pode ser definida a seguinte expressão do logaritmo da função de verossimilhança, que deve ser maximizada (*MLE*):

$$LL=-\frac{1}{2}.\left[n.\ln(2.\pi)+\ln|\mathbf{V}|+(\mathbf{Y}-\mathbf{A}.\gamma)'.\mathbf{V}^{-1}.(\mathbf{Y}-\mathbf{A}.\gamma)\right] \tag{16.27}$$

Ainda segundo os mesmos autores, a expressão do logaritmo da função de verossimilhança restrita é dada, a partir da expressão (16.27), por:

$$LL_r = LL-\frac{1}{2}.\ln\left|\mathbf{A}'.\mathbf{V}^{-1}.\mathbf{A}\right| \tag{16.28}$$

O fato de o método *REML* gerar estimações não viesadas das variâncias dos termos de erro em modelos multinível pode fazer com que o pesquisador opte incondicionalmente por seu uso. Entretanto, **os testes de razão de verossimilhança baseados nas estimações obtidas por *REML* não são apropriados para se compararem modelos com diferentes especificações dos efeitos fixos** e, para essas situações em que há o intuito de se elaborarem tais testes, recomendamos que as variâncias dos termos de erro sejam estimadas por *MLE*, já que é o método utilizado para a estimação dos parâmetros do modelo. Além disso, é importante comentar que as diferenças entre as estimações das variâncias dos termos de erro obtidas por *REML* ou por *MLE* são praticamente inexistentes para grandes amostras.

Na próxima seção, apresentaremos a especificação dos modelos hierárquicos lineares de três níveis com medidas repetidas, mantendo a lógica proposta.

16.3.2. Modelos hierárquicos lineares de três níveis com medidas repetidas (HLM3)

Seguindo a lógica proposta na seção anterior, vamos apresentar a especificação de um modelo hierárquico linear de três níveis, em que há a presença de dados com medidas repetidas, ou seja, com evolução temporal na variável dependente.

De maneira geral, e seguindo a lógica apresentada em Raudenbush *et al.* (2004), um modelo hierárquico de três níveis apresenta três submodelos, sendo um para cada nível de análise da estrutura aninhada de dados. Logo, com base nas expressões (16.22) e (16.23), podemos definir, da seguinte maneira, um modelo geral de três níveis de análise com dados aninhados, em que o primeiro nível apresenta as variáveis explicativas $Z_1, ..., Z_P$ referentes às unidades i ($i = 1, ..., n$) de nível 1, o segundo nível, as variáveis explicativas $X_1, ..., X_Q$ referentes às unidades j ($j = 1, ..., J$) de nível 2, e o terceiro nível, as variáveis explicativas $W_1, ..., W_S$ referentes às unidades k ($k = 1, ..., K$) de nível 3:

Nível 1:

$$Y_{ijk}=\pi_{0jk}+\sum_{p=1}^{P}\pi_{pjk}.Z_{pjk}+e_{ijk} \tag{16.29}$$

em que π_{pjk} ($p = 0, 1, ..., P$) referem-se aos coeficientes de nível 1, Z_{pjk} é uma p-ésima variável explicativa de nível 1 para a observação i na unidade de nível 2 j e na unidade de nível 3 k, e e_{ijk} refere-se aos termos de erro do nível 1 com distribuição normal, com média igual a zero e variância igual a σ^2.

Nível 2:

$$\pi_{pjk}=b_{p0k}+\sum_{q=1}^{Q_p}b_{pqk}.X_{qjk}+r_{pjk} \tag{16.30}$$

em que b_{pqk} ($q = 0, 1, ..., Q_p$) referem-se aos coeficientes de nível 2, X_{qjk} é uma q-ésima variável explicativa de nível 2 para a unidade j na unidade de nível 3 k, e r_{pjk} são os efeitos aleatórios do nível 2, assumindo-se, para cada unidade j, que o vetor $(r_{0jk}, r_{1jk}, ..., r_{Pjk})'$ apresenta distribuição normal multivariada com cada elemento possuindo média zero e variância $\tau_{r\pi pp}$.

Nível 3:

$$b_{pqk} = \gamma_{pq0} + \sum_{s=1}^{S_{pq}} \gamma_{pqs} . W_{sk} + u_{pqk} \tag{16.31}$$

em que γ_{pqs} ($s = 0, 1, ..., S_{pq}$) referem-se aos coeficientes de nível 3, W_{sk} é uma s-ésima variável explicativa de nível 3 para a unidade k, e u_{pqk} são os efeitos aleatórios do nível 3, assumindo-se que para cada unidade k, o vetor composto pelos termos u_{pqk} apresenta distribuição normal multivariada com cada elemento possuindo média zero e variância $\tau_{u\pi pp}$, que resulta na matriz de variância-covariância $\mathbf{T_b}$ com dimensão máxima igual a:

$$\text{Dim}_{\text{máx}}\, \mathbf{T_b} = \sum_{p=0}^{P} (Q_p + 1) . \sum_{p=0}^{P} (Q_p + 1) \tag{16.32}$$

que depende da quantidade de coeficientes do nível 3 especificados com termos aleatórios.

A fim de mantermos a lógica apresentada na seção anterior, e com o intuito de facilitar a compreensão do exemplo que será elaborado nas seções 16.4.2 e 16.5.2, imaginemos agora que exista uma única variável explicativa de nível 1, correspondente aos períodos de tempo em que são monitorados os dados da variável dependente. Em outras palavras, as unidades j do nível 2, aninhadas às unidades k do nível 3, são monitoradas por um período de tempo t ($t = 1, ..., T_j$), o que faz com que o banco de dados apresente j séries de tempo, conforme já mostrava a Tabela 16.2. O intuito é verificar se existem discrepâncias na evolução temporal dos dados da variável dependente e, em caso afirmativo, se essas ocorrem em função de características das unidades de nível 2 e de nível 3. Esta evolução temporal é o que caracteriza o termo **medidas repetidas**.

Neste sentido, a expressão (16.29) pode ser reescrita conforme segue, em que os subscritos i passam a ser subscritos t:

$$Y_{tjk} = \pi_{0jk} + \pi_{1jk} . \text{período}_{jk} + e_{tjk} \tag{16.33}$$

em que π_{0jk} representa o intercepto do modelo correspondente à evolução temporal da variável dependente da unidade j do nível 2 aninhada à unidade k do nível 3, e π_{1jk} corresponde à evolução média (inclinação) da variável dependente para a mesma unidade ao longo do período analisado. Os submodelos correspondentes aos níveis 2 e 3 permanecem com as mesmas especificações daquelas apresentadas, respectivamente, nas expressões (16.30) e (16.31).

O gráfico da Figura 16.4 apresenta, de maneira conceitual, a plotagem do conjunto de modelos representados pela expressão (16.33) e, por meio dele, verificamos que os modelos individuais que representam as unidades j do nível 2 podem apresentar interceptos e inclinações diferentes ao longo do período t, fato que pode ocorrer em função de determinadas características das próprias unidades j do nível 2, ou de características das unidades k do nível 3.

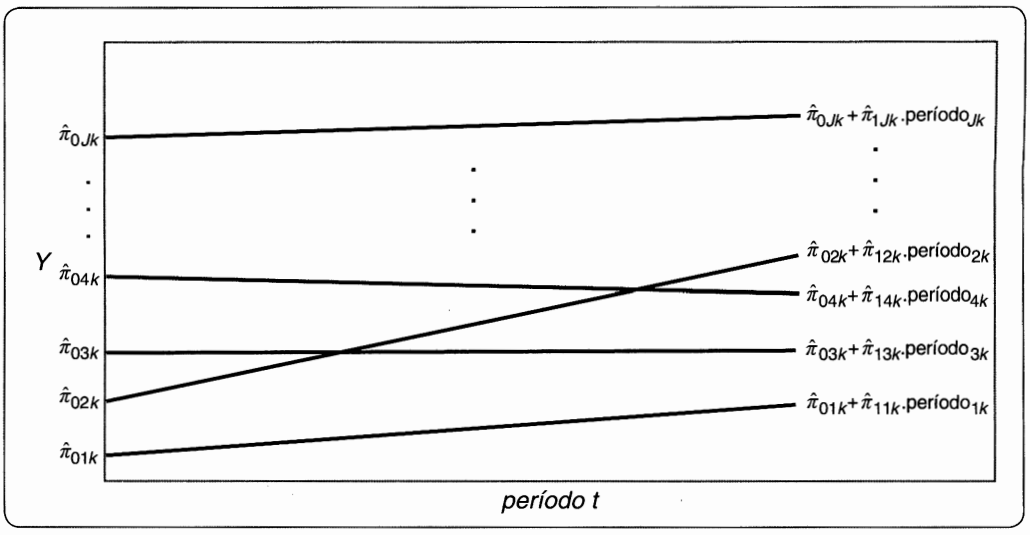


Figura 16.4 Modelos individuais que representam as evoluções temporais da variável dependente para cada uma das J unidades do nível 2.

Logo, devem existir características das unidades *j* do nível 2, invariantes temporalmente, e das unidades *k* do nível 3, invariantes também para as unidades *j* do nível 2 aninhadas a cada unidade *k* do nível 3 (conforme explícito na Tabela 16.2), que podem explicar as diferenças nos interceptos e nas inclinações dos modelos $\hat{Y}_{tjk} = \hat{\pi}_{0jk} + \hat{\pi}_{1jk} . \text{período}_{jk}$ representados na Figura 16.4.

Neste sentido, supondo existir uma única variável explicativa *X*, que representa uma característica das *j* unidades do nível 2, e uma única variável explicativa *W*, que representa uma característica das *k* unidades do nível 3, podemos definir, a partir da expressão (16.33) e com base nas expressões (16.30) e (16.31), o seguinte modelo com três níveis de análise, em que o primeiro nível refere-se à medida repetida e contém apenas a variável temporal:

Nível 1:

$$Y_{tjk} = \pi_{0jk} + \pi_{1jk} . \text{período}_{jk} + e_{tjk} \tag{16.34}$$

Nível 2:

$$\pi_{0jk} = b_{00k} + b_{01k} . X_{jk} + r_{0jk} \tag{16.35}$$

$$\pi_{1jk} = b_{10k} + b_{11k} . X_{jk} + r_{1jk} \tag{16.36}$$

Nível 3:

$$b_{00k} = \gamma_{000} + \gamma_{001} . W_k + u_{00k} \tag{16.37}$$

$$b_{01k} = \gamma_{010} + \gamma_{011} . W_k + u_{01k} \tag{16.38}$$

$$b_{10k} = \gamma_{100} + \gamma_{101} . W_k + u_{10k} \tag{16.39}$$

$$b_{11k} = \gamma_{110} + \gamma_{111} . W_k + u_{11k} \tag{16.40}$$

Combinando as expressões (16.34) a (16.39), chegamos à seguinte expressão:

$$Y_{tjk} = \underbrace{\left(\gamma_{000} + \gamma_{001} . W_k + \gamma_{010} . X_{jk} + \gamma_{011} . W_k . X_{jk} + u_{00k} + u_{01k} . X_{jk} + r_{0jk}\right)}_{\text{intercepto com efeitos aleatórios}} + \underbrace{\left(\gamma_{100} + \gamma_{101} . W_k + \gamma_{110} . X_{jk} + \gamma_{111} . W_k . X_{jk} + u_{10k} + u_{11k} . X_{jk} + r_{1jk}\right) . \text{período}_{jk}}_{\text{inclinação com efeitos aleatórios}} + e_{tjk} \tag{16.41}$$

em que γ_{000} representa o valor esperado da variável dependente no instante inicial e quando $X = W = 0$ (intercepto geral), γ_{001} representa o incremento no valor esperado da variável dependente no instante inicial (alteração no intercepto) para determinada unidade *j* de nível 2 pertencente a uma unidade *k* de nível 3 quando houver alteração unitária na característica *W* de *k*, *ceteris paribus*, γ_{010} representa o incremento no valor esperado da variável dependente no instante inicial para determinada unidade *jk* quando houver alteração unitária na característica *X* de *j*, *ceteris paribus*, e γ_{011} representa o incremento no valor esperado da variável dependente no instante inicial para determinada unidade *jk* quando houver alteração unitária no produto *W.X*, também *ceteris paribus*. Além disso, u_{00k} e u_{01k} representam os termos de erro que indicam a **existência de aleatoriedade nos interceptos**, sendo que o último incide sobre alterações na variável *X*.

Além disso, γ_{100} representa a alteração no valor esperado da variável dependente quando houver alteração unitária no período de análise (mudança na inclinação em razão da evolução temporal unitária), *ceteris paribus*, γ_{101} representa a alteração no valor esperado da variável dependente em razão da evolução temporal unitária para determinada unidade *jk* quando houver alteração unitária na característica *W*, *ceteris paribus*, γ_{110} representa a alteração no valor esperado da variável dependente em razão da evolução temporal unitária para determinada unidade *jk* quando houver alteração unitária na característica *X*, *ceteris paribus*, e γ_{111} representa a alteração no valor esperado da variável dependente em razão da evolução temporal unitária para determinada unidade *jk* quando houver alteração unitária no produto *W.X*, também *ceteris paribus*. Por fim, u_{10k} e u_{11k} representam os termos de erro que indicam a **existência de aleatoriedade nas inclinações**, sendo que o último também incide sobre alterações na variável *X*.

A expressão (16.41) facilita a visualização de que o intercepto e a inclinação podem sofrer influência de termos aleatórios decorrentes da existência de comportamentos distintos da variável dependente ao longo do

tempo para cada uma das unidades do nível 2 (distintas séries de tempo), e esse fenômeno pode ser decorrente das características dessas unidades, bem como das características dos grupos a que pertencem tais unidades.

Se o pesquisador desejar elaborar uma análise acerca dos componentes de efeitos fixos e aleatórios que podem influenciar o comportamento da variável dependente, dado que este procedimento inclusive facilita a inserção dos comandos para elaboração de modelagens multinível em Stata e em SPSS, conforme veremos mais adiante, basta rearranjar os termos da expressão (16.41), conforme segue:

$$\left.\begin{aligned} Y_{tjk} = {} & \gamma_{000} + \gamma_{001}.W_k + \gamma_{010}.X_{jk} + \gamma_{011}.W_k.X_{jk} \\ & + \gamma_{100}.\text{período}_{tjk} + \gamma_{101}.W_k.\text{período}_{jk} + \gamma_{110}.X_{jk}.\text{período}_{jk} + \gamma_{111}.W_k.X_{jk}.\text{período}_{jk} \end{aligned}\right\} \textbf{Efeitos Fixos}$$

$$\underbrace{+u_{00k} + u_{01k}.X_{jk} + u_{10k}.\text{período}_{jk} + u_{11k}.X_{jk}.\text{período}_{jk} + r_{0jk} + r_{1jk}.\text{período}_{jk} + e_{tjk}}_{\textbf{Efeitos Aleatórios}} \quad (16.42)$$

Em modelos hierárquicos de três níveis podemos definir duas correlações intraclasse, dada a existência de duas proporções de variância, sendo uma correspondente ao comportamento dos dados pertencentes às mesmas unidades j de nível 2 e mesmas unidades k de nível 3 (**correlação intraclasse de nível 2**), e outra correspondente ao comportamento dos dados pertencentes às mesmas unidades k de nível 3, porém provenientes de diferentes unidades j de nível 2 (**correlação intraclasse de nível 3**). Nas seções 16.4.2 e 16.5.2, elaboraremos os cálculos dessas correlações intraclasse quando da aplicação de exemplos práticos, respectivamente, em Stata e SPSS.

A partir da expressão (16.34), podemos definir, conforme segue, as expressões gerais dos submodelos de níveis 2 e 3 de uma análise hierárquica com três níveis e medidas repetidas, em que o segundo nível oferece as variáveis explicativas $X_1, ..., X_Q$ referentes a cada unidade j, e o terceiro nível, as variáveis explicativas $W_1, ..., W_S$ referentes a cada unidade k:

Nível 2:

$$\pi_{pjk} = b_{p0k} + \sum_{q=1}^{Q_p} b_{pqk}.X_{qjk} + r_{pjk} \quad (16.43)$$

Nível 3:

$$b_{pqk} = \gamma_{pq0} + \sum_{s=1}^{S_{pq}} \gamma_{pqs}.W_{sk} + u_{pqk} \quad (16.44)$$

Analogamente ao apresentado com os modelos hierárquicos de dois níveis na seção anterior, enquanto os parâmetros dos efeitos fixos são estimados tradicionalmente, em softwares como o Stata e o SPSS, por máxima verossimilhança, os componentes de variância dos termos de erro podem ser estimados tanto por máxima verossimilhança, quanto por máxima verossimilhança restrita, conforme veremos nas próximas seções quando estimarmos modelos hierárquicos de três níveis por meio desses softwares.

Frente ao exposto, enquanto na Seção 16.4 elaboraremos modelagens hierárquicas de dois níveis com dados agrupados e de três níveis com medidas repetidas em Stata, na seção 16.5 elaboraremos as mesmas modelagens, porém em SPSS. Os exemplos adotados respeitam a lógica adotada ao longo do livro.

16.4. ESTIMAÇÃO DE MODELOS HIERÁRQUICOS LINEARES NO SOFTWARE STATA

O objetivo desta seção é propiciar ao pesquisador uma oportunidade de elaboração de procedimentos de modelagem multinível por meio do Stata Statistical Software®. A reprodução das imagens nesta seção tem autorização da StataCorp LP©.

16.4.1. Estimação de um modelo hierárquico linear de dois níveis com dados agrupados no software Stata

Apresentaremos um exemplo que segue a mesma lógica dos capítulos anteriores, porém com dados que variam entre indivíduos e entre grupos a que pertencem esses indivíduos, caracterizando uma estrutura aninhada.

Imagine que o nosso sagaz e talentoso professor, que já explorou consideravelmente os efeitos de determinadas variáveis explicativas sobre o tempo de deslocamento de um grupo de alunos até a escola, sobre a probabilidade de se chegar atrasado às aulas, sobre a quantidade de atrasos que ocorrem semanal ou mensalmente e sobre o desempenho escolar desses alunos ao longo do tempo, por meio, respectivamente, de modelos de regressão

múltipla, de regressão logística binária e multinomial, de regressão para dados de contagem e de regressão com dados longitudinais, tenha agora o interesse em ampliar sua pesquisa para outras escolas, investigando se existem diferenças no comportamento do desempenho escolar entre estudantes provenientes de escolas distintas e, em caso afirmativo, se essas diferenças ocorrem em função de características das próprias escolas.

Neste sentido, o professor conseguiu dados sobre o desempenho escolar (nota de 0 a 100 mais um bônus por participação em sala) de 2.000 estudantes provenientes de 46 escolas. Além disso, também conseguiu dados a respeito do comportamento dos estudantes, como quantidade semanal de horas de estudo, e dados referentes à natureza de cada uma das escolas (pública ou privada) e ao tempo médio de experiência docente dos professores em cada uma delas. Parte do banco de dados elaborado encontra-se na Tabela 16.3, porém a base de dados completa pode ser acessada por meio dos arquivos **DesempenhoAlunoEscola.xls** (Excel) e **DesempenhoAlunoEscola.dta** (Stata).

Tabela 16.3 Exemplo: desempenho escolar e características de estudantes (nível 1) e de escolas (nível 2).

Estudante *i* (Nível 1)	Escola *j* (Nível 2)	Desempenho escolar (Y_{ij})	Quantidade semanal de horas de estudo (X_{ij})	Tempo médio, em anos, de experiência dos docentes (W_{1j})	Escola pública ou privada (W_{2j})
1	1	35,4	11	2	pública
2	1	74,9	23	2	pública
			...		
47	1	24,8	9	2	pública
48	2	41,0	13	2	pública
			...		
72	2	65,2	20	2	pública
			...		
121	4	66,4	20	9	privada
			...		
140	4	93,4	27	9	privada
			...		
1.995	46	44,0	15	2	pública
			...		
2.000	46	56,6	17	2	pública

Após abrirmos o arquivo **DesempenhoAlunoEscola.dta**, podemos digitar o comando **desc**, que faz com que seja possível analisarmos as características do banco de dados, como a quantidade de observações, a quantidade de variáveis e a descrição de cada uma delas. A Figura 16.5 apresenta este primeiro *output* do Stata.

```
. desc

  obs:         2,000
 vars:             6
 size:        42,000
-------------------------------------------------------------------------------
              storage  display     value
variable name   type   format      label      variable label
-------------------------------------------------------------------------------
estudante       int    %8.0g                  estudante i (nível 1)
escola          int    %8.0g                  escola j (nível 2)
desempenho      float  %9.1f                  desempenho escolar
horas           byte   %8.0g                  quantidade semanal de horas de estudo do
                                                aluno
texp            float  %9.0g                  tempo médio de experiência docente dos
                                                professores da escola (anos)
priv            float  %9.0g       priv       natureza da escola (pública ou privada)
-------------------------------------------------------------------------------
Sorted by:  estudante
```

Figura 16.5 Descrição do banco de dados **DesempenhoAlunoEscola.dta**.

Inicialmente, podemos obter informações acerca da quantidade de alunos que foram pesquisados pelo professor em cada escola, por meio do seguinte comando:

```
tabulate escola, subpop(estudante)
```

Os *outputs* são apresentados na Figura 16.6 e, por meio destes, podemos verificar que estamos diante de uma **estrutura desequilibrada de dados agrupados**.

```
. tabulate escola, subpop(estudante)

  escola j |
  (nível 2) |      Freq.     Percent        Cum.
------------+-----------------------------------
          1 |         47        2.35        2.35
          2 |         25        1.25        3.60
          3 |         48        2.40        6.00
          4 |         20        1.00        7.00
          5 |         48        2.40        9.40
          6 |         30        1.50       10.90
          7 |         28        1.40       12.30
          8 |         35        1.75       14.05
          9 |         44        2.20       16.25
         10 |         33        1.65       17.90
         11 |         57        2.85       20.75
         12 |         62        3.10       23.85
         13 |         53        2.65       26.50
         14 |         27        1.35       27.85
         15 |         53        2.65       30.50
         16 |         28        1.40       31.90
         17 |         29        1.45       33.35
         18 |         39        1.95       35.30
         19 |         47        2.35       37.65
         20 |         60        3.00       40.65
         21 |         61        3.05       43.70
         22 |         67        3.35       47.05
         23 |         47        2.35       49.40
         24 |         57        2.85       52.25
         25 |         52        2.60       54.85
         26 |         57        2.85       57.70
         27 |         38        1.90       59.60
         28 |         57        2.85       62.45
         29 |         42        2.10       64.55
         30 |         38        1.90       66.45
         31 |         52        2.60       69.05
         32 |         45        2.25       71.30
         33 |         47        2.35       73.65
         34 |         25        1.25       74.90
         35 |         55        2.75       77.65
         36 |         42        2.10       79.75
         37 |         43        2.15       81.90
         38 |         48        2.40       84.30
         39 |         46        2.30       86.60
         40 |         53        2.65       89.25
         41 |         59        2.95       92.20
         42 |         21        1.05       93.25
         43 |         39        1.95       95.20
         44 |         52        2.60       97.80
         45 |         38        1.90       99.70
         46 |          6        0.30      100.00
------------+-----------------------------------
      Total |      2,000      100.00
```

Figura 16.6 Quantidade de estudantes por escola.

O desempenho médio dos estudantes por escola, que pode ser analisado na Figura 16.7, pode ser obtido por meio dos seguintes comandos:

```
bysort escola: egen desempenho_médio = mean(desempenho)

tabstat desempenho_médio, by(escola)
```

```
. bysort escola: egen desempenho_médio = mean(desempenho)
. tabstat desempenho_médio, by(escola)

Summary for variables: desempenho_médio
     by categories of: escola (escola j (nível 2))

  escola |      mean            escola |      mean
---------+----------          ---------+----------
       1 |  50.38936                24 |  58.54211
       2 |    62.796                25 |  52.57116
       3 |  43.94375                26 |  67.31403
       4 |    75.025                27 |  62.13158
       5 |  56.23333                28 |  71.18597
       6 |  56.93667                29 |  41.76429
       7 |  51.73214                30 |  55.77369
       8 |  92.93143                31 |       57.9
       9 |  84.92728                32 |      60.86
      10 |  70.95454                33 |  75.65958
      11 |  66.56842                34 |    54.892
      12 |  64.72258                35 |  57.33636
      13 |  44.24151                36 |  62.98333
      14 |  42.73333                37 |  45.33023
      15 |  69.16415                38 |       89.3
      16 |  65.86072                39 |  51.07391
      17 |  74.81724                40 |  61.02641
      18 |  60.34103                41 |  59.88983
      19 |  58.83617                42 |   77.0619
      20 |      66.77                43 |  49.32564
      21 |  45.14262                44 |    61.125
      22 |  50.40448                45 |  63.06579
      23 |  71.09787                46 |      42.65
                                 ---------+----------
                                   Total |   60.8596
                                 ----------------------
```

Figura 16.7 Desempenho médio dos estudantes por escola.

E, para finalizarmos este diagnóstico inicial, podemos elaborar um gráfico que permite a visualização do desempenho médio dos estudantes por escola. Este gráfico, apresentado na Figura 16.8, e pode ser obtido pela digitação do seguinte comando:

```
graph twoway scatter desempenho escola || connected desempenho_médio
escola, connect(L) || , ytitle(desempenho escolar)
```

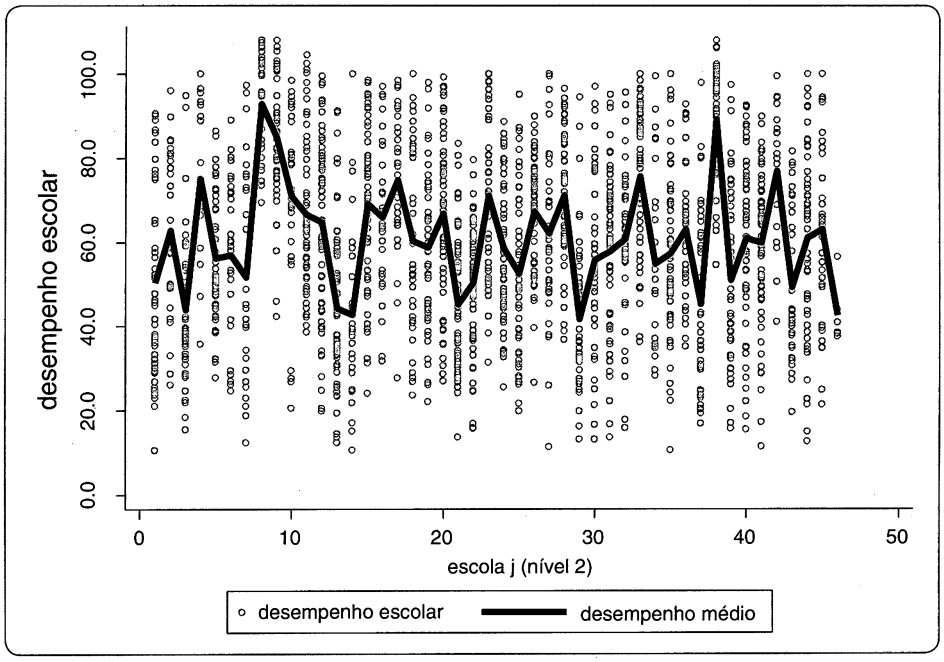


Figura 16.8 Desempenho escolar médio dos estudantes por escola.

Caracterizado o aninhamento dos estudantes em escolas com base nos dados agrupados do nosso exemplo, vamos partir para a modelagem multinível propriamente dita, elaborando os procedimentos com foco na estimação de um modelo hierárquico linear de dois níveis (estudantes e escolas). Na modelagem do desempenho

escolar, embora uma possibilidade seja a inclusão, no componente de efeitos fixos, de variáveis *dummy* que representem escolas, vamos tratar estas unidades de nível 2 como efeitos aleatórios para a estimação destes modelos.

O primeiro modelo a ser estimado, conhecido por **modelo nulo** ou **modelo não condicional**, permite que verifiquemos se existe variabilidade do desempenho escolar entre estudantes provenientes de escolas diferentes, já que nenhuma variável explicativa será inserida na modelagem, que considera apenas a existência de um intercepto e dos termos de erro u_{0j} e r_{ij}, com variâncias respectivamente iguais a τ_{00} e σ^2. O modelo a ser estimado, portanto, apresenta a seguinte expressão:

Modelo Nulo:

$$desempenho_{ij} = b_{0j} + r_{ij}$$
$$b_{0j} = \gamma_{00} + u_{0j}$$

que resulta em:

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

O comando para a estimação do modelo nulo no Stata, para os dados do nosso exemplo, é:

```
xtmixed desempenho || escola: , var nolog reml
```

em que o termo **xtmixed** refere-se à estimação de qualquer modelo hierárquico linear e a primeira variável a ser inserida corresponde à variável dependente, assim como em qualquer outra estimação de um modelo de regressão, com variáveis explicativas podendo ser incluídas em sequência. Além disso, há uma segunda parte do comando **xtmixed**, iniciada pelo termo **||**. Enquanto a primeira parte do comando corresponde aos efeitos fixos, a segunda parte diz respeito aos efeitos aleatórios que podem ser gerados pela existência de um segundo nível de análise, referente, no caso, às escolas (daí a segunda parte iniciar com o termo **escola:**). O termo **var** faz com que sejam apresentados, nos *outputs*, as estimações das variâncias dos termos de erro u_{0j} e r_{ij} (τ_{00} e σ^2, respectivamente), em vez dos desvios-padrão. Já o termo **nolog** apenas faz com que não sejam apresentados, nos *outputs*, os resultados das iterações para a maximização do logaritmo da função de verossimilhança restrita. Por fim, o pesquisador ainda tem a opção de definir o método de estimação a ser utilizado, usando os termos **reml** (máxima verossimilhança restrita) ou **mle** (máxima verossimilhança)[1].

Os *outputs* gerados estão na Figura 16.9.

```
. xtmixed desempenho || escola: , var nolog reml

Mixed-effects REML regression                   Number of obs      =      2000
Group variable: escola                          Number of groups   =        46

                                                Obs per group: min =         6
                                                               avg =      43.5
                                                               max =        67

                                                Wald chi2(0)       =         .
Log restricted-likelihood = -8752.0205          Prob > chi2        =         .

------------------------------------------------------------------------------
  desempenho |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _cons |   61.04901   1.776135    34.37   0.000     57.56785    64.53017
------------------------------------------------------------------------------

------------------------------------------------------------------------------
  Random-effects Parameters  |   Estimate   Std. Err.      [95% Conf. Interval]
-----------------------------+------------------------------------------------
escola: Identity             |
                  var(_cons) |   135.7793   30.75008       87.10859     211.644
-----------------------------+------------------------------------------------
               var(Residual) |   347.5617   11.12078       326.4347     370.056
------------------------------------------------------------------------------
LR test vs. linear regression: chibar2(01) =   486.01 Prob >= chibar2 = 0.0000
```

Figura 16.9 *Outputs* do modelo nulo no Stata.

[1] O comando **xtmixed** passou a estar disponível na versão 9 do Stata (a partir de 2005), e até a versão 12 é o comando para a estimação de modelos hierárquicos lineares, com método padrão de estimação por máxima verossimilhança restrita (*REML*). A partir da versão 13 do Stata, as estimações de modelos hierárquicos lineares podem ser elaboradas por meio dos comandos **xtmixed** ou simplesmente **mixed**, porém o método de estimação padrão, quando não especificado pelo pesquisador, passa a ser o de máxima verossimilhança (*MLE*).

A partir dos *outputs* da Figura 16.9, podemos incialmente verificar que a estimação do parâmetro γ_{00} é igual a 61,049, que corresponde à média dos desempenhos escolares esperados dos estudantes (reta horizontal estimada no modelo nulo, ou intercepto geral)[2]. Além disso, na parte inferior dos *outputs*, são apresentadas as estimações das variâncias dos termos de erro τ_{00} = 135,779 (no Stata, **var(_cons)**) e σ^2 = 347,562 (no Stata, **var(Residual)**). Com base na expressão (16.20), podemos calcular a seguinte correlação intraclasse:

$$rho = \frac{\tau_{00}}{\tau_{00} + \sigma^2} = \frac{135,779}{135,779 + 347,562} = 0,281$$

que indica que aproximadamente 28% da variância total do desempenho escolar é devido à alteração entre escolas, representando um primeiro indício de existência de variabilidade no desempenho escolar dos estudantes provenientes de escolas diferentes. A partir da versão 13 do Stata, é possível obter diretamente essa correlação intraclasse, digitando-se o comando **estat icc** logo após a estimação do correspondente modelo.

Embora o Stata não mostre diretamente o resultado dos testes z com os respectivos níveis de significância para os parâmetros de efeitos aleatórios, o fato de a estimação do componente de variância τ_{00}, correspondente ao intercepto aleatório u_{0j}, ser consideravelmente superior ao seu erro-padrão indica variação significante no desempenho escolar entre escolas. Estatisticamente, podemos verificar que z = 135,779 / 30,750 = 4,416 > 1,96, sendo 1,96 o valor crítico da distribuição normal padrão que resulta em um nível de significância de 5%.

Essa informação é bastante importante para embasar a escolha da modelagem hierárquica, em detrimento de uma modelagem tradicional de regressão por MQO, e é a principal razão para que seja estimado sempre um modelo nulo na elaboração de análises multinível.

Na parte inferior da Figura 16.9 podemos comprovar esse fato, analisando o resultado do teste de razão de verossimilhança (**LR test**, ou *likelihood ratio test*). Como *Sig.* χ^2 = 0,000, podemos rejeitar a hipótese nula de que os interceptos aleatórios sejam iguais a zero (H_0: u_{0j} = 0), o que faz com que a estimação de um modelo tradicional de regressão linear seja descartada para os dados agrupados do nosso exemplo.

Vamos primeiramente investigar se a variável explicativa de nível 1, *horas*, apresenta relação com o comportamento do desempenho escolar dos estudantes provenientes de uma mesma escola (variação entre estudantes) e provenientes de escolas distintas (variação entre escolas). Um primeiro diagnóstico pode ser elaborado por meio da digitação do seguinte comando, que gera o gráfico da Figura 16.10:

```
statsby intercept=_b[_cons] slope=_b[horas], by(escola) saving(ols,
replace): reg desempenho horas
sort escola
merge escola using ols
drop _merge
gen yhat_ols= intercept + slope*horas
sort escola horas
separate desempenho, by(escola)
separate yhat_ols, by(escola)
graph twoway connected yhat_ols1-yhat_ols46 horas || lfit desempenho
horas, clwidth(thick) clcolor(black) legend(off) ytitle(desempenho
escolar)
```

[2] Um pesquisador mais curioso poderá verificar este fato, digitando o comando **predict yhat** logo após a estimação do modelo nulo. Uma nova variável (*yhat*) será gerada no banco de dados, com todos os valores iguais a 61,049 (na realidade, uma constante).

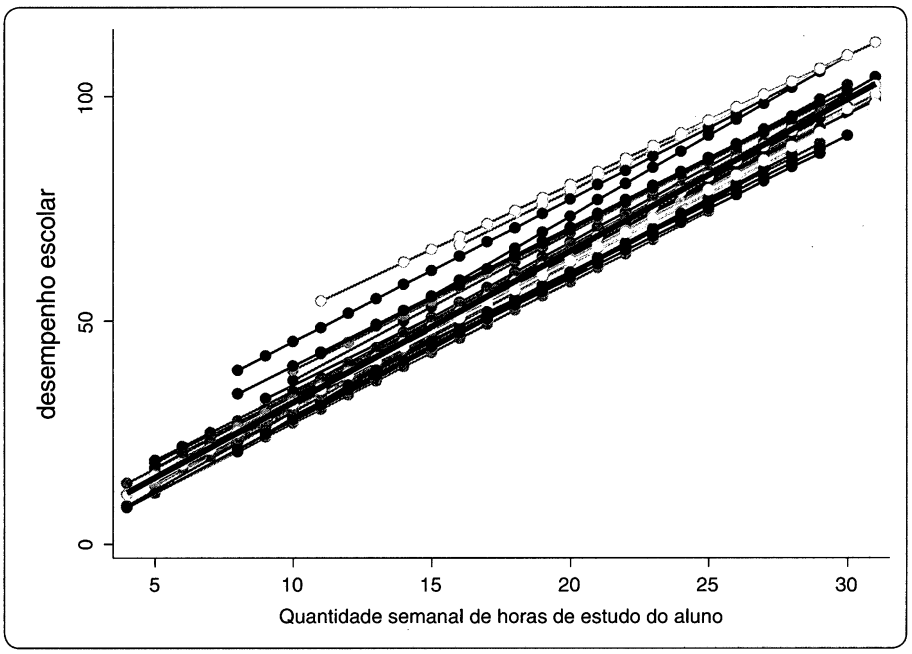


Figura 16.10 Desempenho escolar em função da variável *horas* (variação entre estudantes de uma mesma escola e entre escolas diferentes).

O gráfico da Figura 16.10 apresenta o ajuste linear por MQO, para cada escola, do comportamento do desempenho escolar de cada estudante em função da quantidade semanal de horas de estudo. Podemos verificar que, embora haja melhoria substancial no desempenho escolar à medida que a quantidade semanal de horas de estudo aumenta (felizmente), essa relação não é a mesma para todas as escolas. Mais do que isso, os interceptos de cada modelo são nitidamente distintos.

Portanto, nosso dever passa a ser o de investigar se ocorrem efeitos aleatórios nos interceptos e nas inclinações gerados pela variável *horas*, em decorrência da existência de diversas escolas. Em caso afirmativo, deveremos, posteriormente, investigar se determinadas características das escolas podem responder por tal fato. Note que este último comando também gera um novo arquivo em Stata (**ols.dta**), em que podem ser analisadas as diferenças entre as escolas.

Caso o pesquisador optasse por não incluir efeitos aleatórios na modelagem, ou seja, caso o teste de razão de verossimilhança elaborado na estimação do modelo nulo não rejeitasse H_0 ($u_{0j} = 0$), bastaria que fosse digitado o seguinte comando, conforme estudamos no Capítulo 12, para que os parâmetros do nosso modelo fossem estimados:

```
reg desempenho horas
```

Apenas para fins didáticos, os parâmetros estimados na digitação deste último comando (**reg**), cujos *outputs* não são apresentados aqui, são iguais aos que seriam obtidos por meio do seguinte comando:

```
xtmixed desempenho horas, reml
```

já que o termo **xtmixed** sem a especificação de efeitos aleatórios faz com que sejam estimados, por máxima verossimilhança restrita (termo **reml**), parâmetros com valores idênticos aos que são estimados por mínimos quadrados ordinários (regressão linear apenas com efeitos fixos).

Com base na lógica proposta, vamos, inicialmente, inserir efeitos aleatórios de intercepto no nosso modelo multinível, que passará a ter a seguinte especificação:

Modelo com Interceptos Aleatórios:

$$desempenho_{ij} = b_{0j} + b_{1j}.horas_{ij} + r_{ij}$$

$$b_{0j} = \gamma_{00} + u_{0j}$$

$$b_{1j} = \gamma_{10}$$

que resulta na seguinte expressão:

$$desempenho_{ij} = \gamma_{00} + \gamma_{10}.horas_{ij} + u_{0j} + r_{ij}$$

O comando para a estimação do modelo com interceptos aleatórios no Stata, para os dados do nosso exemplo, é:

```
xtmixed desempenho horas || escola: , var nolog reml
```

que gera os *outputs* da Figura 16.11.

```
. xtmixed desempenho horas || escola: , var nolog reml

Mixed-effects REML regression                   Number of obs      =      2000
Group variable: escola                          Number of groups   =        46

                                                Obs per group: min =         6
                                                               avg =      43.5
                                                               max =        67

                                                Wald chi2(1)       =  19709.41
Log restricted-likelihood = -6372.1643          Prob > chi2        =    0.0000

------------------------------------------------------------------------------
  desempenho |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       horas |   3.251924   .0231635   140.39   0.000     3.206525    3.297324
       _cons |   .5344677   .7875305     0.68   0.497    -1.009064    2.077999
------------------------------------------------------------------------------

------------------------------------------------------------------------------
  Random-effects Parameters  |   Estimate   Std. Err.      [95% Conf. Interval]
-----------------------------+------------------------------------------------
escola: Identity             |
                  var(_cons) |   19.12534   4.199479       12.4367    29.41123
-----------------------------+------------------------------------------------
               var(Residual) |   31.76378   1.016389      29.83288    33.81966
------------------------------------------------------------------------------
LR test vs. linear regression: chibar2(01) =   816.88 Prob >= chibar2 = 0.0000
```

Figura 16.11 *Outputs* do modelo com interceptos aleatórios.

Da mesma forma, a parte superior dos *outputs* mostra os efeitos fixos do nosso modelo, que contempla 46 interceptos separados (um para cada escola), embora não diretamente apresentados. Já a parte inferior corresponde à estimação das variâncias dos termos de erro τ_{00} = 19,125 e σ^2 = 31,764. A correlação intraclasse deste modelo é calculada da seguinte forma:

$$rho = \frac{\tau_{00}}{\tau_{00} + \sigma^2} = \frac{19{,}125}{19{,}125 + 31{,}764} = 0{,}376$$

que mostra um incremento da proporção do componente de variância correspondente ao intercepto em relação ao modelo nulo, demonstrando a importância da inclusão da variável *horas* para o estudo do comportamento do desempenho escolar na comparação entre escolas. Assim como já verificado no modelo nulo, a estimação do componente de variância τ_{00} é quase cinco vezes superior ao seu erro-padrão (z = 19,125/4,199 = 4,555 > 1,96), indicando haver variação significante no desempenho escolar médio entre escolas em decorrência da existência de interceptos aleatórios (os interceptos variam de maneira estatisticamente significante de escola para escola).

Por meio da análise do resultado do teste de razão de verossimilhança (`LR test`, ou *likelihood ratio test*), podemos aqui também rejeitar a hipótese nula de que os interceptos aleatórios sejam iguais a zero (H_0: u_{0j} = 0), já que *Sig.* χ^2 = 0,000, comprovando que a estimação de um modelo tradicional de regressão linear apenas com efeitos fixos seja descartada.

O nosso modelo, portanto, passa a ter, no presente momento, a seguinte especificação:

$$desempenho_{ij} = 0{,}534 + 3{,}252.horas_{ij} + u_{0j} + r_{ij}$$

em que o efeito fixo do intercepto corresponde agora à média esperada dos desempenhos escolares, entre escolas, dos alunos que, por alguma razão, não estudam ($horas_{ij}$ = 0). Por outro lado, uma hora a mais de estudo semanal,

em média, faz com que a média esperada dos desempenhos escolares, entre escolas, seja incrementada em 3,252 pontos, sendo este parâmetro estatisticamente significante.

Apenas para fins didáticos, como esta última estimação representa um modelo em que o componente aleatório contém apenas interceptos, o método de máxima verossimilhança (não restrita) geraria estimações dos parâmetros idênticas às que seriam obtidas por uma estimação tradicional considerando dados em painel (conforme estudamos no Capítulo 15). Além disso, um pesquisador ainda mais curioso poderia verificar que a elaboração de um **modelo linear generalizado multinível** (ou, em inglês, ***generalized linear latent and mixed model - GLLAMM***) também geraria as mesmas estimações dos parâmetros. Em outras palavras, os três comandos a seguir geram estimativas idênticas dos parâmetros e das variâncias dos termos de erro:

Modelo Multinível com Estimação por Máxima Verossimilhança:

```
xtmixed desempenho horas || escola: , var nolog mle
```

em que o termo `mle` significa ***maximum likelihood estimation***.

Modelo para Dados em Painel com Estimação por Máxima Verossimilhança:

```
xtset escola estudante
xtreg desempenho horas, mle
```

Modelo Linear Generalizado Multinível:

```
gllamm desempenho horas, i(escola) adapt
```

em que a opção `adapt` faz com que seja utilizado o processo de **quadratura adaptativa** em vez do processo padrão de **quadratura ordinária de Gauss-Hermite**.

É importante mencionar que os modelos lineares generalizados multinível (*GLLAMM*) são análogos aos modelos lineares generalizados (*GLM*) estudados nos Capítulos 12, 13 e 14, ou seja, também são bastante úteis para a elaboração de modelagens em que a variável dependente apresenta-se de maneira categórica ou com dados de contagem, e existe uma estrutura aninhada de dados. No apêndice deste capítulo, apresentaremos exemplos de modelos hierárquicos não lineares dos tipos logístico, Poisson e binomial negativo. Para um aprofundamento do tema, recomendamos também o estudo de Rabe-Hesketh, Skrondal e Pickles (2002) e de Rabe-Hesketh e Skrondal (2012a, 2012b).

Voltando ao nosso modelo com interceptos aleatórios (*outputs* da Figura 16.11), podemos arquivar (comando `estimates store`) as estimações obtidas para futura comparação com as que serão geradas na estimação de um modelo com interceptos e inclinações aleatórias. Além disso, podemos também obter, por meio do comando `predict, reffects`, os valores esperados dos efeitos aleatórios u_{0j}, conhecidos por ***BLUPS*** (***best linear unbiased predictions***), já que o comando `xtmixed` não os apresenta diretamente. Para tanto, podemos digitar a seguinte sequência de comandos:

```
quietly xtmixed desempenho horas || escola: , var nolog reml
estimates store interceptoaleat
predict u0, reffects
desc u0
by estudante, sort: generate tolist = (_n==1)
list estudante u0 if estudante <= 10 | estudante > 1990 & tolist
```

A Figura 16.12 apresenta os valores dos termos de intercepto aleatório u_{0j} para os primeiros e últimos 10 estudantes da base de dados. Podemos verificar que estes termos de erro são invariantes para estudantes da mesma escola, porém variam entre escolas, o que caracteriza a existência de um intercepto para cada escola.

A fim de propiciar melhor visualização dos interceptos aleatórios por escola, podemos gerar um gráfico (Figura 16.13) digitando o seguinte comando:

```
graph hbar (mean) u0, over(escola) ytitle("Interceptos Aleatórios por
Escola")
```

```
. quietly xtmixed desempenho horas || escola: , var nolog reml
. estimates store interceptoaleat
. predict u0, reffects
. desc u0

              storage  display     value
variable name   type   format      label      variable label
------------------------------------------------------------------------------
u0              float  %9.0g                 BLUP r.e. for escola: _cons

. by estudante, sort: generate tolist = (_n==1)
. list estudante u0 if estudante <= 10 | estudante > 1990 & tolist

     +----------------------+              +----------------------+
     | estuda~e          u0 |              | estuda~e          u0 |
     |----------------------|              |----------------------|
  1. |        1     -2.5026 |       1991.  |     1991   -2.238187 |
  2. |        2     -2.5026 |       1992.  |     1992   -2.238187 |
  3. |        3     -2.5026 |       1993.  |     1993   -2.238187 |
  4. |        4     -2.5026 |       1994.  |     1994   -2.238187 |
  5. |        5     -2.5026 |       1995.  |     1995   -3.096321 |
     |----------------------|              |----------------------|
  6. |        6     -2.5026 |       1996.  |     1996   -3.096321 |
  7. |        7     -2.5026 |       1997.  |     1997   -3.096321 |
  8. |        8     -2.5026 |       1998.  |     1998   -3.096321 |
  9. |        9     -2.5026 |       1999.  |     1999   -3.096321 |
 10. |       10     -2.5026 |       2000.  |     2000   -3.096321 |
     |----------------------|              +----------------------+
```

Figura 16.12 Termos de intercepto aleatório u_{0j}.

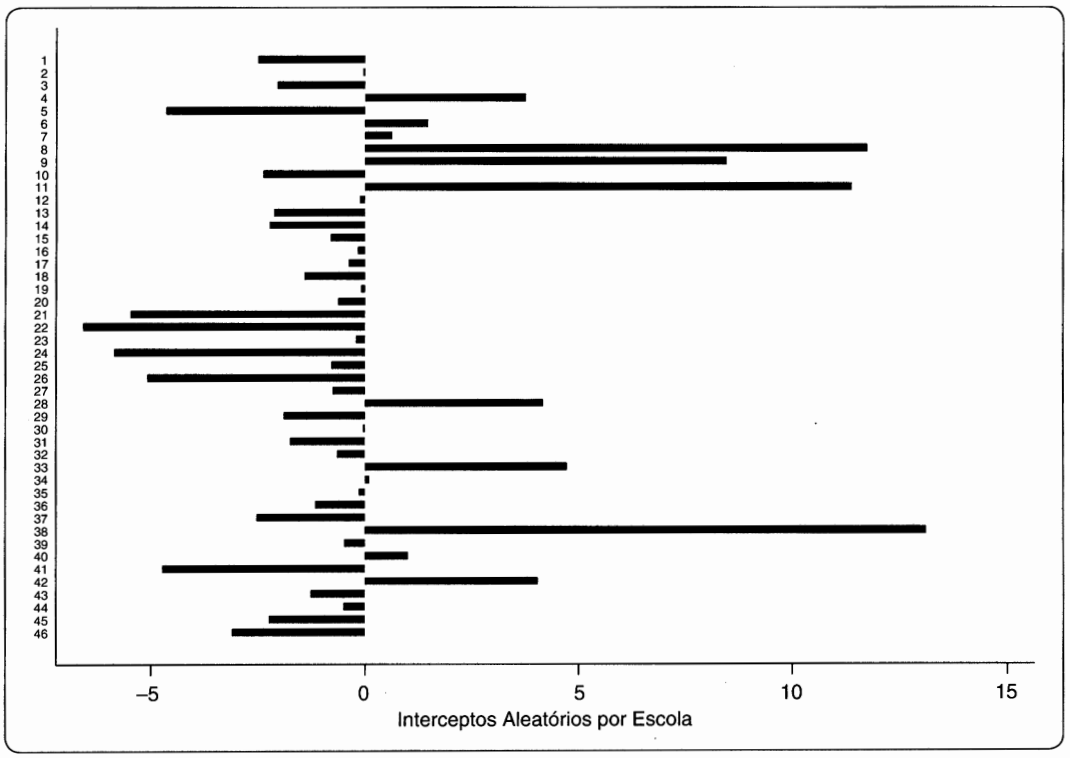


Figura 16.13 Interceptos aleatórios por escola.

Como ainda realizaremos algumas estimações adicionais, a fim de chegarmos a um modelo mais completo e com a presença de variáveis explicativas de nível 2, não vamos, neste momento, apresentar os comandos para gerar os valores previstos do desempenho escolar por estudante. Esse procedimento será realizado mais adiante.

Elaborada a verificação de que o desempenho escolar sofre influência da quantidade de horas de estudo por semana, e de que há diferenças nos interceptos dos modelos entre escolas, vamos, neste momento, estudar se as inclinações também são diferentes entre escolas. Embora os gráficos das Figuras 16.10 e 16.13 permitam que visualizemos, de fato, interceptos discrepantes entre escolas, o mesmo não pode ser dito em relação às inclinações dos 46 ajustes lineares. Entretanto, é nosso dever avaliar tal situação do ponto de vista estatístico. Portanto, vamos

inserir efeitos aleatórios de inclinação no nosso modelo multinível que, com a manutenção dos efeitos aleatórios de intercepto, passará a ter a seguinte expressão:

Modelo com Interceptos e Inclinações Aleatórias:

$$desempenho_{ij} = b_{0j} + b_{1j}.horas_{ij} + r_{ij}$$
$$b_{0j} = \gamma_{00} + u_{0j}$$
$$b_{1j} = \gamma_{10} + u_{1j}$$

que resulta em:

$$desempenho_{ij} = \gamma_{00} + \gamma_{10}.horas_{ij} + u_{0j} + u_{1j}.horas_{ij} + r_{ij}$$

O comando para a estimação do modelo com interceptos e inclinações aleatórias no Stata, para os dados do nosso exemplo, é:

```
xtmixed desempenho horas || escola: horas, var nolog reml
```

Note que a variável *horas* inserida após o termo **escola:** (componente aleatório do comando **xtmixed**) é decorrente do termo $u_{1j}.horas_{ij}$ presente na especificação do modelo multinível. Os resultados obtidos nesta estimação estão na Figura 16.14.

```
. xtmixed desempenho horas || escola: horas, var nolog reml

Mixed-effects REML regression                   Number of obs      =      2000
Group variable: escola                          Number of groups   =        46

                                                Obs per group: min =         6
                                                               avg =      43.5
                                                               max =        67

                                                Wald chi2(1)       =  19709.41
Log restricted-likelihood = -6372.1643          Prob > chi2        =    0.0000

------------------------------------------------------------------------------
  desempenho |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       horas |   3.251924   .0231635   140.39   0.000     3.206525    3.297324
       _cons |    .534468   .7875314     0.68   0.497    -1.009065    2.078001
------------------------------------------------------------------------------

------------------------------------------------------------------------------
  Random-effects Parameters  |   Estimate   Std. Err.      [95% Conf. Interval]
-----------------------------+------------------------------------------------
escola: Independent          |
                  var(horas) |   8.37e-14   8.99e-11            0           .
                  var(_cons) |    19.1254   4.199523       12.4367    29.41142
-----------------------------+------------------------------------------------
               var(Residual) |   31.76378   1.016389      29.83287    33.81966
------------------------------------------------------------------------------
LR test vs. linear regression:       chi2(2) =   816.88   Prob > chi2 = 0.0000

Note: LR test is conservative and provided only for reference.
```

Figura 16.14 *Outputs* do modelo com interceptos e inclinações aleatórias.

Podemos verificar que as estimações dos parâmetros e das variâncias no modelo com interceptos e inclinações aleatórias são praticamente idênticas aos obtidos na estimação dos parâmetros do modelo apenas com interceptos aleatórios (Figura 16.11). Isso decorre do fato de que a estimação da variância τ_{11} dos termos de inclinação aleatória u_{1j} ser estatisticamente igual a zero (valor muito baixo e erro-padrão consideravelmente superior, com valores iguais a zero para os intervalos de confiança).

Embora esse fato seja nítido neste caso, o pesquisador tem a opção de elaborar o teste de razão de verossimilhança para comparar as estimações obtidas pelo modelo com interceptos aleatórios e pelo modelo com interceptos e inclinações aleatórias. Para tanto, deve ser digitado o seguinte comando:

```
estimates store inclinaçãoaleat
```

e, na sequência, o comando que irá elaborar o teste:

```
lrtest inclinaçãoaleat interceptoaleat
```

visto que o termo **interceptoaleat** refere-se à estimação já realizada anteriormente. O resultado do teste é apresentado na Figura 16.15.

```
. lrtest inclinaçãoaleat interceptoaleat

Likelihood-ratio test                                 LR chi2(1)  =     -0.00
(Assumption: interceptoal~t nested in inclinaçãoal~t) Prob > chi2 =    1.0000

Note: The reported degrees of freedom assumes the null hypothesis is not on the
boundary of the parameter space.  If this is not true, then the reported test
is conservative.
Note: LR tests based on REML are valid only when the fixed-effects specification
is identical for both models.
```

Figura 16.15 Teste de razão de verossimilhança para comparar as estimações dos modelos com interceptos aleatórios e com interceptos e inclinações aleatórias.

Sendo o nível de significância do teste igual a 1,000 (muito maior do que 0,05) em decorrência do fato de que os logaritmos das duas funções de verossimilhança restrita são idênticos (LL_r = -6.372,164), fazendo com que **LR chi2** para um grau de liberdade seja igual a 0, é favorecido o modelo apenas com efeitos aleatórios no intercepto, comprovando que os termos de erro aleatório u_{1j} são estatisticamente iguais a zero. É importante mencionar, conforme também explicita a nota na parte inferior da Figura 16.15, que **este teste de razão de verossimilhança somente é válido quando for feita a comparação das estimações obtidas por máxima verossimilhança restrita (*REML*) de dois modelos com especificação idêntica do componente de efeitos fixos**. Como, no nosso caso, os dois modelos, que foram estimados por *REML*, apresentam a mesma especificação $\gamma_{00} + \gamma_{10}.horas_{ij}$ no componente de efeitos fixos, o teste é considerado válido[3].

Apenas para fins didáticos, outro modo de analisar a significância estatística dos termos de erro do modelo multinível é inserir o termo **estmetric** ao final do comando **xtmixed**, conforme segue:

```
xtmixed desempenho horas || escola: horas, estmetric nolog reml
```

Os *outputs* gerados são apresentados na Figura 16.16.

```
. xtmixed desempenho horas || escola: horas, estmetric nolog reml

Mixed-effects REML regression                   Number of obs      =      2000
Group variable: escola                          Number of groups   =        46

                                                Obs per group: min =         6
                                                               avg =      43.5
                                                               max =        67

                                                Wald chi2(1)       =  19709.41
Log restricted-likelihood = -6372.1643          Prob > chi2        =    0.0000

------------------------------------------------------------------------------
  desempenho |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
desempenho   |
       horas |   3.251924   .0231635   140.39   0.000     3.206525    3.297324
       _cons |    .534468   .7875314     0.68   0.497    -1.009065    2.078001
-------------+----------------------------------------------------------------
lns1_1_1     |
       _cons |  -15.05597   537.5352    -0.03   0.978    -1068.606    1038.494
-------------+----------------------------------------------------------------
lns1_1_2     |
       _cons |   1.475509   .1097892    13.44   0.000     1.260326    1.690691
-------------+----------------------------------------------------------------
lnsig_e      |
       _cons |   1.729163   .0159992   108.08   0.000     1.697805    1.760521
------------------------------------------------------------------------------
```

Figura 16.16 Estimação dos parâmetros do modelo com interceptos e inclinações aleatórias, com uso do termo **estmetric**.

[3] Se um pesquisador mais curioso desejar elaborar um teste de razão de verossimilhança para comparar as estimações dos modelos nulo e com interceptos aleatórios, cujas especificações dos componentes fixos são obviamente diferentes, deverá fazê-lo estimando estes dois modelos por máxima verossimilhança (*MLE*), em vez de por máxima verossimilhança restrita (*REML*). Assim, deverá digitar a seguinte sequência de comandos:

```
quietly xtmixed desempenho || escola: , var nolog mle
estimates store nulomle
quietly xtmixed desempenho horas || escola: , var nolog mle
estimates store interceptoaleatmle
lrtest nulomle interceptoaleatmle
```

cujo resultado obtido favorece o modelo com efeitos aleatórios no intercepto em relação ao modelo nulo.

As estimações dos parâmetros de efeitos fixos são idênticas às obtidas anteriormente, porém o termo **estmetric** faz com que sejam apresentadas as estimações do logaritmo natural dos desvios-padrão dos termos de erro, em vez das variâncias desses termos, com as respectivas estatísticas z e seus níveis de significância, o que facilita a interpretação da significância estatística de cada termo aleatório.

Para o termo r_{ij}, por exemplo, em vez de ser apresentada a estimação da sua variância $\sigma^2 = 31{,}764$ (Figura 16.14), é apresentada a estimação do logaritmo natural do desvio-padrão de r_{ij}, de modo que:

$$\ln\left(\sqrt{31{,}764}\right) = 1{,}729$$

Neste sentido, podemos comprovar, portanto, que os termos de inclinação aleatória u_{1j} são estatisticamente iguais a zero ao nível de confiança de, por exemplo, 95%, já que *Sig.* $z = 0{,}978 > 0{,}05$.

Outra discussão pertinente neste momento diz respeito à estrutura da matriz de variância-covariância dos efeitos aleatórios u_{0j} e u_{1j}. Como não especificamos nenhuma estrutura de covariância para estes termos de erro, o Stata pressupõe, por meio do comando **xtmixed**, que essa estrutura seja independente, ou seja, que cov $(u_{0j}, u_{1j}) = \sigma_{01} = 0$. Em outras palavras, com base na expressão (16.18) e nos *outputs* da Figura 16.14, temos que:

$$\mathbf{G} = var[\mathbf{u}] = var\begin{bmatrix} u_{0j} \\ u_{1j} \end{bmatrix} = \begin{bmatrix} \tau_{00} & 0 \\ 0 & \tau_{11} \end{bmatrix} = \begin{bmatrix} 19{,}125 & 0 \\ 0 & 8{,}37x10^{-14} \end{bmatrix}$$

Entretanto, podemos generalizar a estrutura da matriz **G**, permitindo que u_{0j} e u_{1j} sejam correlacionados, ou seja, que cov$(u_{0j}, u_{1j}) = \sigma_{01} \neq 0$. Para tanto, basta que adicionemos o termo **covariance(unstructured)** ao comando **xtmixed**, de modo que:

```
xtmixed desempenho horas || escola: horas, covariance(unstructured)
var nolog reml
```

Os novos *outputs* gerados são apresentados na Figura 16.17.

```
. xtmixed desempenho horas || escola: horas, covariance(unstructured) var nolog reml

Mixed-effects REML regression                   Number of obs      =      2000
Group variable: escola                          Number of groups   =        46

                                                Obs per group: min =         6
                                                               avg =      43.5
                                                               max =        67

                                                Wald chi2(1)       =  19620.62
Log restricted-likelihood = -6372.1111          Prob > chi2        =    0.0000

------------------------------------------------------------------------------
  desempenho |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       horas |   3.251008   .0232093   140.07   0.000     3.205519    3.296498
       _cons |   .5615094   .8100559     0.69   0.488    -1.026171     2.14919
------------------------------------------------------------------------------

------------------------------------------------------------------------------
  Random-effects Parameters  |   Estimate   Std. Err.      [95% Conf. Interval]
-----------------------------+------------------------------------------------
escola: Unstructured         |
                  var(horas) |   .0000759    .000075      .0000109    .0005268
                  var(_cons) |   20.74997   4.425246      13.66111    31.51731
             cov(horas,_cons) |  -.0396861    .019402     -.0777133    -.001659
-----------------------------+------------------------------------------------
               var(Residual) |   31.75566   1.02383       29.81108    33.82709
------------------------------------------------------------------------------
LR test vs. linear regression:       chi2(3) =   816.99   Prob > chi2 = 0.0000

Note: LR test is conservative and provided only for reference.
```

Figura 16.17 Estimação dos parâmetros do modelo com interceptos e inclinações aleatórias, com termos aleatórios u_{0j} e u_{1j} correlacionados.

As novas estimações das variâncias dos termos de erro geram a seguinte matriz de variância-covariância:

$$var[\mathbf{u}] = var\begin{bmatrix} u_{0j} \\ u_{1j} \end{bmatrix} = \begin{bmatrix} \tau_{00} & \sigma_{01} \\ \sigma_{01} & \tau_{11} \end{bmatrix} = \begin{bmatrix} 20{,}750 & -0{,}040 \\ -0{,}040 & 7{,}59x10^{-5} \end{bmatrix}$$

que também pode ser obtida por meio do seguinte comando:

```
estat recovariance
```

cujos *outputs* encontram-se na Figura 16.18.

```
. estat recovariance

Random-effects covariance matrix for level escola

             |     horas       _cons
-------------+----------------------
       horas |  .0000759
       _cons | -.0396861   20.74997
```

Figura 16.18 Matriz de variância-covariância com termos aleatórios u_{0j} e u_{1j} correlacionados.

Embora a estimação da covariância entre u_{0j} e u_{1j} $\text{cov}(u_{0j}, u_{1j}) = \sigma_{01} = -0{,}040 \neq 0$, um pesquisador mais curioso verificará, por meio da inclusão do termo **estmetric** ao final do último comando **xtmixed** digitado (sem o termo **var**), que esta covariância não é estatisticamente significante (na realidade, o *output*, não apresentado aqui, mostrará a não significância do arco tangente hiperbólico da correlação entre estes dois termos de erro).

Outro modo para verificar a não significância da correlação entre os termos de erro é por meio de um novo teste de razão de verossimilhança, que compara as estimações do modelo com interceptos e inclinações aleatórias com termos de erro u_{0j} e u_{1j} independentes (Figura 16.14) com o mesmo modelo, porém com termos de erro correlacionados (Figura 16.17), ou seja, com matriz de variância-covariância *unstructured*. Para tanto, devemos digitar a seguinte sequência de comandos:

```
estimates store inclinaçãoaleatunstructured

lrtest inclinaçãoaleatunstructured inclinaçãoaleat
```

O resultado deste teste está na Figura 16.19.

```
. lrtest inclinaçãoaleatunstructured inclinaçãoaleat

Likelihood-ratio test                                 LR chi2(1)  =      0.11
(Assumption: inclinaçãoal~t nested in inclinaçãoal~d) Prob > chi2 =    0.7442

Note: LR tests based on REML are valid only when the fixed-effects specification
is identical for both models.
```

Figura 16.19 Teste de razão de verossimilhança para comparar as estimações dos modelos com interceptos e inclinações aleatórias com termos de erro u_{0j} e u_{1j} independentes e correlacionados.

A estatística χ^2 deste teste, com 1 grau de liberdade, também pode ser obtida por meio da seguinte expressão:

$$\chi^2_1 = [-2.LL_{\text{r-ind}} - (-2.LL_{\text{r-unstruc}})] = \{-2.(-6.372{,}164) - [-2.(-6.372{,}111)]\} = 0{,}11$$

Ou seja, temos que *Sig.* $\chi^2_1 = 0{,}744 > 0{,}05$. Portanto, podemos afirmar que a estrutura da matriz de variância-covariância entre u_{0j} e u_{1j} pode ser considerada independente neste exemplo.

Porém, mais do que isso, verificamos que a variância estimada de u_{1j} é estatisticamente igual a zero, fazendo com que o modelo com interceptos aleatórios seja mais adequado do que o modelo com interceptos e inclinações aleatórias para os nossos dados.

Vamos neste momento, portanto, inserir as variáveis *texp* e *priv* (variáveis explicativas do nível 2 - escola) no nosso modelo com interceptos aleatórios, de modo que a nova especificação do modelo hierárquico fique conforme segue:

Modelo Completo com Interceptos Aleatórios:

$$desempenho_{ij} = b_{0j} + b_{1j}.horas_{ij} + r_{ij}$$

$$b_{0j} = \gamma_{00} + \gamma_{01}.texp_j + \gamma_{02}.priv_j + u_{0j}$$

$$b_{1j} = \gamma_{10} + \gamma_{11}.texp_j + \gamma_{12}.priv_j$$

que resulta na seguinte expressão:

$$desempenho_{ij} = \gamma_{00} + \gamma_{10}.horas_{ij} + \gamma_{01}.texp_j + \gamma_{02}.priv_j + \gamma_{11}.texp_j.horas_{ij} + \gamma_{12}.priv_j.horas_{ij} + u_{0j} + r_{ij}$$

Desta forma, precisamos, inicialmente, gerar duas novas variáveis, que correspondem à multiplicação de *texp* por *horas* e de *priv* por *horas*. Os comandos a seguir geram estas duas variáveis (*texphoras* e *privhoras*):

```
gen texphoras = texp*horas
gen privhoras = priv*horas
```

Na sequência, podemos estimar o nosso modelo completo com interceptos aleatórios, digitando o seguinte comando:

```
xtmixed desempenho horas texp priv texphoras privhoras || escola: , var nolog reml
```

Os *outputs* são apresentados na Figura 16.20.

```
. xtmixed desempenho horas texp priv texphoras privhoras || escola: , var nolog reml

Mixed-effects REML regression                   Number of obs      =      2000
Group variable: escola                          Number of groups   =        46

                                                Obs per group: min =         6
                                                               avg =      43.5
                                                               max =        67

                                                Wald chi2(5)       =  19953.89
Log restricted-likelihood = -6363.6519          Prob > chi2        =    0.0000

------------------------------------------------------------------------------
  desempenho |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       horas |   3.284991   .0332137    98.90   0.000     3.219893    3.350088
        texp |   .9073246   .2316582     3.92   0.000      .4532829    1.361366
        priv |  -6.067564   2.921377    -2.08   0.038    -11.79336   -.3417699
   texphoras |  -.0019725   .0078371    -0.25   0.801    -.0173328    .0133879
   privhoras |  -.0579369   .1002329    -0.58   0.563    -.2543899    .1385161
       _cons |  -2.792594   .9512356    -2.94   0.003    -4.656982    -.928207
------------------------------------------------------------------------------

------------------------------------------------------------------------------
  Random-effects Parameters  |   Estimate   Std. Err.      [95% Conf. Interval]
-----------------------------+------------------------------------------------
escola: Identity             |
                  var(_cons) |    11.0621    2.56052      7.027675    17.41258
-----------------------------+------------------------------------------------
               var(Residual) |   31.73555   1.015985      29.80544    33.79064
------------------------------------------------------------------------------
LR test vs. linear regression: chibar2(01) =   466.96 Prob >= chibar2 = 0.0000
```

Figura 16.20 *Outputs* do modelo completo com interceptos aleatórios.

Ao analisarmos os parâmetros estimados do componente de efeitos fixos, podemos verificar que aqueles correspondentes às variáveis *texphoras* e *privhoras* não são estatisticamente diferentes de zero, ao nível de significância de 5%. Como não há procedimento *Stepwise* correspondente ao comando **xtmixed** no Stata, vamos manualmente excluir a variável *texphoras* (ou seja, a variável *texp* da expressão da inclinação b_{1j}), por ser aquela cujo parâmetro estimado apresentou maior *Sig. z*. O novo modelo, portanto, apresenta a seguinte expressão:

$$desempenho_{ij} = b_{0j} + b_{1j}.horas_{ij} + r_{ij}$$
$$b_{0j} = \gamma_{00} + \gamma_{01}.texp_j + \gamma_{02}.priv_j + u_{0j}$$
$$b_{1j} = \gamma_{10} + \gamma_{11}.priv_j$$

que resulta em:

$$desempenho_{ij} = \gamma_{00} + \gamma_{10}.horas_{ij} + \gamma_{01}.texp_j + \gamma_{02}.priv_j + \gamma_{11}.priv_j.horas_{ij} + u_{0j} + r_{ij}$$

cuja estimação pode ser obtida por meio da digitação do seguinte comando:

```
xtmixed desempenho horas texp priv privhoras || escola: , var nolog reml
```

Os novos *outputs* são apresentados na Figura 16.21.

```
. xtmixed desempenho horas texp priv privhoras || escola: , var nolog reml

Mixed-effects REML regression                   Number of obs      =      2000
Group variable: escola                          Number of groups   =        46

                                                Obs per group: min =         6
                                                               avg =      43.5
                                                               max =        67

                                                Wald chi2(4)       =  19963.20
Log restricted-likelihood = -6359.7535          Prob > chi2        =    0.0000

------------------------------------------------------------------------------
  desempenho |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       horas |   3.281046   .0292757   112.07   0.000     3.223666    3.338425
        texp |   .8662029   .1641964     5.28   0.000     .5443839    1.188022
        priv |  -5.610535   2.288086    -2.45   0.014    -10.0951    -1.12597
   privhoras |  -.0801207   .0477218    -1.68   0.093    -.1736538    .0134124
       _cons |   -2.71035   .8931607    -3.03   0.002    -4.460913   -.9597874
------------------------------------------------------------------------------

------------------------------------------------------------------------------
  Random-effects Parameters  |   Estimate   Std. Err.      [95% Conf. Interval]
-----------------------------+------------------------------------------------
escola: Identity             |
                  var(_cons) |   11.05778   2.559528      7.024925    17.40582
-----------------------------+------------------------------------------------
               var(Residual) |    31.7206   1.015254      29.79187     33.7742
------------------------------------------------------------------------------
LR test vs. linear regression: chibar2(01) =   467.10 Prob >= chibar2 = 0.0000
```

Figura 16.21 *Outputs* do modelo final completo com interceptos aleatórios sem a variável *texphoras*.

Note que, embora o parâmetro estimado γ_{11} referente à variável *privhoras* não seja estatisticamente significante ao nível de significância de 5%, o é ao nível de significância de 10%. Apenas para fins didáticos, consideraremos este maior nível de significância neste momento, a fim de darmos sequência à análise com a presença de ao menos uma variável de nível 2 (*priv*) na expressão da inclinação b_{1j}, ainda que sem termos aleatórios nesta inclinação. Portanto, a expressão do nosso modelo final estimado com interceptos aleatórios e variáveis explicativas dos níveis 1 e 2 é:

$$desempenho_{ij} = -2,710 + 3,281.horas_{ij} + 0,866.texp_j - 5,610.priv_j - 0,080.priv_j.horas_{ij} + u_{0j} + r_{ij}$$

Um pesquisador mais investigativo poderia questionar o fato de o parâmetro estimado da variável *priv* apresentar sinal negativo. Lembramos que esse fato somente ocorre na presença das demais variáveis explicativas, pois a correlação entre *desempenho* e *priv* é positiva e estatisticamente significante, ao nível de significância de 5%, o que comprova que estudantes provenientes de escolas de natureza privada acabam por apresentar, em média, desempenhos escolares superiores aos dos estudantes provenientes de escolas públicas.

Na sequência, podemos obter os valores esperados *BLUPS* (*best linear unbiased predictions*) dos efeitos aleatórios u_{0j} do nosso modelo final, digitando:

```
predict u0final, reffects
```

que gera no banco de dados uma nova variável, denominada *u0final*. Além disso, também podemos obter os valores esperados do desempenho escolar de cada estudante, por meio da digitação do seguinte comando:

```
predict yhat, fitted
```

que define a variável *yhat*, que também pode ser obtida pelo comando:

```
gen yhat = -2.71035 + 3.281046*horas + .8662029*texp - 5.610535*priv -
.0801207*privhoras + u0final
```

O comando a seguir faz com que seja gerado um gráfico (Figura 16.22) com os valores previstos do desempenho escolar de cada estudante em função da quantidade semanal de horas de estudo para as 46 escolas em análise e, por meio do qual, podemos visualizar que os interceptos são distintos (efeitos aleatórios), porém sem que haja discrepância nas inclinações.

```
graph twoway connected yhat horas, connect(L)
```

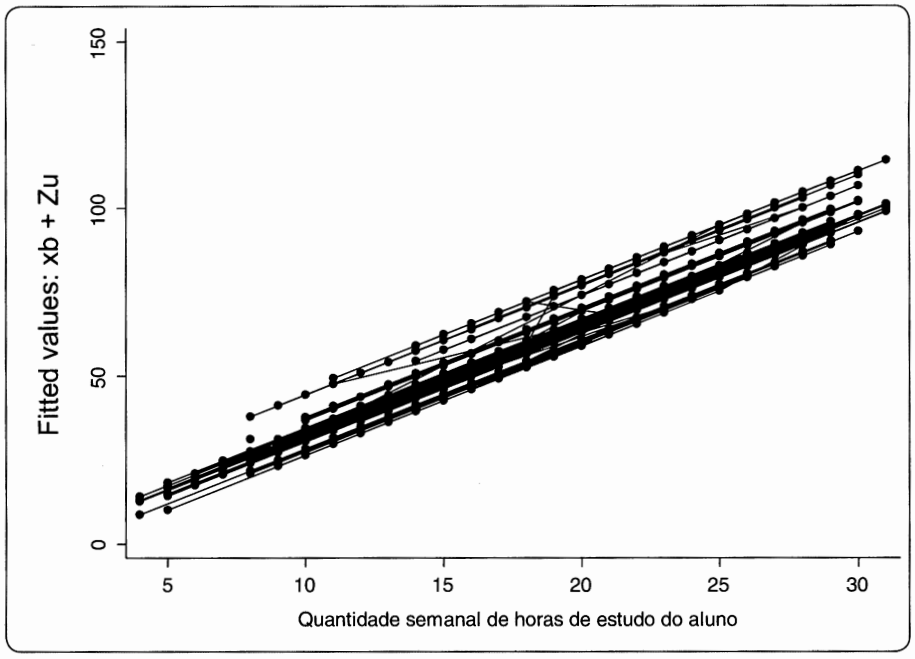


Figura 16.22 Valores previstos do desempenho escolar em função da variável *horas* para o modelo final completo com interceptos aleatórios.

Por fim, a Figura 16.23 apresenta os valores dos interceptos e das inclinações dos ajustes lineares dos valores previstos do desempenho escolar médio para cada uma das 46 escolas, em que é possível comprovar a existência de efeitos aleatórios nos interceptos e apenas de efeitos fixos nas inclinações. Essa figura pode ser obtida com a digitação da seguinte sequência de comandos:

```
generate interceptfinal = _b[_cons] + u0final
generate slopefinal = _b[horas] + _b[texp] + _b[priv] + _b[privhoras]
by escola, sort: generate grupo = (_n==1)
list escola interceptfinal slopefinal if grupo == 1
```

Portanto, podemos concluir que existem diferenças no comportamento do desempenho escolar entre estudantes provenientes de mesmas escolas e de escolas distintas, e essas diferenças ocorrem, respectivamente, em função da quantidade semanal de horas de estudo de cada estudante, da natureza (pública ou privada) e do tempo médio de experiência docente dos professores de cada escola.

```
. generate interceptfinal = _b[_cons] + u0final
. generate slopefinal = _b[horas] + _b[texp] + _b[priv] + _b[privhoras]
. by escola, sort: generate grupo = (_n==1)
. list escola interceptfinal slopefinal if grupo == 1
```

	escola	intercept	slope
1.	1	-4.16957	-1.543407
48.	2	-1.894821	-1.543407
73.	3	-3.666173	-1.543407
121.	4	2.755683	-1.543407
141.	5	-5.345044	-1.543407
189.	6	-.3607166	-1.543407
219.	7	-1.135043	-1.543407
247.	8	1.99781	-1.543407
282.	9	-1.299724	-1.543407
326.	10	-4.221467	-1.543407
359.	11	1.197181	-1.543407
416.	12	-8.295818	-1.543407
478.	13	-3.741182	-1.543407
531.	14	-3.841384	-1.543407
558.	15	-1.455961	-1.543407
611.	16	-2.030933	-1.543407
639.	17	-2.306067	-1.543407
668.	18	-3.19111	-1.543407
707.	19	-1.866918	-1.543407
754.	20	-1.314391	-1.543407
814.	21	-7.131632	-1.543407
875.	22	-8.121008	-1.543407
942.	23	-2.087642	-1.543407
989.	24	-6.462057	-1.543407
1046.	25	-2.490379	-1.543407
1098.	26	-5.595652	-1.543407
1155.	27	-2.556698	-1.543407
1193.	28	-4.038416	-1.543407
1250.	29	-3.504889	-1.543407
1292.	30	-1.804854	-1.543407
1330.	31	-3.479754	-1.543407
1382.	32	-1.441315	-1.543407
1427.	33	3.12553	-1.543407
1474.	34	-1.68581	-1.543407
1499.	35	-1.887107	-1.543407
1554.	36	-2.94762	-1.543407
1596.	37	-4.148458	-1.543407
1639.	38	3.211197	-1.543407
1687.	39	-2.189148	-1.543407
1733.	40	-.7969732	-1.543407
1786.	41	-13.63122	-1.543407
1845.	42	3.058528	-1.543407
1866.	43	-2.950832	-1.543407
1905.	44	-2.277107	-1.543407
1957.	45	-4.016261	-1.543407
1995.	46	-4.640889	-1.543407

Figura 16.23 Efeitos aleatórios nos interceptos e efeitos fixos nas inclinações (em destaque, a identificação da primeira observação em cada escola).

Optamos por elaborar a estratégica de análise multinível proposta por Raudenbush e Bryk (2002) e Snijders e Bosker (2011), ou seja, primeiramente **estudamos a decomposição de variância a partir da definição de um modelo nulo** (modelo não condicional) para, na sequência, **serem construídos um modelo com interceptos aleatórios e um modelo com interceptos e inclinações aleatórias**. Por fim, a partir da definição do caráter de aleatoriedade dos termos de erro, **construímos o modelo completo com a inclusão das variáveis de nível 2** na análise. Esse procedimento é conhecido por ***multilevel step-up strategy***.

Em seguida, iremos elaborar uma modelagem hierárquica linear de três níveis, em que será caracterizado o aninhamento dos dados pela presença de medidas repetidas, ou seja, pela existência de evolução temporal no comportamento da variável dependente.

16.4.2. Estimação de um modelo hierárquico linear de três níveis com medidas repetidas no software Stata

Apresentaremos um exemplo que segue a mesma lógica da seção anterior, porém, neste momento, com dados que variam ao longo do tempo, entre indivíduos e entre grupos a que pertencem esses indivíduos, caracterizando uma estrutura aninhada com medidas repetidas.

Imagine que o nosso versado e matraqueado professor tenha agora o interesse em ampliar sua pesquisa, monitorando o desempenho escolar dos estudantes por determinado período, a fim de investigar se existe variabilidade nesse desempenho ao longo do tempo entre estudantes provenientes de uma mesma escola e entre aqueles provenientes de escolas distintas e, em caso afirmativo, se existem características dos estudantes e das escolas que explicam essa variabilidade.

Neste sentido, 15 escolas se dispuseram a fornecer os dados referentes ao desempenho escolar (nota de 0 a 100) de seus alunos nos últimos quatro anos, totalizando 610 estudantes. Além disso, o professor também incluiu na base o sexo de cada um deles, a fim de verificar se existem diferenças decorrentes dessa variável no desempenho escolar. A variável referente ao tempo médio de experiência docente em cada uma das escolas permanece no estudo. Parte do banco de dados elaborado encontra-se na Tabela 16.4, porém a base de dados completa pode ser acessada por meio dos arquivos **DesempenhoTempoAlunoEscola.xls** (Excel) e **DesempenhoTempoAlunoEscola.dta** (Stata).

Tabela 16.4 Exemplo: desempenho escolar ao longo do tempo (nível 1 – medida repetida) e características de estudantes (nível 2) e de escolas (nível 3).

Estudante j (Nível 2)	Escola k (Nível 3)	Desempenho escolar (Y_{tjk})	Ano t (Nível 1)	Sexo (X_{jk})	Tempo médio, em anos, de experiência dos docentes (W_k)
1	1	35,4	1	masculino	2
1	1	44,4	2	masculino	2
1	1	46,4	3	masculino	2
1	1	52,4	4	masculino	2
...					
121	4	66,4	1	feminino	9
121	4	66,4	2	feminino	9
121	4	74,4	3	feminino	9
121	4	79,4	4	feminino	9
...					
610	15	87,6	1	feminino	9
610	15	92,6	2	feminino	9
610	15	94,6	3	feminino	9
610	15	100,0	4	feminino	9

Após abrirmos o arquivo **DesempenhoTempoAlunoEscola.dta**, podemos digitar o comando **desc**, que permite que analisemos as características do banco de dados, como a quantidade de observações, a quantidade de variáveis e a descrição de cada uma delas. A Figura 16.24 apresenta este *output* do Stata.

```
. desc

  obs:         2,440
 vars:             6
 size:        56,120
-------------------------------------------------------------------------------
              storage  display     value
variable name   type   format      label      variable label
-------------------------------------------------------------------------------
estudante       int    %8.0g                  estudante j (nível 2)
escola          byte   %8.0g                  escola k (nível 3)
desempenho      float  %9.0g                  desempenho escolar
ano             float  %9.0g                  período de monitoramento (ano 1 a 4)
sexo            float  %9.0g       sexo       sexo
texp            float  %9.0g                  tempo médio de experiência docente dos
                                               professores da escola (anos)
-------------------------------------------------------------------------------
Sorted by:
```

Figura 16.24 Descrição do banco de dados **DesempenhoTempoAlunoEscola.dta**.

Seguindo a lógica proposta na seção anterior, vamos inicialmente analisar a quantidade de estudantes monitorados pelo professor em cada período de tempo (*ano*), por meio do seguinte comando:

```
tabulate ano, subpop(estudante)
```

Os *outputs* são apresentados na Figura 16.25 e, por meio desses, podemos verificar que estamos diante de um **painel balanceado de dados**, já que cada um dos 610 estudantes é monitorado nos quatro períodos de tempo.

```
. tabulate ano, subpop(estudante)

 período de |
monitoramen |
to (ano 1 a |
         4) |      Freq.     Percent        Cum.
------------+-----------------------------------
          1 |        610       25.00       25.00
          2 |        610       25.00       50.00
          3 |        610       25.00       75.00
          4 |        610       25.00      100.00
------------+-----------------------------------
      Total |      2,440      100.00
```

Figura 16.25 Quantidade de estudantes monitorados em cada período.

O gráfico da Figura 16.26, obtido por meio da digitação do seguinte comando, permite que seja analisada a evolução temporal do desempenho escolar dos 50 primeiros estudantes da amostra:

```
graph twoway connected desempenho ano if estudante <= 50, connect(L)
```

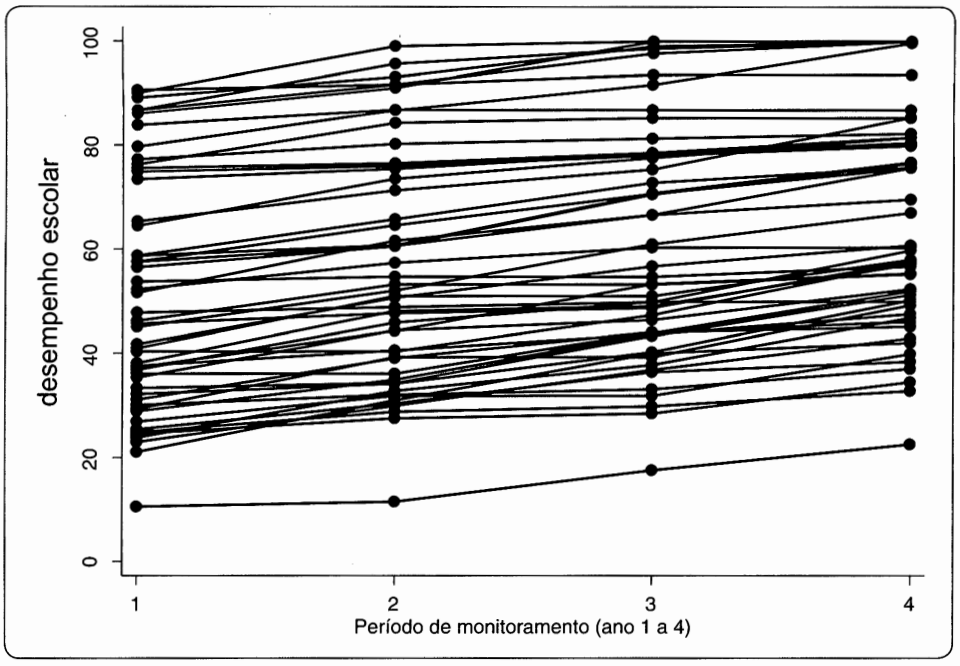


Figura 16.26 Evolução temporal do desempenho escolar dos 50 primeiros estudantes da amostra.

Este gráfico já permite que visualizemos que as evoluções temporais dos desempenhos escolares apresentam interceptos e inclinações distintas entre estudantes, o que justifica a adoção da modelagem multinível e **oferece subsídios à inclusão de efeitos aleatórios de intercepto e de inclinação no nível 2** dos modelos que serão estimados.

Além disso, os desempenhos médios dos estudantes nos quatro períodos podem ser analisados nas Figuras 16.27 e 16.28, obtidas a partir dos comandos a seguir. Por meio delas, é possível verificar que existe um comportamento crescente, aproximadamente linear, do desempenho escolar dos estudantes ao longo do tempo, e **essa é a razão para que também seja inserida a variável *ano*, com especificação linear, no nível 1 da modelagem**, conforme veremos adiante.

```
bysort ano: egen desempenho_médio = mean(desempenho)

tabstat desempenho_médio, by(ano)

graph twoway scatter desempenho ano || connected desempenho_médio ano,
connect(L) || , ytitle(desempenho escolar)
```

```
. bysort ano: egen desempenho_médio = mean(desempenho)
. tabstat desempenho_médio, by(ano)

Summary for variables: desempenho_médio
     by categories of: ano (período de monitoramento (ano 1 a 4))

     ano |      mean
---------+----------
       1 |  61.65492
       2 |  66.36607
       3 |  70.61115
       4 |  74.73328
---------+----------
   Total |  68.34135
--------------------
```

Figura 16.27 Desempenho escolar médio dos estudantes em cada período.

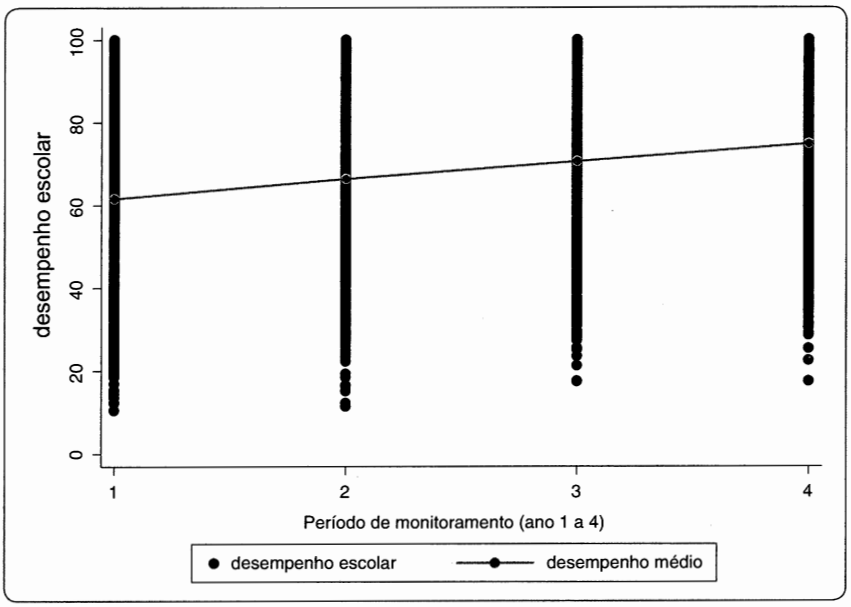


Figura 16.28 Evolução do desempenho escolar médio dos estudantes em cada período.

A fim de justificar mais fortemente as razões para que seja estimado um modelo hierárquico de três níveis, vamos elaborar um gráfico (Figura 16.29) que apresenta as evoluções temporais dos desempenhos escolares médios. Para tanto, podemos digitar a seguinte sequência de comandos:

```
  statsby intercept=_b[_cons] slope=_b[ano], by(escola) saving(ols,
replace): reg desempenho ano
  sort escola
  merge escola using ols
  drop _merge
  gen yhat_ols= intercept + slope*ano
  sort escola ano
  separate desempenho, by(escola)
  separate yhat_ols, by(escola)
  graph twoway connected yhat_ols1-yhat_ols15 ano || lfit desempenho
ano, clwidth(thick) clcolor(black) legend(off) ytitle(desempenho escolar)
```

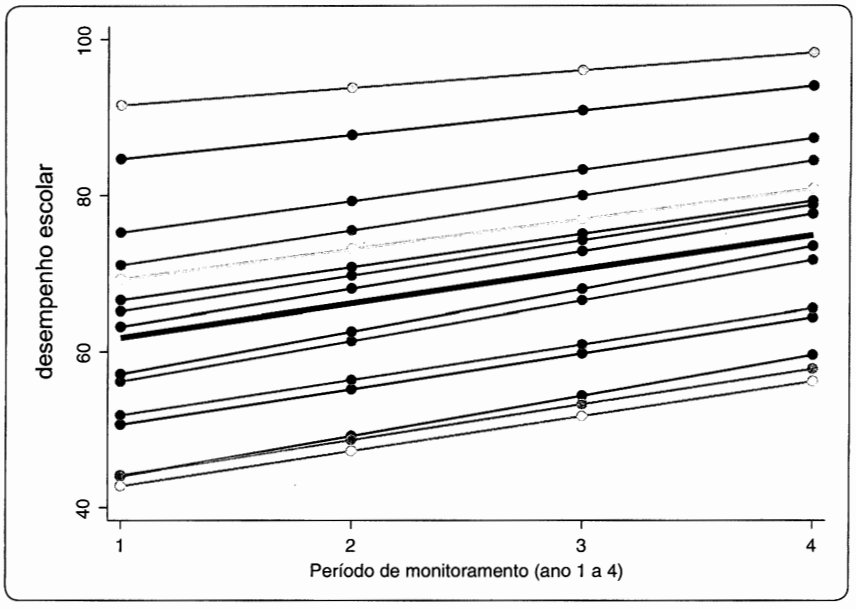


Figura 16.29 Evolução temporal do desempenho escolar médio dos estudantes de cada escola (ajuste linear por MQO).

Este gráfico apresenta o ajuste linear por MQO, para cada escola, do comportamento do desempenho escolar ao longo do tempo e também **oferece subsídios à inclusão de efeitos aleatórios de intercepto e de inclinação no nível 3** dos modelos que serão estimados, já que as evoluções temporais dos desempenhos escolares apresentam interceptos e inclinações distintas também entre as escolas. Note que a última sequência de comandos gera um novo arquivo em Stata (**ols.dta**), em que podem ser analisadas as diferenças no comportamento do desempenho escolar, em termos de interceptos e inclinações temporais, entre as escolas.

Caracterizado o aninhamento temporal dos estudantes pertencentes a diferentes escolas nos dados com medidas repetidas do nosso exemplo, vamos inicialmente estimar um modelo nulo (modelo não condicional), que permite que verifiquemos se existe variabilidade no desempenho escolar entre estudantes provenientes de uma mesma escola e entre aqueles provenientes de escolas distintas. Nenhuma variável explicativa será inserida na modelagem, que considera apenas a existência de um intercepto e dos termos de erro u_{00k}, r_{0jk} e e_{tjk}, com variâncias respectivamente iguais a τ_{u000}, τ_{r000} e σ^2. O modelo a ser estimado apresenta a seguinte expressão:

Modelo Nulo:

$$desempenho_{tjk} = \pi_{0jk} + e_{tjk}$$
$$\pi_{0jk} = b_{00k} + r_{0jk}$$
$$b_{00k} = \gamma_{000} + u_{00k}$$

que resulta em:

$$desempenho_{tjk} = \gamma_{000} + u_{00k} + r_{0jk} + e_{tjk}$$

O comando para a estimação deste modelo nulo no Stata é:

```
xtmixed desempenho || escola: || estudante: , var nolog reml
```

que, conforme podemos observar, apresenta agora dois componentes de efeitos aleatórios, sendo um correspondente ao nível 3 (escola) e outro ao nível 2 (estudante). É importante frisar que **a ordem de inserção dos componentes de efeitos aleatórios no comando `xtmixed` é decrescente na existência de mais de dois níveis**, ou seja, devemos iniciar com o nível superior de aninhamento dos dados e seguir até o nível inferior (nível 2). Os *outputs* obtidos são apresentados na Figura 16.30.

```
. xtmixed desempenho || escola: || estudante: , var nolog reml

Mixed-effects REML regression                   Number of obs      =      2440

------------------------------------------------------------
                |   No. of       Observations per Group
 Group Variable |   Groups    Minimum    Average    Maximum
----------------+-------------------------------------------
         escola |       15         80      162.7        248
      estudante |      610          4        4.0          4
------------------------------------------------------------

                                                Wald chi2(0)       =          .
Log restricted-likelihood = -9092.1387          Prob > chi2        =          .

------------------------------------------------------------------------------
  desempenho |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _cons |   68.71395   3.553167    19.34   0.000     61.74987    75.67803
------------------------------------------------------------------------------

------------------------------------------------------------------------------
  Random-effects Parameters  |   Estimate   Std. Err.     [95% Conf. Interval]
-----------------------------+------------------------------------------------
escola: Identity             |
                  var(_cons) |   180.1941   71.60437      82.69809    392.6319
-----------------------------+------------------------------------------------
estudante: Identity          |
                  var(_cons) |   325.7989   19.49574      289.7436    366.3408
-----------------------------+------------------------------------------------
               var(Residual) |    41.6494   1.376887      39.03632    44.43739
------------------------------------------------------------------------------
LR test vs. linear regression:       chi2(2) =  4036.13   Prob > chi2 = 0.0000

Note: LR test is conservative and provided only for reference.
```

Figura 16.30 *Outputs* do modelo nulo no Stata.

Na parte superior da Figura 16.30, podemos inicialmente comprovar que estamos diante de um painel balanceado, já que, para cada estudante, temos quantidades mínima e máxima de períodos de monitoramento iguais a quatro, com média também igual a quatro.

Em relação ao componente de efeitos fixos, podemos verificar que a estimação do parâmetro γ_{000} é igual a 68,714, que corresponde à média dos desempenhos escolares anuais esperados dos estudantes (reta horizontal estimada no modelo nulo, ou intercepto geral).

Já na parte inferior dos *outputs*, são apresentadas as estimações das variâncias dos termos de erro τ_{u000} = 180,194 (no Stata, **var(_cons)** para **escola**), τ_{r000} = 325,799 (no Stata, **var(_cons)** para **estudante**) e σ^2 = 41,649 (no Stata, **var(Residual)**).

Logo, podemos definir duas correlações intraclasse, dada a existência de duas proporções de variância, em que a primeira delas refere-se à correlação entre os dados da variável *desempenho* em t e em t' ($t \neq t'$) de determinado estudante j pertencente a determinada escola k (correlação intraclasse de nível 2), e a outra refere-se à correlação entre os dados da variável *desempenho* em t e em t' ($t \neq t'$) de diferentes estudantes j e j' ($j \neq j'$) pertencentes a determinada escola k (correlação intraclasse de nível 3). Neste sentido, temos que:

- **Correlação intraclasse de nível 2:**

$$rho_{estudante|escola} = corr\left(Y_{tjk}, Y_{t'jk}\right) = \frac{\tau_{u000} + \tau_{r000}}{\tau_{u000} + \tau_{r000} + \sigma^2} = \frac{180{,}194 + 325{,}799}{180{,}194 + 325{,}799 + 41{,}649} = 0{,}924$$

- **Correlação intraclasse de nível 3:**

$$rho_{escola} = corr\left(Y_{tjk}, Y_{t'j'k}\right) = \frac{\tau_{u000}}{\tau_{u000} + \tau_{r000} + \sigma^2} = \frac{180{,}194}{180{,}194 + 325{,}799 + 41{,}649} = 0{,}329$$

A partir da versão 13 do Stata, é possível obter diretamente essas correlações intraclasse, digitando-se o comando **estat icc** logo após a estimação do modelo correspondente.

Neste sentido, a correlação entre os desempenhos escolares anuais, para uma mesma escola, é igual a 32,9% (rho_{escola}) e a correlação entre os desempenhos escolares anuais, para um mesmo estudante de determinada escola, é igual a 92,4% ($rho_{estudante|escola}$). Para o modelo sem variáveis explicativas, portanto, enquanto o desempenho escolar anual é levemente correlacionado entre escolas, o mesmo passa a ser fortemente correlacionado quando o cálculo é feito para o mesmo estudante proveniente de determinada escola. Nesse último caso, estimamos que os efeitos aleatórios de estudantes e escolas compõem aproximadamente 92% da variância total dos resíduos!

Em relação à significância estatística dessas variâncias, o fato de os valores estimados de τ_{u000}, τ_{r000} e σ^2 serem consideravelmente superiores aos respectivos erros-padrão indica haver variação significativa no desempenho escolar anual entre estudantes e entre escolas. Mais especificamente, podemos verificar que todas essas relações são maiores do que 1,96, sendo esse o valor crítico da distribuição normal padrão que resulta em um nível de significância de 5%.

Conforme discutido na seção 16.4.1, essa informação é fundamental para embasar a escolha da modelagem multinível neste exemplo, em vez de uma simples e tradicional modelagem de regressão por MQO. Na parte inferior da Figura 16.30 podemos comprovar esse fato, analisando o resultado do teste de razão de verossimilhança (**LR test**). Como *Sig.* χ^2 = 0,000, podemos rejeitar a hipótese nula de que os interceptos aleatórios sejam iguais a zero (H_0: $u_{00k} = r_{0jk} = 0$), o que faz com que a estimação de um modelo tradicional de regressão linear seja descartada para os dados com medidas repetidas do nosso exemplo.

Embora pesquisadores frequentemente desprezem a estimação de modelos nulos, a análise dos resultados pode auxiliar na rejeição ou não de hipóteses de pesquisa e até mesmo propiciar ajustes em relação aos constructos propostos. Para os dados do nosso exemplo, os resultados do modelo nulo permitem que afirmemos que há variabilidade significativa no desempenho escolar ao longo dos quatro anos da análise, que há variabilidade significativa no desempenho escolar, ao longo do tempo, entre estudantes de uma mesma escola, e que há variabilidade significativa no desempenho escolar, ao longo do tempo, entre estudantes provenientes de escolas distintas. Esses achados podem, por si só, rejeitar ou comprovar hipóteses de pesquisa e

ser utilizados para a estruturação ser determinado trabalho, sem que, dependendo dos objetivos do pesquisador, seja necessária a elaboração de modelagens adicionais.

Como o nosso objetivo, além do exposto, é verificar se existem características dos estudantes e das escolas que explicam a variabilidade do desempenho escolar entre estudantes de uma mesma escola e entre aqueles provenientes de escolas distintas, seguiremos com os próximos passos da modelagem, respeitando a *multilevel step-up strategy*.

Neste sentido, assim como já preliminarmente visualizado por meio dos gráficos das Figuras 16.28 e 16.29, vamos inserir a variável de nível 1, *ano*, na análise, com o intuito de investigar se a variável temporal apresenta relação com o comportamento do desempenho escolar dos estudantes e, mais do que isso, se o desempenho escolar apresenta comportamento linear ao longo do tempo.

Modelo de Tendência Linear com Interceptos Aleatórios:

$$desempenho_{tjk} = \pi_{0jk} + \pi_{1jk}.ano_{jk} + e_{tjk}$$
$$\pi_{0jk} = b_{00k} + r_{0jk}$$
$$\pi_{1jk} = b_{10k}$$
$$b_{00k} = \gamma_{000} + u_{00k}$$
$$b_{10k} = \gamma_{100}$$

que resulta na seguinte expressão:

$$desempenho_{tjk} = \gamma_{000} + \gamma_{100}.ano_{jk} + u_{00k} + r_{0jk} + e_{tjk}$$

O comando para a estimação do modelo de tendência linear com interceptos aleatórios no Stata, para os dados do nosso exemplo, é:

```
xtmixed desempenho ano || escola: || estudante: , var nolog reml
```

cujos *outputs* são apresentados na Figura 16.31.

```
. xtmixed desempenho ano || escola: || estudante: , var nolog reml

Mixed-effects REML regression                   Number of obs      =      2440

------------------------------------------------------------------
                |   No. of       Observations per Group
 Group Variable |   Groups    Minimum    Average    Maximum
----------------+-------------------------------------------------
         escola |       15         80      162.7        248
      estudante |      610          4        4.0          4
------------------------------------------------------------------

                                                Wald chi2(1)       =   5683.02
Log restricted-likelihood = -7801.4202          Prob > chi2        =    0.0000

------------------------------------------------------------------------------
  desempenho |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         ano |   4.348016   .0576768    75.39   0.000     4.234972    4.461061
       _cons |   57.84391   3.556109    16.27   0.000     50.87407    64.81376
------------------------------------------------------------------------------

------------------------------------------------------------------------------
  Random-effects Parameters  |   Estimate   Std. Err.      [95% Conf. Interval]
-----------------------------+------------------------------------------------
escola: Identity             |
                  var(_cons) |   180.1959   71.60532      82.69876    392.6368
-----------------------------+------------------------------------------------
estudante: Identity          |
                  var(_cons) |   333.6753   19.49293      297.5759    374.1539
-----------------------------+------------------------------------------------
               var(Residual) |   10.14618   .3355141      9.509446    10.82556
------------------------------------------------------------------------------
LR test vs. linear regression:       chi2(2) =  6505.83   Prob > chi2 = 0.0000

Note: LR test is conservative and provided only for reference.
```

Figura 16.31 *Outputs* do modelo de tendência linear com interceptos aleatórios.

Inicialmente, podemos verificar que a média de crescimento anual do desempenho escolar é estatisticamente significante e com parâmetro estimado de $\gamma_{100} = 4{,}348$, *ceteris paribus*.

Em relação aos componentes de efeitos aleatórios, também verificamos a existência de significância estatística das variâncias de u_{00k}, r_{0jk} e e_{tjk}, pelo fato de as estimações de τ_{u000}, τ_{r000} e σ^2 serem consideravelmente superiores aos respectivos erros-padrão. Neste sentido, novas correlações intraclasse podem ser calculadas, conforme segue:

- **Correlação intraclasse de nível 2:**

$$rho_{estudante|escola} = corr\left(Y_{tjk}, Y_{t'jk}\right) = \frac{\tau_{u000} + \tau_{r000}}{\tau_{u000} + \tau_{r000} + \sigma^2} = \frac{180{,}196 + 333{,}675}{180{,}196 + 333{,}675 + 10{,}146} = 0{,}981$$

- **Correlação intraclasse de nível 3:**

$$rho_{escola} = corr\left(Y_{tjk}, Y_{t'j'k}\right) = \frac{\tau_{u000}}{\tau_{u000} + \tau_{r000} + \sigma^2} = \frac{180{,}196}{180{,}196 + 333{,}675 + 10{,}146} = 0{,}344$$

As duas proporções de variância são mais elevadas do que aquelas obtidas na estimação do modelo nulo, o que demonstra a importância da inclusão da variável correspondente à medida repetida no nível 1. Além disso, o resultado do teste de razão de verossimilhança (**LR test**) na parte inferior da Figura 16.31 permite que comprovemos que seja descartada a estimação de um modelo tradicional de regressão linear simples (*desempenho* em função de *ano*) apenas com efeitos fixos.

O nosso modelo, portanto, passa a ter, no presente momento, a seguinte especificação:

$$desempenho_{tjk} = 57{,}844 + 4{,}348.ano_{jk} + u_{00k} + r_{0jk} + e_{tjk}$$

Na sequência, podemos arquivar (comando **estimates store**) as estimações obtidas para futura comparação com as que serão geradas na estimação de um modelo de tendência linear com interceptos e inclinações aleatórias. Podemos também obter, por meio do comando **predict, reffects**, os valores esperados dos efeitos aleatórios *BLUPS* (*best linear unbiased predictions*) u_{00k} e r_{0jk}. Mantendo a lógica proposta na seção anterior, vamos digitar a seguinte sequência de comandos:

```
estimates store interceptoaleat
predict u00 r0, reffects
desc u00 r0
by estudante, sort: generate tolist = (_n==1)
list estudante escola u00 r0 if escola <=2 & tolist
```

A Figura 16.32 apresenta os valores dos termos de interceptos aleatórios u_{00k} e r_{0jk} para os estudantes das duas primeiras escolas da base de dados. Podemos verificar que, enquanto os termos de erro u_{00k} são invariantes para estudantes da mesma escola e ao longo do tempo (variável *u00* gerada na base de dados), os termos r_{0jk} variam entre estudantes, porém são invariantes para um mesmo estudante ao longo do tempo (variável *r0* gerada na base de dados), o que caracteriza a existência de um intercepto para cada estudante e um intercepto para cada escola.

```
. estimates store interceptoaleat
. predict u00 r0, reffects
. desc u00 r0

              storage  display     value
variable name   type   format      label      variable label
------------------------------------------------------------------------------
u00             float  %9.0g                  BLUP r.e. for escola: _cons
r0              float  %9.0g                  BLUP r.e. for estudante: _cons

. by estudante, sort: generate tolist = (_n==1)
. list estudante escola u00 r0 if escola <=2 & tolist

     +-----------------------------------------+     +-----------------------------------------+
     | estuda~e   escola        u00         r0 |     | estuda~e   escola        u00         r0 |
     |-----------------------------------------|     |-----------------------------------------|
  1. |        1        1   -10.8088   -13.15515 |  161. |       41        1   -10.8088   -17.86931 |
  5. |        2        1   -10.8088    19.09966 |  165. |       42        1   -10.8088   -25.06462 |
  9. |        3        1   -10.8088    35.84734 |  169. |       43        1   -10.8088    16.27116 |
 13. |        4        1   -10.8088   -2.932857 |  173. |       44        1   -10.8088   -17.42271 |
 17. |        5        1   -10.8088    31.30685 |  177. |       45        1   -10.8088    36.07064 |
     |-----------------------------------------|     |-----------------------------------------|
 21. |        6        1   -10.8088   -5.413996 |  181. |       46        1   -10.8088    8.728494 |
 25. |        7        1   -10.8088   -42.08523 |  185. |       47        1   -10.8088   -28.63746 |
 29. |        8        1   -10.8088   -24.61801 |  189. |       48        2    1.580118   -19.89285 |
 33. |        9        1   -10.8088   -24.56839 |  193. |       49        2    1.580118   -3.169975 |
 37. |       10        1   -10.8088    39.09763 |  197. |       50        2    1.580118    6.556092 |
     |-----------------------------------------|     |-----------------------------------------|
 41. |       11        1   -10.8088   -7.895134 |  201. |       51        2    1.580118    24.07293 |
 45. |       12        1   -10.8088    16.22153 |  205. |       52        2    1.580118   -16.56812 |
 49. |       13        1   -10.8088    37.13753 |  209. |       53        2    1.580118   -1.979025 |
 53. |       14        1   -10.8088    24.60778 |  213. |       54        2    1.580118    20.99632 |
 57. |       15        1   -10.8088    37.08791 |  217. |       55        2    1.580118   -13.78925 |
     |-----------------------------------------|     |-----------------------------------------|
 61. |       16        1   -10.8088    22.12664 |  221. |       56        2    1.580118   -16.86586 |
 65. |       17        1   -10.8088    27.93251 |  225. |       57        2    1.580118    13.65215 |
 69. |       18        1   -10.8088   -11.41835 |  229. |       58        2    1.580118   -26.49268 |
 73. |       19        1   -10.8088   -25.06462 |  233. |       59        2    1.580118   -34.33308 |
 77. |       20        1   -10.8088    19.94324 |  237. |       60        2    1.580118    15.04158 |
     |-----------------------------------------|     |-----------------------------------------|
 81. |       21        1   -10.8088    7.140564 |  241. |       61        2    1.580118    -14.7817 |
 85. |       22        1   -10.8088   -10.27703 |  245. |       62        2    1.580118   -38.65026 |
 89. |       23        1   -10.8088   -5.910223 |  249. |       63        2    1.580118    18.46556 |
 93. |       24        1   -10.8088    -15.4378 |  253. |       64        2    1.580118    22.68349 |
 97. |       25        1   -10.8088   -18.56403 |  257. |       65        2    1.580118    6.357596 |
     |-----------------------------------------|     |-----------------------------------------|
101. |       26        1   -10.8088    34.18498 |  261. |       66        2    1.580118    14.54535 |
105. |       27        1   -10.8088   -17.22422 |  265. |       67        2    1.580118    26.15709 |
109. |       28        1   -10.8088   -12.16269 |  269. |       68        2    1.580118   -10.86151 |
113. |       29        1   -10.8088   -9.731179 |  273. |       69        2    1.580118    19.50763 |
117. |       30        1   -10.8088   -1.642665 |  277. |       70        2    1.580118   -23.06871 |
     |-----------------------------------------|     |-----------------------------------------|
121. |       31        1   -10.8088   -18.46479 |  281. |       71        2    1.580118    28.48936 |
125. |       32        1   -10.8088   -24.22103 |  285. |       72        2    1.580118    6.853824 |
129. |       33        1   -10.8088    4.411312 |     +-----------------------------------------+
133. |       34        1   -10.8088    8.033776 |
137. |       35        1   -10.8088    10.21718 |
     |-----------------------------------------|
141. |       36        1   -10.8088   -17.67082 |
145. |       37        1   -10.8088    -.352474 |
149. |       38        1   -10.8088   -22.43461 |
153. |       39        1   -10.8088   -28.93519 |
157. |       40        1   -10.8088   -6.307207 |
     |-----------------------------------------|
```

Figura 16.32 Termos de interceptos aleatórios u_{00k} e r_{0jk} para as duas primeiras escolas da amostra (em destaque, a identificação da observação correspondente ao primeiro período de tempo de cada estudante).

A fim de propiciar melhor visualização dos interceptos aleatórios por escola e por estudante, podemos gerar dois gráficos (Figuras 16.33 e 16.34), digitando os seguintes comandos:

```
graph hbar (mean) u00, over(escola) ytitle("Interceptos Aleatórios por Escola")
```

```
graph hbar (mean) r0, over(estudante) ytitle("Interceptos Aleatórios por Estudante")
```

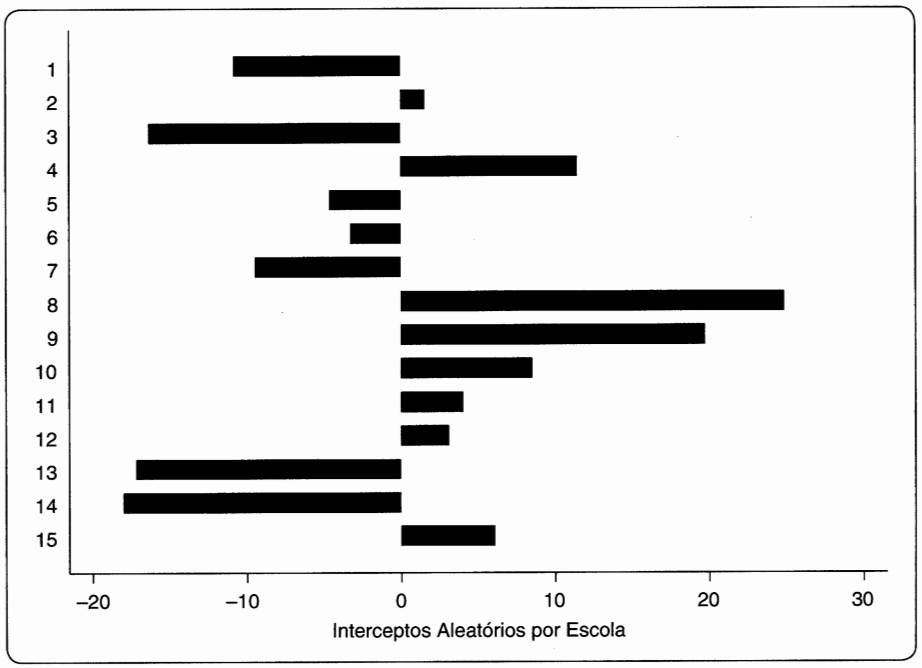


Figura 16.33 Interceptos aleatórios por escola.

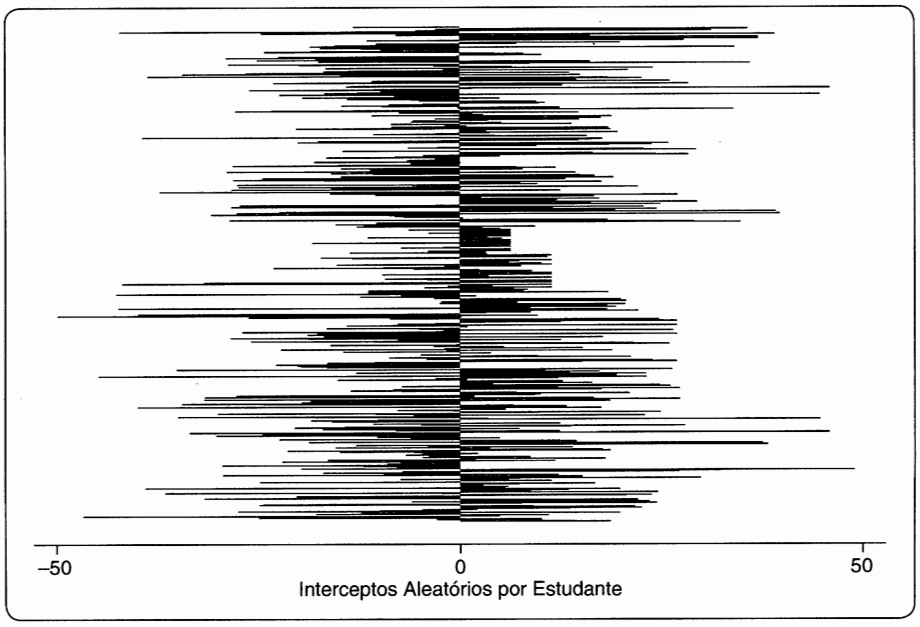


Figura 16.34 Interceptos aleatórios por estudante.

Neste momento da modelagem, portanto, temos condições de afirmar que o desempenho escolar dos estudantes segue uma tendência linear ao longo do tempo, existindo variância significativa de interceptos entre aqueles que estudam na mesma escola e entre aqueles que estudam em escolas distintas.

Precisamos, assim, também verificar se existe variância significativa de inclinações do desempenho escolar ao longo do tempo entre os diferentes estudantes, já que os gráficos das Figuras 16.26 e 16.29 já nos ofereciam indícios de ocorrência desse fenômeno. Portanto, vamos inserir efeitos aleatórios de inclinação nos níveis 2 e 3 do nosso modelo multinível que, com a manutenção dos efeitos aleatórios de intercepto, passará a ter a seguinte expressão:

Modelo de Tendência Linear com Interceptos e Inclinações Aleatórias:

$$desempenho_{tjk} = \pi_{0jk} + \pi_{1jk}.ano_{jk} + e_{tjk}$$
$$\pi_{0jk} = b_{00k} + r_{0jk}$$
$$\pi_{1jk} = b_{10k} + r_{1jk}$$
$$b_{00k} = \gamma_{000} + u_{00k}$$
$$b_{10k} = \gamma_{100} + u_{10k}$$

que resulta em:

$$desempenho_{tjk} = \gamma_{000} + \gamma_{100}.ano_{jk} + u_{00k} + u_{10k}.ano_{jk} + r_{0jk} + r_{1jk}.ano_{jk} + e_{tjk}$$

O comando para estimação deste modelo de tendência linear com interceptos e inclinações aleatórias no Stata é:

```
xtmixed desempenho ano || escola: ano || estudante: ano, var nolog reml
```

Note agora que a variável *ano* está presente no componente de efeitos fixos e nos componentes de efeitos aleatórios de nível 3 (multiplicando o termo de erro u_{10k}) e de nível 2 (multiplicando o termo de erro r_{1jk}). Os *outputs* obtidos são apresentados na Figura 16.35.

```
. xtmixed desempenho ano || escola: ano || estudante: ano, var nolog reml

Mixed-effects REML regression                   Number of obs      =      2440

------------------------------------------------------------
                |   No. of       Observations per Group
 Group Variable |   Groups    Minimum    Average    Maximum
----------------+-------------------------------------------
         escola |       15         80      162.7        248
      estudante |      610          4        4.0          4
------------------------------------------------------------

                                                Wald chi2(1)       =    424.89
Log restricted-likelihood =  -7464.819          Prob > chi2        =    0.0000

------------------------------------------------------------------------------
  desempenho |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         ano |   4.343297   .2107073    20.61   0.000     3.930318    4.756276
       _cons |   57.85776   3.955816    14.63   0.000      50.1045    65.61102
------------------------------------------------------------------------------

------------------------------------------------------------------------------
  Random-effects Parameters  |   Estimate   Std. Err.      [95% Conf. Interval]
-----------------------------+------------------------------------------------
escola: Independent          |
                   var(ano) |   .5600495   .2519118      .2319283    1.352381
                 var(_cons) |   224.3434   88.72199       103.344     487.014
-----------------------------+------------------------------------------------
estudante: Independent       |
                   var(ano) |   3.157275   .2305444      2.736261    3.643067
                 var(_cons) |   374.2847   22.00905      333.5408    420.0058
-----------------------------+------------------------------------------------
              var(Residual) |   3.867725   .1595253      3.567365    4.193374
------------------------------------------------------------------------------
LR test vs. linear regression:       chi2(4) =  7179.03   Prob > chi2 = 0.0000

Note: LR test is conservative and provided only for reference.
```

Figura 16.35 *Outputs* do modelo de tendência linear com interceptos e inclinações aleatórias.

Podemos verificar que, embora as estimações dos parâmetros de efeitos fixos não se alterem consideravelmente em relação ao modelo anterior, as estimações das variâncias são diferentes, o que gera novas correlações intraclasse, conforme segue:

- **Correlação intraclasse de nível 2:**

$$rho_{estudante|escola} = corr\left(Y_{tjk}, Y_{t'jk}\right) = \frac{\tau_{u000} + \tau_{u100} + \tau_{r000} + \tau_{r100}}{\tau_{u000} + \tau_{u100} + \tau_{r000} + \tau_{r100} + \sigma^2}$$
$$= \frac{224,343 + 0,560 + 374,285 + 3,157}{224,343 + 0,560 + 374,285 + 3,157 + 3,868} = 0,994$$

- **Correlação intraclasse de nível 3:**

$$rho_{escola} = corr\left(Y_{tjk}, Y_{t'j'k}\right) = \frac{\tau_{u000} + \tau_{u100}}{\tau_{u000} + \tau_{u100} + \tau_{r000} + \tau_{r100} + \sigma^2}$$
$$= \frac{224,343 + 0,560}{224,343 + 0,560 + 374,285 + 3,157 + 3,868} = 0,371$$

Logo, para este modelo, estimamos que os efeitos aleatórios de estudantes e escolas compõem aproximadamente 99% da variância total dos resíduos!

Vamos digitar o seguinte comando, a fim de que possamos comprovar a melhor adequação dessa estimação sobre a estimação anterior, sem inclinações aleatórias:

```
estimates store inclinaçãoaleat
```

Na sequência, podemos digitar o comando que irá elaborar o teste de razão de verossimilhança:

```
lrtest inclinaçãoaleat interceptoaleat
```

já que o termo **interceptoaleat** refere-se à estimação já realizada anteriormente. O resultado do teste é apresentado na Figura 16.36.

```
. lrtest inclinaçãoaleat interceptoaleat

Likelihood-ratio test                                 LR chi2(2)  =    673.20
(Assumption: interceptoal~t nested in inclinaçãoal~t) Prob > chi2 =    0.0000

Note: The reported degrees of freedom assumes the null hypothesis is not on
the boundary of the parameter space.  If this is not true, then the reported
test is conservative.
Note: LR tests based on REML are valid only when the fixed-effects specification
is identical for both models.
```

Figura 16.36 Teste de razão de verossimilhança para comparar as estimações dos modelos de tendência linear com interceptos aleatórios e com interceptos e inclinações aleatórias.

Fazendo uso dos valores obtidos da função de verossimilhança restrita nas Figuras 16.31 e 16.35, chegamos à seguinte estatística χ^2 do teste, com 2 graus de liberdade:

$$\chi^2_2 = [-2.LL_{r\text{-}interceptoaleat} - (-2.LL_{r\text{-}inclinaçãoaleat})] = \{-2.(-7.801,420) - [-2.(-7.464,819)]\} = 673,20$$

que resulta em um *Sig.* $\chi^2_2 = 0,000 < 0,05$ e acaba por favorecer o modelo de tendência linear com interceptos e inclinações aleatórias.Vale novamente frisar, conforme também explicita a nota na parte inferior da Figura 16.36, que este teste de razão de verossimilhança somente é válido quando for feita a comparação das estimações obtidas por máxima verossimilhança restrita (*REML*) de dois modelos com especificação idêntica do componente de efeitos fixos. Como, no nosso caso, os dois modelos, que foram estimados por *REML*, apresentam a mesma especificação $\gamma_{000} + \gamma_{100}.ano_{jk}$ no componente de efeitos fixos, o teste é considerado válido.

Portanto, o nosso modelo passa a ter a seguinte especificação:

$$desempenho_{tjk} = 57,858 + 4,343.ano_{jk} + u_{00k} + u_{10k}.ano_{jk} + r_{0jk} + r_{1jk}.ano_{jk} + e_{tjk}$$

Na presente situação, temos condições de afirmar que o desempenho escolar dos estudantes segue uma tendência linear ao longo do tempo, existindo variância significativa de interceptos e de inclinações entre aqueles que estudam na mesma escola e entre aqueles que estudam em escolas distintas.

Desta forma, vamos inserir a variável *sexo*, de nível 2, na análise, a fim de verificarmos se essa característica explica a variação no desempenho escolar anual entre os estudantes.

Modelo de Tendência Linear com Interceptos e Inclinações Aleatórias e a Variável *sexo* de Nível 2:

$$desempenho_{tjk} = \pi_{0jk} + \pi_{1jk}.ano_{jk} + e_{tjk}$$
$$\pi_{0jk} = b_{00k} + b_{01k}.sexo_{jk} + r_{0jk}$$
$$\pi_{1jk} = b_{10k} + b_{11k}.sexo_{jk} + r_{1jk}$$
$$b_{00k} = \gamma_{000} + u_{00k}$$
$$b_{01k} = \gamma_{010}$$
$$b_{10k} = \gamma_{100} + u_{10k}$$
$$b_{11k} = \gamma_{110}$$

que resulta na seguinte expressão:

$$desempenho_{tjk} = \gamma_{000} + \gamma_{100}.ano_{jk} + \gamma_{010}.sexo_{jk} + \gamma_{110}.sexo_{jk}.ano_{jk} + u_{00k} + u_{10k}.ano_{jk} + r_{0jk} + r_{1jk}.ano_{jk} + e_{tjk}$$

Precisamos, inicialmente, gerar uma nova variável que corresponde à multiplicação de *sexo* por *ano*. O comando a seguir gera esta variável (*sexoano*):

```
gen sexoano = sexo*ano
```

Na sequência, podemos estimar o nosso modelo de tendência linear com interceptos e inclinações aleatórias e a variável *sexo* de nível 2, digitando o seguinte comando:

```
xtmixed desempenho ano sexo sexoano || escola: ano || estudante: ano, var nolog reml
```

Os *outputs* gerados são apresentados na Figura 16.37.

```
. xtmixed desempenho ano sexo sexoano || escola: ano || estudante: ano, var
nolog reml

Mixed-effects REML regression                   Number of obs      =      2440

------------------------------------------------------------
                |   No. of       Observations per Group
 Group Variable |   Groups    Minimum    Average    Maximum
----------------+-------------------------------------------
         escola |       15         80      162.7        248
      estudante |      610          4        4.0          4
------------------------------------------------------------

                                                Wald chi2(3)       =    633.54
Log restricted-likelihood = -7424.2732          Prob > chi2        =    0.0000

------------------------------------------------------------------------------
  desempenho |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         ano |   4.028844   .2024281    19.90   0.000     3.632092    4.425595
        sexo |  -15.03265   1.766749    -8.51   0.000    -18.49542   -11.56989
     sexoano |   .7050945   .1827647     3.86   0.000     .3468824    1.063307
       _cons |   64.49828   3.465572    18.61   0.000     57.70589    71.29068
------------------------------------------------------------------------------

------------------------------------------------------------------------------
  Random-effects Parameters  |   Estimate   Std. Err.      [95% Conf. Interval]
-----------------------------+------------------------------------------------
escola: Independent          |
                   var(ano) |   .4113062   .1977923      .1602627    1.055597
                 var(_cons) |   161.6346   64.79808      73.67059    354.6293
-----------------------------+------------------------------------------------
estudante: Independent       |
                   var(ano) |   3.096463   .2272074      2.681685    3.575395
                 var(_cons) |   337.7062    19.9023       300.867    379.0562
-----------------------------+------------------------------------------------
              var(Residual) |   3.867745   .1594995      3.567432    4.193339
------------------------------------------------------------------------------
LR test vs. linear regression:       chi2(4) =  6850.06   Prob > chi2 = 0.0000

Note: LR test is conservative and provided only for reference.
```

Figura 16.37 *Outputs* do modelo de tendência linear com interceptos e inclinações aleatórias e a variável *sexo* de nível 2.

Este modelo apresenta estimações significantes, tanto dos parâmetros de efeitos fixos, quanto das variâncias dos termos de efeitos aleatórios, ao nível de significância de 5%, e, neste momento da modelagem, temos condições de afirmar que o desempenho escolar dos estudantes segue uma tendência linear ao longo do tempo, existindo variância significativa de interceptos e de inclinações entre aqueles que estudam na mesma escola e entre aqueles que estudam em escolas distintas e, mais do que isso, o fato de determinado estudante ser do sexo feminino ou masculino é parte da razão de existência dessa variação no desempenho escolar.

O modelo passa a ter a seguinte especificação:

$$desempenho_{tjk} = 64,498 + 4,029.ano_{jk} - 15,033.sexo_{jk} + 0,705.sexo_{jk}.ano_{jk} + u_{00k} + u_{10k}.ano_{jk} + r_{0jk} + r_{1jk}.ano_{jk} + e_{tjk}$$

e, pela qual, podemos verificar que estudantes do sexo masculino (*dummy sexo* = 1) apresentam, em média e *ceteris paribus*, desempenhos piores do que os do sexo feminino.

Vamos, por fim, investigar se a variável *texp*, de nível 3 (tempo médio de experiência docente dos professores da escola, em anos), também explica a variação no desempenho escolar anual entre os estudantes. Após algumas análises intermediárias, partiremos para a estimação do modelo hierárquico de três níveis com a seguinte especificação:

Modelo de Tendência Linear com Interceptos e Inclinações Aleatórias e as Variáveis *sexo* de Nível 2 e *texp* de Nível 3 (Modelo Completo):

$$\begin{aligned}
desempenho_{tjk} &= \pi_{0jk} + \pi_{1jk}.ano_{jk} + e_{tjk} \\
\pi_{0jk} &= b_{00k} + b_{01k}.sexo_{jk} + r_{0jk} \\
\pi_{1jk} &= b_{10k} + b_{11k}.sexo_{jk} + r_{1jk} \\
b_{00k} &= \gamma_{000} + \gamma_{001}.texp_k + u_{00k} \\
b_{01k} &= \gamma_{010} \\
b_{10k} &= \gamma_{100} + \gamma_{101}.texp_k + u_{10k} \\
b_{11k} &= \gamma_{110}
\end{aligned}$$

que resulta na seguinte expressão:

$$\begin{aligned}
desempenho_{tjk} = {} & \gamma_{000} + \gamma_{100}.ano_{jk} + \gamma_{010}.sexo_{jk} + \gamma_{001}.texp_k \\
& + \gamma_{110}.sexo_{jk}.ano_{jk} + \gamma_{101}.texp_k.ano_{jk} \\
& + u_{00k} + u_{10k}.ano_{jk} + r_{0jk} + r_{1jk}.ano_{jk} + e_{tjk}
\end{aligned}$$

Para estimarmos esse modelo, é preciso que criemos mais uma nova variável (*texpano*), correspondente à multiplicação de *texp* por *ano*. Vamos então digitar o seguinte comando:

```
gen texpano = texp*ano
```

Assim, podemos estimar o modelo proposto digitando o seguinte comando:

```
xtmixed desempenho ano sexo texp sexoano texpano || escola: ano ||
estudante: ano, var nolog reml
```

cujos *outputs* são apresentados na Figura 16.38.

```
. xtmixed desempenho ano sexo texp sexoano texpano || escola: ano || estudante:
ano, var nolog reml

Mixed-effects REML regression                   Number of obs      =      2440

------------------------------------------------------------
                |   No. of       Observations per Group
 Group Variable |   Groups    Minimum    Average    Maximum
----------------+-------------------------------------------
         escola |       15         80      162.7        248
      estudante |      610          4        4.0          4
------------------------------------------------------------

                                                Wald chi2(5)       =    883.26
Log restricted-likelihood = -7419.6785          Prob > chi2        =    0.0000

------------------------------------------------------------------------------
  desempenho |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         ano |    4.528292   .2586443    17.51   0.000     4.021359    5.035226
        sexo |   -14.69529   1.762759    -8.34   0.000    -18.15024   -11.24035
        texp |    1.179424    .343969     3.43   0.001     .5052567    1.85359
    sexoano |    .6485018   .1828469     3.55   0.000     .2901286    1.006875
    texpano |   -.0570213   .0211086    -2.70   0.007    -.0983934   -.0156491
       _cons |    54.72215   3.925206    13.94   0.000     47.02889    62.41541
------------------------------------------------------------------------------

------------------------------------------------------------------------------
  Random-effects Parameters  |   Estimate   Std. Err.     [95% Conf. Interval]
-----------------------------+------------------------------------------------
escola: Independent          |
                  var(ano)   |    .262667   .1394859      .0927653    .7437469
                var(_cons)   |   87.99372   37.97699       37.7645     205.031
-----------------------------+------------------------------------------------
estudante: Independent       |
                  var(ano)   |   3.092474   .2267585      2.678496    3.570436
                var(_cons)   |   337.6269   19.89377      300.8031    378.9587
-----------------------------+------------------------------------------------
             var(Residual)   |   3.867764   .1595005      3.567449     4.19336
------------------------------------------------------------------------------
LR test vs. linear regression:       chi2(4) =  6557.63   Prob > chi2 = 0.0000

Note: LR test is conservative and provided only for reference.
```

Figura 16.38 *Outputs* do modelo de tendência linear com interceptos e inclinações aleatórias e as variáveis *sexo* de nível 2 e *texp* de nível 3.

Embora as estimações dos parâmetros de efeitos fixos e das variâncias dos termos aleatórios sejam significantes, ao nível de significância de 5%, é preciso que estudemos a estrutura das matrizes de variância-covariância dos efeitos aleatórios (u_{00k}, u_{10k}) e (r_{0jk}, r_{1jk}). Com base nos *outputs* da Figura 16.38, temos que:

- **Matriz de variância-covariância dos efeitos aleatórios para o nível *escola*:**

$$var\begin{bmatrix} u_{00k} \\ u_{10k} \end{bmatrix} = \begin{bmatrix} 87,994 & 0 \\ 0 & 0,263 \end{bmatrix}$$

- **Matriz de variância-covariância dos efeitos aleatórios para o nível *estudante*:**

$$var\begin{bmatrix} r_{0jk} \\ r_{1jk} \end{bmatrix} = \begin{bmatrix} 337,627 & 0 \\ 0 & 3,092 \end{bmatrix}$$

Vamos arquivar os resultados desta estimação, digitando:

```
estimates store finalindependente
```

Como não especificamos nenhuma estrutura de covariância para esses termos de erro, o Stata pressupõe, na elaboração do comando **xtmixed**, que esta estrutura seja independente, ou seja, que cov(u_{00k}, u_{10k}) = 0 e que cov(r_{0jk}, r_{1jk}) = 0. Entretanto, podemos generalizar a estrutura dessas matrizes, permitindo que u_{00k} e u_{10k} sejam correlacionados e que r_{0jk} e r_{1jk} também sejam correlacionados. Para tanto, é preciso que adicionemos, no comando **xtmixed**, o termo **covariance(unstructured)** nos componentes de efeitos aleatórios do nível *escola* e do nível *estudante*, de modo que:

```
xtmixed desempenho ano sexo texp sexoano texpano || escola: ano,
covariance(unstructured) || estudante: ano, covariance(unstructured)
var nolog reml
```

que gera os *outputs* da Figura 16.39.

```
. xtmixed desempenho ano sexo texp sexoano texpano || escola: ano,
covariance(unstructured) || estudante: ano, covariance(unstructured) var nolog reml

Mixed-effects REML regression                   Number of obs      =      2440

-------------------------------------------------------------
                |   No. of       Observations per Group
 Group Variable |   Groups    Minimum    Average    Maximum
----------------+--------------------------------------------
         escola |       15         80      162.7        248
      estudante |      610          4        4.0          4
-------------------------------------------------------------

                                                Wald chi2(5)       =    868.08
Log restricted-likelihood = -7376.7147          Prob > chi2        =    0.0000

------------------------------------------------------------------------------
  desempenho |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         ano |   4.515641   .2583749    17.48   0.000     4.009236    5.022047
        sexo |  -14.70213   1.795536    -8.19   0.000    -18.22131   -11.18294
        texp |   1.178656   .3459065     3.41   0.001     .5006918    1.856621
    sexoano |   .6518855   .1847166     3.53   0.000     .2898477    1.013923
    texpano |  -.0566496   .0209988    -2.70   0.007    -.0978065   -.0154928
      _cons |   54.73435   3.951437    13.85   0.000     46.98968    62.47902
------------------------------------------------------------------------------

------------------------------------------------------------------------------
  Random-effects Parameters  |   Estimate   Std. Err.      [95% Conf. Interval]
-----------------------------+------------------------------------------------
escola: Unstructured         |
                   var(ano) |   .2554224   .1378072      .0887183    .7353682
                 var(_cons) |   88.7366    38.40337      37.99447    207.2456
              cov(ano,_cons) |  -3.185306   1.904226      -6.91752    .5469079
-----------------------------+------------------------------------------------
estudante: Unstructured      |
                   var(ano) |     3.2575   .2350138      2.827965    3.752276
                 var(_cons) |   350.9127   20.68884      312.6185    393.8978
              cov(ano,_cons) |  -13.25089   1.673704     -16.53129   -9.970494
-----------------------------+------------------------------------------------
               var(Residual) |   3.795043   .1536567      3.505521    4.108476
------------------------------------------------------------------------------
LR test vs. linear regression:       chi2(6) =  6643.55   Prob > chi2 = 0.0000

Note: LR test is conservative and provided only for reference.
```

Figura 16.39 *Outputs* do modelo de tendência linear com interceptos e inclinações aleatórias e as variáveis *sexo* de nível 2 e *texp* de nível 3, com termos aleatórios (u_{00k}, u_{10k}) e (r_{0jk}, r_{1jk}) correlacionados.

As estimações dos parâmetros de efeitos fixos são bastante próximas daquelas obtidas na estimação do modelo que considera a existência de estrutura independente das matrizes de variância-covariância dos termos aleatórios (Figura 16.38).

Já em relação aos parâmetros de efeitos aleatórios, com exceção das estimações de u_{10k} e de $\text{cov}(u_{00k}, u_{10k})$, que são estatisticamente significantes ao nível de significância de 10% (já que os respectivos $|z| > 1{,}64$, sendo esse o valor crítico da distribuição normal padrão que resulta em um nível de significância de 10%), todas as demais estimações são significantes ao nível de significância de 5%. Com finalidade didática, adotaremos o nível de confiança de 90% para darmos sequência à análise.

Neste sentido, considerando que $\text{cov}(u_{00k}, u_{10k})$ e $\text{cov}(r_{0jk}, r_{1jk})$ sejam estatisticamente diferentes de zero, com base nos *outputs* da Figura 16.39 podemos escrever que:

- **Matriz de variância-covariância dos efeitos aleatórios para o nível *escola*:**

$$var\begin{bmatrix} u_{00k} \\ u_{10k} \end{bmatrix} = \begin{bmatrix} 88{,}737 & -3{,}185 \\ -3{,}185 & 0{,}255 \end{bmatrix}$$

- **Matriz de variância-covariância dos efeitos aleatórios para o nível *estudante*:**

$$var\begin{bmatrix} r_{0jk} \\ r_{1jk} \end{bmatrix} = \begin{bmatrix} 350,913 & -13,251 \\ -13,251 & 3,258 \end{bmatrix}$$

O pesquisador também obterá essas matrizes caso digite o seguinte comando logo após a última estimação:

```
estat recovariance
```

cujos *outputs* são apresentados na Figura 16.40.

```
. estat recovariance

Random-effects covariance matrix for level escola

             |        ano       _cons
-------------+----------------------
         ano |   .2554224
       _cons |  -3.185306     88.7366

Random-effects covariance matrix for level estudante

             |        ano       _cons
-------------+----------------------
         ano |     3.2575
       _cons |  -13.25089    350.9127
```

Figura 16.40 Matrizes de variância-covariância com termos aleatórios (u_{00k}, u_{10k}) e (r_{0jk}, r_{1jk}) correlacionados.

Mesmo estatisticamente diferentes de zero as estimações das covariâncias dos termos aleatórios nos dois níveis da análise, se o pesquisador desejar comprovar a melhor adequação deste último modelo sobre aquele que considera a matriz com termos de erro independentes, basta que elabore um teste de razão de verossimilhança para comparar as duas estimações.

Com tal finalidade, vamos primeiramente digitar o seguinte comando, referente à estimação com termos aleatórios *unstructured*:

```
estimates store finalunstructured
```

Na sequência, podemos digitar o comando para realização do referido teste:

```
lrtest finalunstructured finalindependente
```

O resultado é apresentado na Figura 16.41.

```
. lrtest finalunstructured finalindependente

Likelihood-ratio test                                 LR chi2(2)  =     85.93
(Assumption: finalindepen~e nested in finalunstruc~d) Prob > chi2 =    0.0000

Note: LR tests based on REML are valid only when the fixed-effects specification
is identical for both models.
```

Figura 16.41 Teste de razão de verossimilhança para comparar as estimações dos modelos completos com termos aleatórios (u_{00k}, u_{10k}) e (r_{0jk}, r_{1jk}) independentes e correlacionados.

A estatística χ^2 deste teste, com 2 graus de liberdade, também pode ser obtida por meio da seguinte expressão:

$$\chi^2_2 = [-2.LL_{r\text{-}ind} - (-2.LL_{r\text{-}unstruc})] = \{-2.(-7.419,679) - [-2.(-7.376,715)]\} = 85,93$$

que resulta em um *Sig.* $\chi^2_2 = 0,000 < 0,05$. Portanto, podemos afirmar que a estrutura das matrizes de variância-covariância dos termos aleatórios pode ser considerada *unstructured* neste exemplo, ou seja, podemos considerar que os termos de erro u_{00k} e u_{10k} sejam correlacionados ($\text{cov}(u_{00k}, u_{10k}) \neq 0$) e que os termos de erro r_{0jk} e r_{1jk} também sejam correlacionados ($\text{cov}(r_{0jk}, r_{1jk}) \neq 0$).

Chegamos ao nosso modelo final, com a seguinte especificação:

$$desempenho_{tjk} = 54,734 + 4,516.ano_{jk} - 14,702.sexo_{jk} + 1,179.texp_k + 0,652.sexo_{jk}.ano_{jk} - 0,057.texp_k.ano_{jk} + u_{00k} + u_{10k}.ano_{jk} + r_{0jk} + r_{1jk}.ano_{jk} + e_{tjk}$$

Na sequência, podemos obter os valores esperados *BLUPS* (*best linear unbiased predictions*) dos efeitos aleatórios u_{10k}, u_{00k}, r_{1jk} e r_{0jk} do nosso modelo final, digitando:

```
predict u10final u00final r1final r0final, reffects
```

que gera no banco de dados quatro novas variáveis, denominadas *u10final*, *u00final*, *r1final* e *r0final* que correspondem, respectivamente, aos efeitos aleatórios de inclinação e de intercepto do nível *escola* e aos efeitos aleatórios de inclinação e de intercepto do nível *estudante*. O seguinte comando, cujos *outputs* encontram-se na Figura 16.42, faz com que sejam apresentadas as descrições destes termos aleatórios:

```
desc u10final u00final r1final r0final
```

```
. desc u10final u00final r1final r0final

              storage  display     value
variable name   type   format      label      variable label
------------------------------------------------------------------------------
--
u10final        float  %9.0g                  BLUP r.e. for escola: ano
u00final        float  %9.0g                  BLUP r.e. for escola: _cons
r1final         float  %9.0g                  BLUP r.e. for estudante: ano
r0final         float  %9.0g                  BLUP r.e. for estudante: _cons
```

Figura 16.42 Descrição dos termos aleatórios u_{10k}, u_{00k}, r_{1jk} e r_{0jk}.

Além disso, também podemos obter os valores esperados do desempenho escolar de cada estudante em cada um dos períodos monitorados, por meio da digitação do seguinte comando:

```
predict yhatestudante, fitted level(estudante)
```

que define a variável *yhatestudante*, que também pode ser obtida por meio do seguinte comando:

```
gen yhatestudante = 54.73435 + 4.515641*ano - 14.70213*sexo +
1.178656*texp + .6518855*sexoano - .0566496*texpano + u00final +
u10final*ano + r0final + r1final*ano
```

que corresponde à expressão:

$$\widehat{desempenho\_estudante}_{jk} = 54,734 + 4,516.ano_{jk} - 14,702.sexo_{jk} + 1,179.texp_k + 0,652.sexo_{jk}.ano_{jk} - 0,057.texp_k.ano_{jk} + u_{00k} + u_{10k}.ano_{jk} + r_{0jk} + r_{1jk}.ano_{jk}$$

Se o pesquisador digitar o seguinte comando, irá obter os valores esperados do desempenho escolar de cada estudante em cada um dos períodos monitorados, porém sem a consideração de efeitos aleatórios no nível *estudante*:

```
predict yhatescola, fitted level(escola)
```

que define a variável *yhatescola* no banco de dados, que também pode ser obtida por meio do seguinte comando:

```
gen yhatescola = 54.73435 + 4.515641*ano - 14.70213*sexo +
1.178656*texp + .6518855*sexoano - .0566496*texpano + u00final +
u10final*ano
```

que corresponde à expressão:

$$\begin{aligned} desempênho\_estudante_{jk} = {} & 54,734 + 4,516.ano_{jk} - 14,702.sexo_{jk} + 1,179.texp_k \\ & + 0,652.sexo_{jk}.ano_{jk} - 0,057.texp_k.ano_{jk} \\ & + u_{00k} + u_{10k}.ano_{jk} \end{aligned}$$

Os termos de erro e_{tjk} podem ser obtidos por meio da digitação do comando **`predict etjk, res`** (que equivale a *desempenho - yhatestudante*).

Neste momento, portanto, temos condições de finalizar a análise, verificando que, além do desempenho escolar dos estudantes seguir uma tendência linear ao longo do tempo, existindo variância significativa de interceptos e de inclinações entre aqueles que estudam na mesma escola e entre aqueles que estudam em escolas distintas, e o sexo dos estudantes ser significante para explicar parte dessa variação, o próprio tempo médio de experiência docente em cada escola (variável de nível 3) também explica parte das discrepâncias no desempenho escolar anual entre os estudantes provenientes de diferentes escolas.

O comando a seguir, digitado após o comando **`sort estudante ano`**, faz com que seja gerado um gráfico (Figura 16.43) com os valores previstos do desempenho escolar ao longo do tempo para os 50 primeiros estudantes da amostra (*yhatestudante*) e, por meio do qual, podemos visualizar distintos interceptos e inclinações ao longo do tempo para diferentes estudantes.

```
sort estudante ano

graph twoway connected yhatestudante ano if estudante <= 50, connect(L)
```

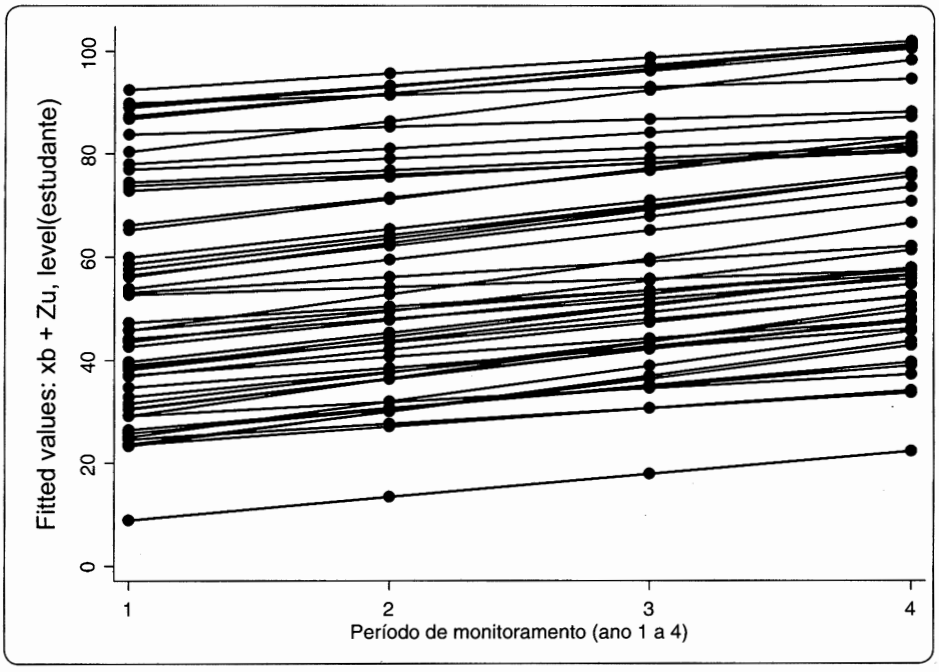


Figura 16.43 Valores previstos do desempenho escolar ao longo do tempo para os 50 primeiros estudantes da amostra.

Por fim, um pesquisador curioso, com o intuito de questionar a superioridade dos modelos multinível em relação aos modelos tradicionais de regressão estimados por MQO na existência de bases de dados com estruturas aninhadas, decide elaborar um gráfico em que é possível comparar os valores previstos do desempenho escolar gerados por esta modelagem hierárquica de três níveis (HLM3) com aqueles gerados por meio de

uma estimação por MQO, para todos os estudantes da amostra em cada um dos períodos analisados, usando as mesmas variáveis explicativas *ano*, *sexo*, *texp*, *sexoano* e *texpano* (obviamente, existem somente efeitos fixos na estimação por MQO).

Neste sentido, é digitada a seguinte sequência de comandos, que gera o gráfico da Figura 16.44:

```
quietly reg desempenho ano sexo texp sexoano texpano
predict yhatreg
graph twoway mspline yhatreg desempenho || mspline yhatestudante
desempenho || lfit desempenho desempenho ||, legend(label(1 "MQO")
label(2 "HLM3") label(3 "Valores Observados"))
```

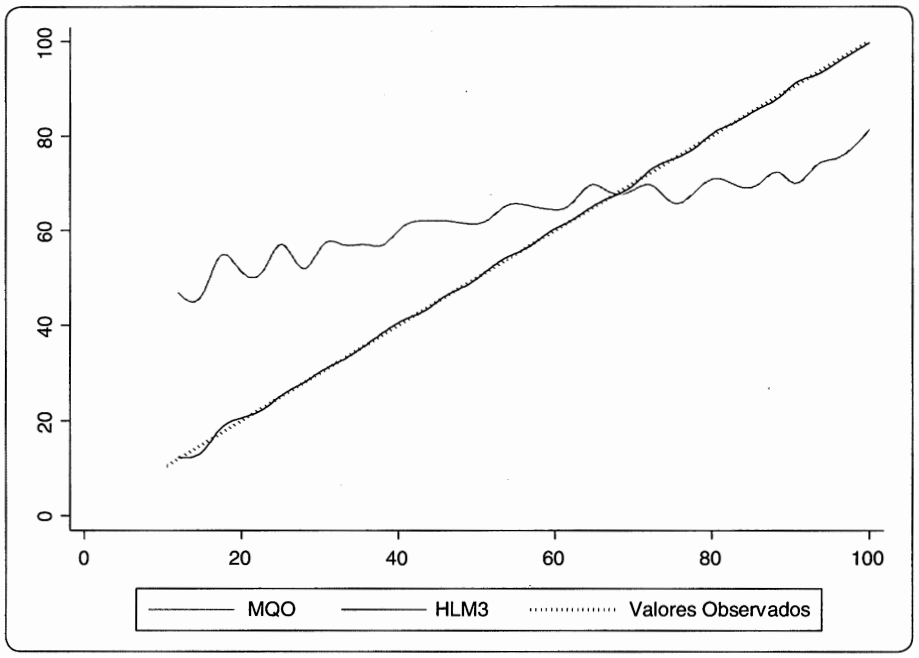


Figura 16.44 Valores previstos por MQO e por HLM3 x valores observados do desempenho escolar.

A reta pontilhada, a 45°, mostra os valores observados do desempenho escolar de cada um dos estudantes da amostra em cada um dos períodos analisados (*desempenho* x *desempenho*). Por meio do gráfico da Figura 16.44, podemos comprovar, nitidamente, a superioridade do nosso modelo de tendência linear com variáveis explicativas e com interceptos e inclinações aleatórias nos níveis 2 e 3 (modelo HLM3 completo) sobre o modelo de regressão linear múltipla estimado por MQO com as mesmas variáveis explicativas, o que demonstra a importância de se considerarem componentes de efeitos aleatórios na existência de estruturas aninhadas de dados.

O Quadro 16.1 apresenta, de forma consolidada, os comandos gerais, em Stata, para elaboração da modelagem hierárquica linear de dois níveis com dados agrupados e da modelagem hierárquica linear de três níveis com medidas repetidas, conforme estudado nas seções 16.4.1 e 16.4.2, respectivamente. O assunto é realmente vasto e novos modelos intermediários podem ser estimados sempre pelo pesquisador, em função de seus objetivos de pesquisa e dos constructos propostos.

Feitas essas considerações, e respeitada a *multilevel step-up strategy* ao longo de toda esta seção, vamos elaborar os mesmos exemplos por meio do software SPSS, a fim de propiciar ao pesquisador a oportunidade de comparação do manuseio dos softwares, dos procedimentos e rotinas para estimação dos modelos e das lógicas com que são apresentados os *outputs*.

QUADRO 16.1 Modelagens hierárquicas, modelos intermediários (*multilevel step-up strategy*) e comandos em Stata.

Modelagem	Modelo Intermediário	Comando em Stata
Hierárquica Linear de Dois Níveis com Dados Agrupados	Modelo Nulo (Modelo Não Condicional)	`xtmixed Y || var(nível 2):`
	Modelo com Interceptos Aleatórios	`xtmixed Y X || var(nível 2):`
	Modelo com Interceptos e Inclinações Aleatórias	`xtmixed Y X || var(nível 2): X`
	Modelo com Interceptos e Inclinações Aleatórias e Termos de Erro Correlacionados	`xtmixed Y X || var(nível 2): X covariance(unstructured)`
Hierárquica Linear de Três Níveis com Medidas Repetidas	Modelo Nulo (Modelo Não Condicional)	`xtmixed Y || var(nível 3): || var(nível 2):`
	Modelo de Tendência Linear com Interceptos Aleatórios	`xtmixed Y t || var(nível 3): || var(nível 2):`
	Modelo de Tendência Linear com Interceptos e Inclinações Aleatórias	`xtmixed Y t || var(nível 3): t || var(nível 2): t`
	Modelo de Tendência Linear com Interceptos e Inclinações Aleatórias e Variável de Nível 2	`xtmixed Y t X Xt || var(nível 3): t || var(nível 2): t`
	Modelo de Tendência Linear com Interceptos e Inclinações Aleatórias e Variáveis de Níveis 2 e 3	`xtmixed Y t X W Xt Wt WXt || var(nível 3): t || var(nível 2): t`
	Modelo de Tendência Linear com Interceptos e Inclinações Aleatórias e Variáveis de Níveis 2 e 3 e Termos de Erro Correlacionados	`xtmixed Y t X W Xt Wt WXt || var(nível 3): t, covariance(unstructured) || var(nível 2): t, covariance(unstructured)`

Nota: Considerada uma variável **X** de nível 2, uma variável **W** de nível 3 (quando houver) e **t** como variável temporal. Além disso, **Y** refere-se à variável dependente. Em todos os casos, foi omitido o termo correspondente ao método de estimação. Conforme discutido, enquanto o método de estimação padrão adotado pelo Stata até a versão 12 é o de máxima verossimilhança restrita (`reml`), o método padrão passa a ser o de máxima verossimilhança (`mle`) a partir da versão 13.

16.5. ESTIMAÇÃO DE MODELOS HIERÁRQUICOS LINEARES NO SOFTWARE SPSS

Apresentaremos agora o passo a passo para a elaboração dos nossos exemplos por meio do IBM SPSS Statistics Software®. A reprodução das imagens nesta seção tem autorização da International Business Machines Corporation©.

O maior objetivo, neste momento, é propiciar ao pesquisador uma oportunidade de elaborar as técnicas de modelagem multinível no SPSS. A cada apresentação de um *output*, faremos menção ao respectivo resultado obtido na elaboração das técnicas por meio do Stata, a fim de que o pesquisador possa compará-los e, dessa forma, decidir qual software utilizar, em função das características de cada um e da própria acessibilidade para uso.

16.5.1. Estimação de um modelo hierárquico linear de dois níveis com dados agrupados no software SPSS

Voltando ao exemplo utilizado na seção 16.4.1, lembremos que o nosso professor levantou dados sobre o desempenho escolar (nota de 0 a 100 mais um bônus por participação em sala) de 2.000 estudantes provenientes de 46 escolas, bem como dados sobre a quantidade semanal de horas de estudo (variável explicativa de nível 1) e sobre a natureza das escolas (pública ou privada) e o tempo médio de experiência docente dos professores em cada uma delas (variáveis explicativas de nível 2). A base de dados completa está no arquivo **DesempenhoAlunoEscola.sav**.

Mantendo a lógica apresentada, vamos, inicialmente, estimar o modelo nulo, conforme segue:

Modelo Nulo:

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

Embora seja possível elaborar modelagens multinível fazendo uso do menu **Analyze → Mixed Models** do SPSS, com base em procedimentos *point-and-click*, optamos, nesta seção, por estimar os modelos por meio de sintaxes, a fim de propiciar uma melhor comparação com as estimações elaboradas na seção 16.4.1 e facilitar a compreensão sobre a lógica de inclusão das variáveis nos componentes de efeitos fixos e aleatórios. Para tanto, com o arquivo **DesempenhoAlunoEscola.sav** aberto, devemos clicar em **File → New → Syntax**. Para o modelo nulo, devemos digitar a seguinte sintaxe na janela que será aberta:

```
MIXED desempenho
/METHOD = REML
/PRINT = SOLUTION TESTCOV
/FIXED = INTERCEPT
/RANDOM = INTERCEPT | SUBJECT(escola) .
```

em que a primeira linha (**MIXED**)[4] apresenta apenas a variável dependente *desempenho* e as duas linhas seguintes (**METHOD** e **PRINT**) determinam, respectivamente, o método de estimação adotado (no caso, máxima verossimilhança restrita, ou *REML*) e que sejam apresentados, nos *outputs*, as estimações de efeitos fixos com correspondentes erros-padrão. Por fim, nas duas últimas linhas (**FIXED** e **RANDOM**) podem ser especificadas, além do termo de intercepto, as variáveis que farão parte dos componentes de efeitos fixos e aleatórios, respectivamente, em que o termo **SUBJECT** inserido após a barra vertical | identifica a variável de grupo correspondente ao nível 2 (no nosso caso, a variável *escola*).

A Figura 16.45 apresenta a janela do SPSS com a inclusão da sintaxe correspondente ao modelo nulo, com destaque para o botão **Run Selection** que deverá ser clicado a fim de que a modelagem multinível seja elaborada.

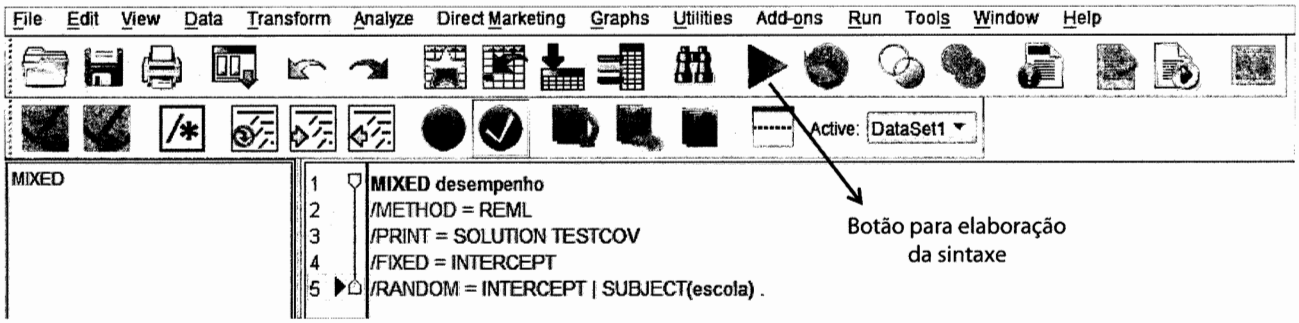


Figura 16.45 Janela com inclusão da sintaxe para estimação do modelo nulo no SPSS.

A seguir, na Figura 16.46, são apresentados os *outputs* gerados pelo SPSS.

Inicialmente, podemos verificar que o *output* **Model Dimension** apresenta a quantidade de níveis considerados na modelagem (no caso, 2) e a quantidade de parâmetros estimados (no caso, 3, incluindo o termo de erro). O termo **Variance Components** informa que está sendo considerada uma estrutura da matriz de variância-covariância com termos de erro aleatórios independentes.

Em **Information Criteria**, é apresentado o valor de **-2 Restricted Log Likelihood**, que corresponde a -2 vezes o valor máximo obtido do logaritmo da função de verossimilhança restrita para a estimação dos parâmetros do modelo. Podemos verificar que o *output* do SPSS mostra que $-2.LL_r$ = 17.504,04, que é exatamente igual a -2 vezes o valor apresentado pelo Stata (Figura 16.8), já que -2.(-8.752,02) = 17.504,04.

[4] O comando **MIXED** passou a estar disponível no SPSS a partir de 2001, na versão 11.0.

Model Dimension[a]

		Number of Levels	Covariance Structure	Number of Parameters	Subject Variables
Fixed Effects	Intercept	1		1	
Random Effects	Intercept	1	Variance Components	1	escola
Residual				1	
Total		2		3	

a. Dependent Variable: desempenho escolar (nota de 0 a 100).

Information Criteria[a]

-2 Restricted Log Likelihood	17504,041
Akaike's Information Criterion (AIC)	17508,041
Hurvich and Tsai's Criterion (AICC)	17508,047
Bozdogan's Criterion (CAIC)	17521,242
Schwarz's Bayesian Criterion (BIC)	17519,242

The information criteria are displayed in smaller-is-better forms.

a. Dependent Variable: desempenho escolar (nota de 0 a 100).

Fixed Effects

Type III Tests of Fixed Effects[a]

Source	Numerator df	Denominator df	F	Sig.
Intercept	1	44,388	1181,424	,000

a. Dependent Variable: desempenho escolar (nota de 0 a 100).

Estimates of Fixed Effects[a]

Parameter	Estimate	Std. Error	df	t	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Intercept	61,049010	1,776134	44,388	34,372	,000	57,470330	64,627689

a. Dependent Variable: desempenho escolar (nota de 0 a 100).

Covariance Parameters

Estimates of Covariance Parameters[a]

Parameter		Estimate	Std. Error	Wald Z	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Residual		347,561691	11,120778	31,253	,000	326,434748	370,055975
Intercept [subject = escola]	Variance	135,779174	30,750059	4,416	,000	87,108516	211,643878

a. Dependent Variable: desempenho escolar (nota de 0 a 100).

Figura 16.46 *Outputs* do modelo nulo no SPSS.

Na sequência, em **Fixed Effects**, é apresentada a estimação do parâmetro γ_{00} (efeito fixo), que corresponde à média dos desempenhos escolares esperados dos estudantes (reta horizontal estimada no modelo nulo, ou intercepto geral). Podemos verificar que a estimação de $\gamma_{00} = 61{,}049$ corresponde àquela obtida na Figura 16.8 na elaboração do modelo nulo no Stata.

Por fim, são apresentadas as estimações dos componentes de variância dos termos de erro (efeitos aleatórios) dos níveis 1 e 2 (**Covariance Parameters**). Podemos aqui também verificar que os *outputs* correspondem aos obtidos pelo Stata, já que as estimações de $\tau_{00} = 135{,}779$ (**Intercept [subject=escola]**) e $\sigma^2 = 347{,}562$ (**Residual**). Note, entretanto, que o SPSS apresenta de maneira direta, ao contrário do Stata, as estatísticas z das estimações das variâncias dos termos de erro, com respectivos níveis de significância. Assim, para os dados do nosso exemplo, podemos comprovar que existe variabilidade no desempenho escolar dos estudantes provenientes de escolas diferentes, visto que *Sig.* z $\tau_{00} < 0{,}05$ (definido o nível de confiança de 95%).

Com base na correlação intraclasse, calculada a seguir, podemos verificar que aproximadamente 28% da variância total do desempenho escolar é devido à alteração entre escolas.

$$rho = \frac{\tau_{00}}{\tau_{00} + \sigma^2} = \frac{135,779}{135,779 + 347,562} = 0,281$$

A fim de mantermos a lógica apresentada na seção 16.4.1, vamos agora estimar o modelo com interceptos aleatórios, incluindo a variável *horas* como explicativa, conforme segue:

Modelo com Interceptos Aleatórios:

$$desempenho_{ij} = \gamma_{00} + \gamma_{10}.horas_{ij} + u_{0j} + r_{ij}$$

A sintaxe para a estimação desse modelo no SPSS é:

```
MIXED desempenho WITH horas
/METHOD = REML
/PRINT = SOLUTION TESTCOV
/FIXED = INTERCEPT horas
/RANDOM = INTERCEPT | SUBJECT(escola) .
```

em que devem ser inseridas todas as variáveis explicativas que o pesquisador desejar após o termo **WITH** na primeira linha da sintaxe. A partir de sua execução, chegamos aos principais *outputs* apresentados na Figura 16.47.

Information Criteria[a]

-2 Restricted Log Likelihood	12744,329
Akaike's Information Criterion (AIC)	12748,329
Hurvich and Tsai's Criterion (AICC)	12748,335
Bozdogan's Criterion (CAIC)	12761,528
Schwarz's Bayesian Criterion (BIC)	12759,528

The information criteria are displayed in smaller-is-better forms.

a. Dependent Variable: desempenho escolar (nota de 0 a 100).

Fixed Effects

Estimates of Fixed Effects[a]

Parameter	Estimate	Std. Error	df	t	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Intercept	,534468	,787530	91,043	,679	,499	-1,029855	2,098790
horas	3,251924	,023163	1984,423	140,390	,000	3,206497	3,297352

a. Dependent Variable: desempenho escolar (nota de 0 a 100).

Covariance Parameters

Estimates of Covariance Parameters[a]

Parameter		Estimate	Std. Error	Wald Z	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Residual		31,763781	1,016389	31,252	,000	29,832877	33,819661
Intercept [subject = escola]	Variance	19,125335	4,199478	4,554	,000	12,436696	29,411223

a. Dependent Variable: desempenho escolar (nota de 0 a 100).

Figura 16.47 Principais *outputs* do modelo com interceptos aleatórios.

Esses *outputs* correspondem aos apresentados na Figura 16.10 (Stata) e, por meio dos mesmos, podemos verificar que existe significância estatística das estimações das variâncias dos termos de erro τ_{00} = 19,125 e σ^2 = 31,764, que resultam na seguinte correlação intraclasse:

$$rho = \frac{\tau_{00}}{\tau_{00} + \sigma^2} = \frac{19{,}125}{19{,}125 + 31{,}764} = 0{,}376$$

Neste sentido, há um incremento da proporção do componente de variância correspondente ao intercepto em relação ao modelo nulo, o que favorece a decisão de inclusão da variável *horas* para o estudo do comportamento do desempenho escolar na comparação entre escolas.

O nosso modelo, portanto, passa a ter, no presente momento, a seguinte especificação:

$$desempenho_{ij} = 0{,}534 + 3{,}252.horas_{ij} + u_{0j} + r_{ij}$$

em que o efeito fixo do intercepto corresponde à média esperada dos desempenhos escolares, entre escolas, dos alunos que, por alguma razão, não estudam ($horas_{ij} = 0$) e a inclinação permite que afirmemos que uma hora a mais de estudo semanal, em média, faz com que a média esperada dos desempenhos escolares, entre escolas, seja incrementada em 3,252 pontos, sendo este parâmetro estatisticamente significante[5].

Neste momento, vamos inserir efeitos aleatórios de inclinação no nosso modelo multinível que, com a manutenção dos efeitos aleatórios de intercepto, passará a ter a seguinte expressão:

Modelo com Interceptos e Inclinações Aleatórias:

$$desempenho_{ij} = \gamma_{00} + \gamma_{10}.horas_{ij} + u_{0j} + u_{1j}.horas_{ij} + r_{ij}$$

A nova sintaxe é:

```
MIXED desempenho WITH horas
/METHOD = REML
/PRINT = SOLUTION TESTCOV
/FIXED = INTERCEPT horas
/RANDOM = INTERCEPT horas | SUBJECT(escola) .
```

que gera os *outputs* apresentados na Figura 16.48.

Analogamente, esses *outputs* correspondem àqueles apresentados na Figura 16.13 (Stata).

Podemos verificar que as estimações dos parâmetros e das variâncias no modelo com interceptos e inclinações aleatórias são idênticas às obtidas na estimação dos parâmetros do modelo apenas com interceptos aleatórios (Figura 16.47). Isso ocorre pelo fato de a estimação da variância τ_{11} (**horas [subject=escola]**) ser estatisticamente igual a zero, o que faz com que o valor obtido de $-2.LL_r$ também seja o mesmo daquele apresentado na Figura 16.47.

---

[5] Se o pesquisador desejar elaborar um teste de razão de verossimilhança para comparar as estimações dos modelos nulo e com interceptos aleatórios, cujas especificações dos componentes fixos são obviamente diferentes, deverá fazê-lo estimando estes dois modelos por máxima verossimilhança (*ML*), em vez de por máxima verossimilhança restrita (*REML*). Assim, deverá digitar as duas sintaxes a seguir, correspondentes, respectivamente, às estimações por máxima verossimilhança (no SPSS, **METHOD = ML**) do modelo nulo e do modelo com interceptos aleatórios:

```
MIXED desempenho
/METHOD = ML
/PRINT = SOLUTION TESTCOV
/FIXED = INTERCEPT
/RANDOM = INTERCEPT | SUBJECT(escola) .
```

```
MIXED desempenho WITH horas
/METHOD = ML
/PRINT = SOLUTION TESTCOV
/FIXED = INTERCEPT horas
/RANDOM = INTERCEPT | SUBJECT(escola) .
```

que, embora não apresentados aqui, geram valores de $-2.LL$ iguais, respetivamente, a 17.507,017 e 12.739,629. Portanto, o teste de razão de verossimilhança apresenta nível de significância Sig. χ^2_1 (17.507,017 - 12.739,629 = 4.767,39) = 0,000 < 0,05, o que favorece a adoção do modelo com efeitos aleatórios no intercepto.

Information Criteria[a]

-2 Restricted Log Likelihood	12744,329
Akaike's Information Criterion (AIC)	12750,329
Hurvich and Tsai's Criterion (AICC)	12750,341
Bozdogan's Criterion (CAIC)	12770,128
Schwarz's Bayesian Criterion (BIC)	12767,128

The information criteria are displayed in smaller-is-better forms.

a. Dependent Variable: desempenho escolar (nota de 0 a 100).

Fixed Effects

Estimates of Fixed Effects[a]

Parameter	Estimate	Std. Error	df	t	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Intercept	,534468	,787530	91,043	,679	,499	-1,029855	2,098790
horas	3,251924	,023163	1984,423	140,390	,000	3,206497	3,297352

a. Dependent Variable: desempenho escolar (nota de 0 a 100).

Covariance Parameters

Estimates of Covariance Parameters[b]

Parameter		Estimate	Std. Error	Wald Z	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Residual		31,763781	1,016389	31,252	,000	29,832877	33,819661
Intercept [subject = escola]	Variance	19,125335	4,199478	4,554	,000	12,436696	29,411223
horas [subject = escola]	Variance	,000000[a]	,000000	.	.	.	.

a. This covariance parameter is redundant. The test statistic and confidence interval cannot be computed.
b. Dependent Variable: desempenho escolar (nota de 0 a 100).

Figura 16.48 Principais *outputs* do modelo com interceptos e inclinações aleatórias.

Portanto, a aplicação de um teste de razão de verossimilhança ofereceria um resultado que obviamente indicaria o favorecimento da adoção do modelo apenas com interceptos aleatórios, já que o nível de significância *Sig.* χ^2_1 (12.744,329 - 12.744,329 = 0) = 1,000 > 0,05, conforme já mostrava a Figura 16.14.

Se o pesquisador desejar generalizar a estrutura da matriz de variância-covariância dos termos de erro aleatórios, permitindo que u_{0j} e u_{1j} sejam correlacionados, basta que estime os parâmetros do modelo usando o termo **COVTYPE(UN)** ao final da linha **RANDOM** da última sintaxe, que passará a ser:

```
MIXED desempenho WITH horas
/METHOD = REML
/PRINT = SOLUTION TESTCOV
/FIXED = INTERCEPT horas
/RANDOM = INTERCEPT horas | SUBJECT(escola) COVTYPE(UN) .
```

em que o termo **COVTYPE(UN)** considera a existência de uma matriz de variância-covariância *unstructured*. Os *outputs* deste modelo não estão apresentados aqui, porém um teste de razão de verossimilhança para comparar as estimações dos modelos com interceptos e inclinações aleatórias com termos de erro u_{0j} e u_{1j} independentes e correlacionados mostrará que a estrutura da matriz de variância-covariância entre u_{0j} e u_{1j} pode ser considerada independente, de forma análoga ao apresentado na Figura 16.18.

Sendo independente a estrutura matriz de variância-covariância dos erros aleatórios e sendo mais adequado o modelo apenas com interceptos aleatórios, vamos partir para a estimação do modelo final completo, que possui a seguinte especificação:

Modelo Final Completo:

$$desempenho_{ij} = \gamma_{00} + \gamma_{10}.horas_{ij} + \gamma_{01}.texp_j + \gamma_{02}.priv_j + \gamma_{11}.priv_j.horas_{ij} + u_{0j} + r_{ij}$$

Note que já partimos para a última estimação obtida na seção 16.4.1. A sintaxe para a elaboração da modelagem é:

```
MIXED desempenho WITH horas texp priv
/METHOD = REML
/PRINT = SOLUTION TESTCOV
/FIXED = INTERCEPT horas texp priv priv*horas
/RANDOM = INTERCEPT | SUBJECT(escola)
/SAVE = PRED FIXPRED .
```

em que a última linha agora apresenta o termo **SAVE = PRED FIXPRED**, que faz com que sejam geradas duas novas variáveis no banco de dados, *PRED_1* e *FXPRED_1*. Enquanto a primeira corresponde aos valores previstos do desempenho escolar por estudante (*yhat* no Stata), inclusive com componentes aleatórios u_{0j} de intercepto, a segunda refere-se aos valores previstos do desempenho escolar decorrentes apenas do componente de efeitos fixos. Os *outputs* gerados são apresentados na Figura 16.49, e os valores esperados *BLUPS* (*best linear unbiased predictions*) dos efeitos aleatórios u_{0j} do nosso modelo final podem, portanto, ser obtidos por meio da seguinte sintaxe:

```
COMPUTE blups=PRED_1-FXPRED_1.
```

que gera no banco de dados uma nova variável, denominada *blups*, igual à variável *u0final* definida na estimação deste modelo em Stata.

Information Criteria[a]

-2 Restricted Log Likelihood	12719,507
Akaike's Information Criterion (AIC)	12723,507
Hurvich and Tsai's Criterion (AICC)	12723,513
Bozdogan's Criterion (CAIC)	12736,704
Schwarz's Bayesian Criterion (BIC)	12734,704

The information criteria are displayed in smaller-is-better forms.

a. Dependent Variable: desempenho escolar (nota de 0 a 100).

Fixed Effects

Estimates of Fixed Effects[a]

Parameter	Estimate	Std. Error	df	t	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Intercept	-2,710350	,893160	94,435	-3,035	,003	-4,483634	-,937066
horas	3,281046	,029276	1988,758	112,074	,000	3,223631	3,338460
texp	,866203	,164196	42,244	5,275	,000	,534898	1,197508
priv	-5,610535	2,288084	58,462	-2,452	,017	-10,189862	-1,031209
horas * priv	-,080121	,047722	1986,701	-1,679	,093	-,173711	,013469

a. Dependent Variable: desempenho escolar (nota de 0 a 100).

Covariance Parameters

Estimates of Covariance Parameters[a]

Parameter		Estimate	Std. Error	Wald Z	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Residual		31,720600	1,015254	31,244	,000	29,791867	33,774199
Intercept [subject = escola]	Variance	11,057762	2,559522	4,320	,000	7,024914	17,405779

a. Dependent Variable: desempenho escolar (nota de 0 a 100).

Figura 16.49 Principais *outputs* do modelo final completo com interceptos aleatórios.

Esses resultados correspondem aos apresentados na Figura 16.20 (Stata). Com estimações significantes das variâncias dos termos de erro aleatórios e dos parâmetros de efeitos fixos, ao nível de confiança de 95% (exceção feita à estimação do parâmetro da variável combinada *horas*priv*, significante ao nível de confiança de 90%), chegamos à seguinte expressão do modelo proposto:

$$desempenho_{ij} = -2,710 + 3,281.horas_{ij} + 0,866.texp_j - 5,610.priv_j - 0,080.priv_j.horas_{ij} + u_{0j} + r_{ij}$$

construído com a inclusão de variáveis explicativas dos níveis 1 e 2 e por meio da *multilevel step-up strategy*. Podemos concluir, portanto, que existem diferenças no comportamento do desempenho escolar entre estudantes provenientes de mesmas escolas e de escolas distintas, e essas diferenças ocorrem, respectivamente, em função da quantidade semanal de horas de estudo de cada estudante, da natureza (pública ou privada) e do tempo médio de experiência dos professores de cada escola.

Na sequência elaboraremos, também em SPSS, um exemplo de modelo hierárquico linear de três níveis com medidas repetidas.

16.5.2. Estimação de um modelo hierárquico linear de três níveis com medidas repetidas no software SPSS

Nesta seção, vamos retomar o exemplo utilizado na seção 16.4.2, lembrando que o nosso professor conseguiu dados sobre o desempenho escolar (nota de 0 a 100) ao longo de quatro anos (variável temporal de nível 1) de 2.000 estudantes provenientes de 15 escolas, bem como dados sobre o sexo de cada estudante (variável explicativa de nível 2) e sobre o tempo médio de experiência docente em cada uma das escolas (variável explicativa de nível 3). A base de dados completa é apresentada no arquivo **DesempenhoTempoAlunoEscola.sav**.

É importante mencionar que o SPSS apresenta tempos de processamento de estimações de modelos multinível, principalmente para uma quantidade de níveis igual ou superior a três, consideravelmente maior do que o Stata.

Mantendo a lógica apresentada na seção 16.4.2, vamos inicialmente estimar o modelo nulo, conforme segue:

Modelo Nulo:

$$desempenho_{tjk} = \gamma_{000} + u_{00k} + r_{0jk} + e_{tjk}$$

Para esse modelo nulo, devemos digitar a seguinte rotina na janela de sintaxes:

```
MIXED desempenho
/METHOD = REML
/PRINT = SOLUTION TESTCOV
/FIXED = INTERCEPT
/RANDOM = INTERCEPT | SUBJECT(estudante)
/RANDOM = INTERCEPT | SUBJECT(escola) .
```

em que a primeira linha (**MIXED**) apresenta apenas a variável dependente *desempenho* e as duas linhas seguintes (**METHOD** e **PRINT**) determinam, respectivamente, o método de estimação adotado (no caso, máxima verossimilhança restrita, ou *REML*) e que sejam apresentados, nos *outputs*, as estimações de efeitos fixos com correspondentes erros-padrão. Na linha seguinte (**FIXED**) pode ser especificada a variável que fará parte dos componentes de efeitos fixos, além do termo de intercepto. Por fim, nas duas últimas linhas da rotina (**RANDOM**) podem ser especificadas, além dos termos de intercepto, as variáveis que farão parte dos componentes de efeitos aleatórios nos diferentes níveis da análise, em que o termo **SUBJECT** inserido após a barra vertical | identifica a variável de grupo correspondente a cada nível (no nosso caso, *estudante* para o nível 2 e *escola* para o nível 3).

A Figura 16.50 apresenta os *outputs* gerados pelo SPSS.

Model Dimension[a]

		Number of Levels	Covariance Structure	Number of Parameters	Subject Variables
Fixed Effects	Intercept	1		1	
Random Effects	Intercept	1	Variance Components	1	estudante
	Intercept	1	Variance Components	1	escola
Residual				1	
Total		3		4	

a. Dependent Variable: desempenho escolar.

Information Criteria[a]

-2 Restricted Log Likelihood	18184,277
Akaike's Information Criterion (AIC)	18190,277
Hurvich and Tsai's Criterion (AICC)	18190,287
Bozdogan's Criterion (CAIC)	18210,675
Schwarz's Bayesian Criterion (BIC)	18207,675

The information criteria are displayed in smaller-is-better forms.

a. Dependent Variable: desempenho escolar.

Fixed Effects

Type III Tests of Fixed Effects[a]

Source	Numerator df	Denominator df	F	Sig.
Intercept	1	13,982	373,992	,000

a. Dependent Variable: desempenho escolar.

Estimates of Fixed Effects[a]

						95% Confidence Interval	
Parameter	Estimate	Std. Error	df	t	Sig.	Lower Bound	Upper Bound
Intercept	68,713953	3,553153	13,982	19,339	,000	61,092286	76,335620

a. Dependent Variable: desempenho escolar.

Covariance Parameters

Estimates of Covariance Parameters[a]

						95% Confidence Interval	
Parameter		Estimate	Std. Error	Wald Z	Sig.	Lower Bound	Upper Bound
Residual		41,649389	1,376886	30,249	,000	39,036312	44,437385
Intercept [subject = estudante]	Variance	325,799148	19,495760	16,711	,000	289,743835	366,341134
Intercept [subject = escola]	Variance	180,192658	71,603650	2,517	,012	82,697580	392,628101

a. Dependent Variable: desempenho escolar.

Figura 16.50 *Outputs* do modelo nulo no SPSS.

Não vamos analisar novamente todos os *outputs* do modelo gerado, visto que são exatamente iguais aos apresentados na Figura 16.30, obtida na estimação deste modelo nulo em Stata.

Entretanto, podemos verificar que a estimação do parâmetro γ_{000} (**Fixed Effects**) é igual a 68,714, que corresponde à média dos desempenhos escolares anuais esperados dos estudantes (reta horizontal estimada no modelo nulo, ou intercepto geral).

Além disso, temos que as estimações das variâncias dos termos de erro (**Covariance Parameters**) τ_{u000} = 180,194 (**Intercept [subject=escola]**), τ_{r000} = 325,799 (**Intercept [subject=estudante]**) e σ^2 = 41,649 (**Residual**) são estatisticamente diferentes de zero, ao nível de significância de 5%. Esse fato permite que afirmemos que há variabilidade significativa no desempenho escolar ao longo dos quatro anos da análise, que há variabilidade significativa no desempenho escolar, ao longo do tempo, entre estudantes de uma mesma escola, e que há variabilidade significativa no desempenho escolar, ao longo do tempo, entre estudantes provenientes de escolas distintas.

As duas correlações intraclasse, correspondentes aos níveis 2 e 3 da análise, podem ser calculadas conforme segue:

- **Correlação intraclasse de nível 2:**

$$rho_{estudante|escola} = corr\left(Y_{tjk}, Y_{t'jk}\right) = \frac{\tau_{u000} + \tau_{r000}}{\tau_{u000} + \tau_{r000} + \sigma^2} = \frac{180{,}194 + 325{,}799}{180{,}194 + 325{,}799 + 41{,}649} = 0{,}924$$

- **Correlação intraclasse de nível 3:**

$$rho_{escola} = corr\left(Y_{tjk}, Y_{t'j'k}\right) = \frac{\tau_{u000}}{\tau_{u000} + \tau_{r000} + \sigma^2} = \frac{180{,}194}{180{,}194 + 325{,}799 + 41{,}649} = 0{,}329$$

Logo, a correlação entre os desempenhos escolares anuais, para uma mesma escola, é igual a 32,9% (rho_{escola}) e a correlação entre os desempenhos escolares anuais, para um mesmo estudante de determinada escola, é igual a 92,4% ($rho_{estudante|escola}$).

A fim de mantermos a lógica apresentada na seção 16.4.2, vamos partir agora para a estimação do modelo de tendência linear com interceptos e inclinações aleatórias, incluindo a variável *ano* (medida repetida) como explicativa no nível 1, conforme segue:

Modelo de Tendência Linear com Interceptos e Inclinações Aleatórias:

$$desempenho_{tjk} = \gamma_{000} + \gamma_{100}.ano_{jk} + u_{00k} + u_{10k}.ano_{jk} + r_{0jk} + r_{1jk}.ano_{jk} + e_{tjk}$$

A sintaxe para a estimação deste modelo no SPSS é:

```
MIXED desempenho WITH ano
/METHOD = REML
/PRINT = SOLUTION TESTCOV
/FIXED = INTERCEPT ano
/RANDOM = INTERCEPT ano | SUBJECT(estudante)
/RANDOM = INTERCEPT ano | SUBJECT(escola) .
```

em que devem ser inseridas todas as variáveis explicativas que o pesquisador desejar após o termo **WITH** na primeira linha da sintaxe. Após nove iterações e alguns minutos de processamento do software, chegamos aos principais *outputs* apresentados na Figura 16.51.

Information Criteria[a]

-2 Restricted Log Likelihood	14929,638
Akaike's Information Criterion (AIC)	14939,638
Hurvich and Tsai's Criterion (AICC)	14939,663
Bozdogan's Criterion (CAIC)	14973,633
Schwarz's Bayesian Criterion (BIC)	14968,633

The information criteria are displayed in smaller-is-better forms.

a. Dependent Variable: desempenho escolar.

Fixed Effects

Estimates of Fixed Effects[a]

Parameter	Estimate	Std. Error	df	t	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Intercept	57,857761	3,955770	13,993	14,626	,000	49,373090	66,342433
ano	4,343297	,210705	13,903	20,613	,000	3,891085	4,795509

a. Dependent Variable: desempenho escolar.

Covariance Parameters

Estimates of Covariance Parameters[a]

Parameter		Estimate	Std. Error	Wald Z	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Residual		3,867728	,159525	24,245	,000	3,567368	4,193377
Intercept [subject = estudante]	Variance	374,284569	22,009042	17,006	,000	333,540633	420,005615
ano [subject = estudante]	Variance	3,157274	,230544	13,695	,000	2,736260	3,643066
Intercept [subject = escola]	Variance	224,337985	88,719082	2,529	,011	103,342179	486,998939
ano [subject = escola]	Variance	,560036	,251904	2,223	,026	,231924	1,352342

a. Dependent Variable: desempenho escolar.

Figura 16.51 Principais *outputs* do modelo de tendência linear com interceptos e inclinações aleatórias.

Esses *outputs* correspondem àqueles apresentados na Figura 16.35 e, por meio dos quais, podemos verificar que os parâmetros estimados dos componentes de efeitos fixos e aleatórios são estatisticamente diferentes de zero, ao nível de significância de 5%, o que nos dá subsídios à afirmação de que o desempenho escolar dos estudantes segue uma tendência linear ao longo do tempo, existindo variância significativa de interceptos e de inclinações entre aqueles que estudam na mesma escola e entre aqueles que estudam em escolas distintas[6]. Por meio da correlação intraclasse de nível 2, calculada a seguir, estimamos que os efeitos aleatórios de estudantes e escolas compõem aproximadamente 99% da variância total dos resíduos!

---

[6] Se o pesquisador desejar comparar os resultados dessa estimação com aqueles provenientes da estimação de um modelo de tendência linear apenas com interceptos aleatórios, assim como realizado em Stata, basta que digite a seguinte rotina na janela de sintaxes do SPSS:

```
MIXED desempenho WITH ano
/METHOD = REML
/PRINT = SOLUTION TESTCOV
/FIXED = INTERCEPT ano
/RANDOM = INTERCEPT | SUBJECT(estudante)
/RANDOM = INTERCEPT | SUBJECT(escola) .
```

Os resultados, embora não apresentados aqui, geram valor de -2*LL* igual a 15.602,840. Portanto, um teste de razão de verossimilhança apresentará nível de significância Sig. χ^2_2 (15.602,840 - 14.929,638 = 673,20) = 0,000 < 0,05, o que favorece a adoção do modelo de tendência linear com interceptos e inclinações aleatórias.

$$rho_{estudante|escola} = corr\left(Y_{tjk}, Y_{t'jk}\right) = \frac{\tau_{u000} + \tau_{u100} + \tau_{r000} + \tau_{r100}}{\tau_{u000} + \tau_{u100} + \tau_{r000} + \tau_{r100} + \sigma^2}$$

$$= \frac{224,343 + 0,560 + 374,285 + 3,157}{224,343 + 0,560 + 374,285 + 3,157 + 3,868} = 0,994$$

Neste momento, o nosso modelo passa a ter a seguinte especificação:

$$desempenho_{tjk} = 57,858 + 4,343.ano_{jk} + u_{00k} + u_{10k}.ano_{jk} + r_{0jk} + r_{1jk}.ano_{jk} + e_{tjk}$$

Por fim, investigaremos se as variáveis *sexo* e *texp*, de níveis 2 e 3, respectivamente, também explicam a variação no desempenho escolar anual entre os estudantes. Após algumas análises intermediárias, partiremos para a estimação do seguinte modelo completo de três níveis:

Modelo de Tendência Linear com Interceptos e Inclinações Aleatórias e as Variáveis *sexo* de Nível 2 e *texp* de Nível 3 (Modelo Completo):

$$\begin{aligned} desempenho_{tjk} = {} & \gamma_{000} + \gamma_{100}.ano_{jk} + \gamma_{010}.sexo_{jk} + \gamma_{001}.texp_k \\ & + \gamma_{110}.sexo_{jk}.ano_{jk} + \gamma_{101}.texp_k.ano_{jk} \\ & + u_{00k} + u_{10k}.ano_{jk} + r_{0jk} + r_{1jk}.ano_{jk} + e_{tjk} \end{aligned}$$

Para a estimação deste modelo, vamos partir para a generalização da estrutura das matrizes de variância-covariância dos termos aleatórios, permitindo que (u_{00k}, u_{10k}) e (r_{0jk}, r_{1jk}) sejam correlacionados (matrizes de variância-covariância *unstructured*). Para tanto, devemos inserir a expressão **COVTYPE(UN)** ao final das linhas **RANDOM**, fazendo com que a sintaxe do SPSS seja:

```
MIXED desempenho WITH ano sexo texp
/METHOD = REML
/PRINT = SOLUTION TESTCOV
/FIXED = INTERCEPT ano sexo texp sexo*ano texp*ano
/RANDOM = INTERCEPT ano | SUBJECT(estudante) COVTYPE(UN)
/RANDOM = INTERCEPT ano | SUBJECT(escola) COVTYPE(UN)
/SAVE = PRED FIXPRED RESID .
```

em que a última linha apresenta agora o termo **SAVE = PRED FIXPRED RESID**, que faz com que sejam geradas três novas variáveis no banco de dados, *PRED_1*, *FXPRED_1* e *RESID_1*, que correspondem, respectivamente, aos valores previstos do desempenho escolar por estudante (*yhatestudante* no Stata), aos valores previstos do desempenho escolar decorrentes apenas do componente de efeitos fixos e aos termos de erro e_{tjk}.

Após cinco iterações e alguns minutos de processamento do software, chegamos aos *outputs* apresentados na Figura 16.52.

Estes *outputs* correspondem àqueles apresentados na Figura 16.39 (Stata) e, por meio dos quais, verificamos que todos os parâmetros estimados para o componente de efeitos fixos são estatisticamente diferentes de zero, ao nível de significância de 5%. Já em relação aos parâmetros dos componentes de efeitos aleatórios, apenas as estimações de u_{10k} e de cov(u_{00k}, u_{10k}) são estatisticamente significantes ao nível de significância de 10%, sendo todas as demais significantes ao nível de significância de 5%. Neste sentido, considerando que cov(u_{00k}, u_{10k}) e cov(r_{0jk}, r_{1jk}) sejam estatisticamente diferentes de zero, podemos escrever que:

- **Matriz de variância-covariância dos efeitos aleatórios para o nível *escola*:**

$$var\begin{bmatrix} u_{00k} \\ u_{10k} \end{bmatrix} = \begin{bmatrix} 88,734 & -3,185 \\ -3,185 & 0,255 \end{bmatrix}$$

- **Matriz de variância-covariância dos efeitos aleatórios para o nível *estudante*:**

$$var\begin{bmatrix} r_{0jk} \\ r_{1jk} \end{bmatrix} = \begin{bmatrix} 350,913 & -13,251 \\ -13,251 & 3,257 \end{bmatrix}$$

Information Criteria[a]

-2 Restricted Log Likelihood	14753,429
Akaike's Information Criterion (AIC)	14767,429
Hurvich and Tsai's Criterion (AICC)	14767,476
Bozdogan's Criterion (CAIC)	14815,010
Schwarz's Bayesian Criterion (BIC)	14808,010

The information criteria are displayed in smaller-is-better forms.

a. Dependent Variable: desempenho escolar.

Fixed Effects

Estimates of Fixed Effects[a]

Parameter	Estimate	Std. Error	df	t	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Intercept	54,734351	3,951390	15,516	13,852	,000	46,336504	63,132198
ano	4,515640	,258373	21,461	17,477	,000	3,979027	5,052254
sexo	-14,702129	1,795535	606,763	-8,188	,000	-18,228348	-11,175911
texp	1,178656	,345902	13,131	3,407	,005	,432135	1,925177
ano * sexo	,651886	,184716	514,048	3,529	,000	,288994	1,014778
ano * texp	-,056650	,020999	13,707	-2,698	,018	-,101777	-,011522

a. Dependent Variable: desempenho escolar.

Covariance Parameters

Estimates of Covariance Parameters[a]

Parameter		Estimate	Std. Error	Wald Z	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Residual		3,795045	,153657	24,698	,000	3,505523	4,108479
Intercept + ano [subject = estudante]	UN (1,1)	350,912601	20,688828	16,961	,000	312,618371	393,897688
	UN (2,1)	-13,250888	1,673703	-7,917	,000	-16,531285	-9,970490
	UN (2,2)	3,257499	,235014	13,861	,000	2,827965	3,752275
Intercept + ano [subject = escola]	UN (1,1)	88,734046	38,402010	2,311	,021	37,993584	207,238439
	UN (2,1)	-3,185216	1,904173	-1,673	,094	-6,917327	,546894
	UN (2,2)	,255415	,137804	1,853	,064	,088715	,735350

a. Dependent Variable: desempenho escolar.

Figura 16.52 Principais *outputs* do modelo de tendência linear com interceptos e inclinações aleatórias e as variáveis *sexo* de nível 2 e *texp* de nível 3, com termos aleatórios (u_{00k}, u_{10k}) e (r_{0jk}, r_{1jk}) correlacionados.

Portanto, a expressão do nosso modelo final apresenta a seguinte especificação[7]:

$$desempenho_{tjk} = 54,734 + 4,516.ano_{jk} - 14,702.sexo_{jk} + 1,179.texp_k + 0,652.sexo_{jk}.ano_{jk} - 0,057.texp_k.ano_{jk} + u_{00k} + u_{10k}.ano_{jk} + r_{0jk} + r_{1jk}.ano_{jk} + e_{tjk}$$

construído com a inclusão de variáveis explicativas dos níveis 2 e 3 e por meio da *multilevel step-up strategy*.

Podemos concluir, portanto, que o desempenho escolar dos estudantes segue uma tendência linear ao longo do tempo, existindo variância significativa de interceptos e de inclinações entre aqueles que estudam na mesma escola e entre aqueles que estudam em escolas distintas, o sexo dos estudantes é significante para explicar parte dessa variação e o tempo médio de experiência docente em cada escola também explica parte das discrepâncias no desempenho escolar anual entre os estudantes provenientes de diferentes escolas.

Analogamente ao Quadro 16.1 apresentado ao final da seção 16.4, o Quadro 16.2 consolida as rotinas gerais para estimação, em SPSS, de modelos multinível.

Quadro 16.2 Modelagens hierárquicas, modelos intermediários (*multilevel step-up strategy*) e rotinas em SPSS.

Modelagem	Modelo Intermediário	Rotina em SPSS
Hierárquica Linear de Dois Níveis com Dados Agrupados	Modelo Nulo (Modelo Não Condicional)	**MIXED Y** **/FIXED = INTERCEPT** **/RANDOM = INTERCEPT | SUBJECT(var_nível2) .**
	Modelo com Interceptos Aleatórios	**MIXED Y WITH X** **/FIXED = INTERCEPT X** **/RANDOM = INTERCEPT | SUBJECT(var_nível2) .**
	Modelo com Interceptos e Inclinações Aleatórias	**MIXED Y WITH X** **/FIXED = INTERCEPT X** **/RANDOM = INTERCEPT X | SUBJECT(var_nível2) .**
	Modelo com Interceptos e Inclinações Aleatórias e Termos de Erro Correlacionados	**MIXED Y WITH X** **/FIXED = INTERCEPT X** **/RANDOM = INTERCEPT X | SUBJECT(var_nível2) COVTYPE(UN) .**

[7] Analogamente, se o pesquisador também desejar comparar os resultados desta estimação com aqueles provenientes de uma estimação de um modelo considerando termos aleatórios independentes, assim como realizado em Stata, basta que ele digite a seguinte rotina na janela de sintaxes do SPSS:

```
MIXED desempenho WITH ano sexo texp
/METHOD = REML
/PRINT = SOLUTION TESTCOV
/FIXED = INTERCEPT ano sexo texp sexo*ano texp*ano
/RANDOM = INTERCEPT ano | SUBJECT(estudante)
/RANDOM = INTERCEPT ano | SUBJECT(escola) .
```

Os resultados, embora não apresentados aqui, geram valor de -2.*LL* igual a 14.839,357. Portanto, um teste de razão de verossimilhança apresentará nível de significância Sig. χ^2_2 (14.839,357 - 14.753,429 = 85,93) = 0,000 < 0,05, o que permite que afirmemos que a estrutura da matriz de variância-covariância entre os termos de erro pode ser considerada *unstructured* neste exemplo, ou seja, podemos considerar que os termos de erro u_{00k} e u_{10k} sejam correlacionados ($\text{cov}(u_{00k}, u_{10k}) \neq 0$) e que os termos de erro r_{0jk} e r_{1jk} também sejam correlacionados ($\text{cov}(r_{0jk}, r_{1jk}) \neq 0$).

Modelagem	Modelo Intermediário	Rotina em SPSS
Hierárquica Linear de Três Níveis com Medidas Repetidas	Modelo Nulo (Modelo Não Condicional)	**MIXED Y** **/FIXED = INTERCEPT** **/RANDOM = INTERCEPT | SUBJECT(var_nível2)** **/RANDOM = INTERCEPT | SUBJECT(var_nível3) .**
	Modelo de Tendência Linear com Interceptos Aleatórios	**MIXED Y WITH t** **/FIXED = INTERCEPT t** **/RANDOM = INTERCEPT | SUBJECT(var_nível2)** **/RANDOM = INTERCEPT | SUBJECT(var_nível3) .**
	Modelo de Tendência Linear com Interceptos e Inclinações Aleatórias	**MIXED Y WITH t** **/FIXED = INTERCEPT t** **/RANDOM = INTERCEPT t | SUBJECT(var_nível2)** **/RANDOM = INTERCEPT t | SUBJECT(var_nível3) .**
	Modelo de Tendência Linear com Interceptos e Inclinações Aleatórias e Variável de Nível 2	**MIXED Y WITH t X** **/FIXED = INTERCEPT t X X*t** **/RANDOM = INTERCEPT t | SUBJECT(var_nível2)** **/RANDOM = INTERCEPT t | SUBJECT(var_nível3) .**
	Modelo de Tendência Linear com Interceptos e Inclinações Aleatórias e Variáveis de Níveis 2 e 3	**MIXED Y WITH t X W** **/FIXED = INTERCEPT t X W X*t W*t W*X*t** **/RANDOM = INTERCEPT t | SUBJECT(var_nível2)** **/RANDOM = INTERCEPT t | SUBJECT(var_nível3) .**
	Modelo de Tendência Linear com Interceptos e Inclinações Aleatórias e Variáveis de Níveis 2 e 3 e Termos de Erro Correlacionados	**MIXED Y WITH t X W** **/FIXED = INTERCEPT t X W X*t W*t W*X*t** **/RANDOM = INTERCEPT t | SUBJECT(var_ nível2) COVTYPE(UN)** **/RANDOM = INTERCEPT t | SUBJECT(var_ nível3) COVTYPE(UN) .**

Nota: Considerada uma variável **X** de nível 2, uma variável **W** de nível 3 (quando houver) e **t** como variável temporal. Além disso, **Y** refere-se à variável dependente. Em todos os comandos, considerada a estimação por máxima verossimilhança restrita (termo omitido **/METHOD = REML)**.

16.6. CONSIDERAÇÕES FINAIS

Os modelos multinível de regressão para dados em painel possibilitam que o pesquisador avalie a relação entre determinada variável de desempenho e uma ou mais variáveis preditoras que caracterizam diferentes níveis de análise, sendo cada nível formado por indivíduos ou grupos aninhados em outros grupos e assim sucessivamente. Como variáveis de determinado grupo são invariantes entre grupos ou indivíduos correspondentes a níveis inferiores que estejam aninhados àquele grupo, é natural que muitas pesquisas usem tais modelos, uma vez que muitas bases apresentam estruturas aninhadas de dados, como aquelas que trazem, simultaneamente, características de estudantes e escolas, empresas e países, municípios e estados da federação ou imóveis e bairros, por exemplo.

Muitas podem ser as características das bases com estruturas aninhadas de dados, sendo as mais comuns aquelas com aninhamento absoluto em que há a presença de dados agrupados ou de dados com medidas repetidas. Neste

capítulo, optamos por apresentar exemplos em que são utilizadas bases para a estimação de modelos hierárquicos lineares de dois níveis com dados agrupados e de três níveis com medidas repetidas. Entretanto, a partir dos quais, acreditamos que o pesquisador tenha condições de estimar modelos, por exemplo, de três níveis com dados agrupados ou até mesmo considerando uma quantidade superior de níveis de análise, decorrentes de estruturas mais complexas de aninhamento.

Os modelos multinível permitem que sejam identificadas e analisadas as heterogeneidades individuais e entre grupos a que pertencem esses indivíduos, tornando possível a especificação de componentes aleatórios em cada nível da análise. E esse fato representa a principal diferença em relação aos tradicionais modelos de regressão estimados por MQO, que não conseguem levar em consideração o aninhamento natural dos dados e, consequentemente, geram estimadores viesados dos parâmetros.

Embora muitos trabalhos façam uso de modelagens multinível estimando apenas modelos nulos para a investigação da decomposição de variância do fenômeno em estudo nos diferentes níveis de análise, a possibilidade de inclusão de variáveis explicativas correspondentes aos distintos níveis nos componentes de efeitos fixos e aleatórios permite que sejam investigadas eventuais relações entre essas variáveis e a variável dependente, o que propicia a determinação de novos objetivos de pesquisa e o estabelecimento de constructos interessantes.

Recentemente, é possível perceber uma crescente preocupação de fabricantes de softwares com relação à capacidade de processamento de comandos e rotinas para a estimação de modelos multinível mais complexos. Não podemos deixar de mencionar o importante e didático software HLM (Hierarchical Linear and Nonlinear Modeling), produzido pela Scientific Software International (SSI) e desenvolvido pelos professores Stephen Raudenbush (University of Michigan), Anthony Bryk (University of Chicago) e Richard Congdon (Harvard University).

Para a estimação de modelos multinível, é necessário, assim como para qualquer outra técnica de modelagem, que a aplicação venha acompanhada de rigor metodológico e de certos cuidados na análise dos resultados, principalmente se estes tiverem como objetivo a elaboração de previsões. A adoção de determinado método de estimação, em detrimento de outro, pode auxiliar o pesquisador na escolha do modelo mais apropriado, valorizando a sua pesquisa e propiciando novos estudos sobre o tema escolhido.

Neste capítulo, procuramos elaborar, por meio da utilização de diferentes bases, algumas modelagens importantes para estruturas aninhadas de dados, adequadas para cada situação de uso. Além disso, também procuramos propiciar ao pesquisador uma oportunidade de aplicar esses diferentes tipos de estimações nos softwares Stata e SPSS, o que acaba por favorecer o seu manuseio.

16.7. EXERCÍCIOS

1) A organização de uma competição internacional de ciências para estudantes do ensino médio provenientes de 24 países ($j = 1, ..., 24$) deseja investigar o comportamento do desempenho dos participantes em função de suas características e das características dos países de onde vieram. Embora os coordenadores do evento saibam que o desempenho é reflexo de diversos fatores, como dedicação dos participantes e das próprias características das escolas em que estudam, o desejo, neste momento, é tentar verificar se há relação entre as notas obtidas na competição, o nível social dos estudantes, traduzido pela renda média familiar, e a importância dispensada pelos países em quesitos como desenvolvimento científico e tecnológico, traduzida aqui pelo investimento em pesquisa e desenvolvimento. A base coletada, que contém dados dos cinco mais bem classificados estudantes de cada país, o que totaliza 120 participantes na competição ($i = 1, ..., 120$) e gera uma estrutura equilibrada de dados agrupados, pode ser acessada por meio do arquivo **Competição de Ciências.dta**. As variáveis presentes nesta base são:

Variável	Descrição
país	Variável *string* que identifica o país.
idpaís	Código do país *j*.
pesqdes	Investimento do país em pesquisa e desenvolvimento, em % do PIB (Fonte: Banco Mundial).
idestudante	Código do estudante *i*.
nota	Nota de ciências obtida pelo estudante na competição (0 a 100).
renda	Renda média mensal da família do estudante (US$).

Por meio do uso desta base de dados, pede-se:

a) Elabore uma tabela que comprove a existência de uma estrutura equilibrada de dados agrupados de estudantes em países.

b) Elabore gráficos que permitam a visualização da nota média obtida na competição de ciências pelos participantes de cada país.

c) Dada existência de dois níveis de análise, com estudantes (nível 1) aninhados em países (nível 2), estime o seguinte modelo nulo:

$$nota_{ij} = b_{0j} + r_{ij}$$
$$b_{0j} = \gamma_{00} + u_{0j}$$

que resulta em:

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

d) Por meio da estimação do modelo nulo, é possível verificar que existe variabilidade da nota obtida entre estudantes provenientes de diferentes países?

e) A partir do resultado do teste de razão de verossimilhança gerado, é possível rejeitar a hipótese nula de que os interceptos aleatórios sejam iguais a zero, ou seja, é possível descartar a estimação de um modelo tradicional de regressão linear para estes dados agrupados?

f) Ainda com base na estimação do modelo nulo, calcule a correlação intraclasse e discuta o resultado.

g) Elabore um gráfico que apresente o ajuste linear por MQO, para cada país, do comportamento da nota de ciências de cada estudante em função da renda média mensal familiar.

h) Estime o seguinte modelo com interceptos aleatórios:

$$nota_{ij} = b_{0j} + b_{1j}.renda_{ij} + r_{ij}$$
$$b_{0j} = \gamma_{00} + u_{0j}$$
$$b_{1j} = \gamma_{10}$$

que resulta em:

$$nota_{ij} = \gamma_{00} + \gamma_{10}.renda_{ij} + u_{0j} + r_{ij}$$

i) Discuta a significância estatística, ao nível de 5% de significância, das estimações dos parâmetros de efeitos fixos e aleatórios.

j) Elabore um gráfico de barras que permita a visualização dos termos de intercepto aleatório u_{0j} por país.

k) Estime o seguinte modelo com interceptos e inclinações aleatórias:

$$nota_{ij} = b_{0j} + b_{1j}.renda_{ij} + r_{ij}$$
$$b_{0j} = \gamma_{00} + u_{0j}$$
$$b_{1j} = \gamma_{10} + u_{1j}$$

que resulta em:

$$nota_{ij} = \gamma_{00} + \gamma_{10}.renda_{ij} + u_{0j} + u_{1j}.renda_{ij} + r_{ij}$$

l) Com base nas estimações do modelo com interceptos aleatórios e do modelo com interceptos e inclinações aleatórias, elabore um teste de razão de verossimilhança e discuta o resultado.

m) Estime o seguinte modelo multinível:

$$nota_{ij} = b_{0j} + b_{1j}.renda_{ij} + r_{ij}$$
$$b_{0j} = \gamma_{00} + u_{0j}$$
$$b_{1j} = \gamma_{10} + \gamma_{11}.pesqdes_j$$

que resulta em:

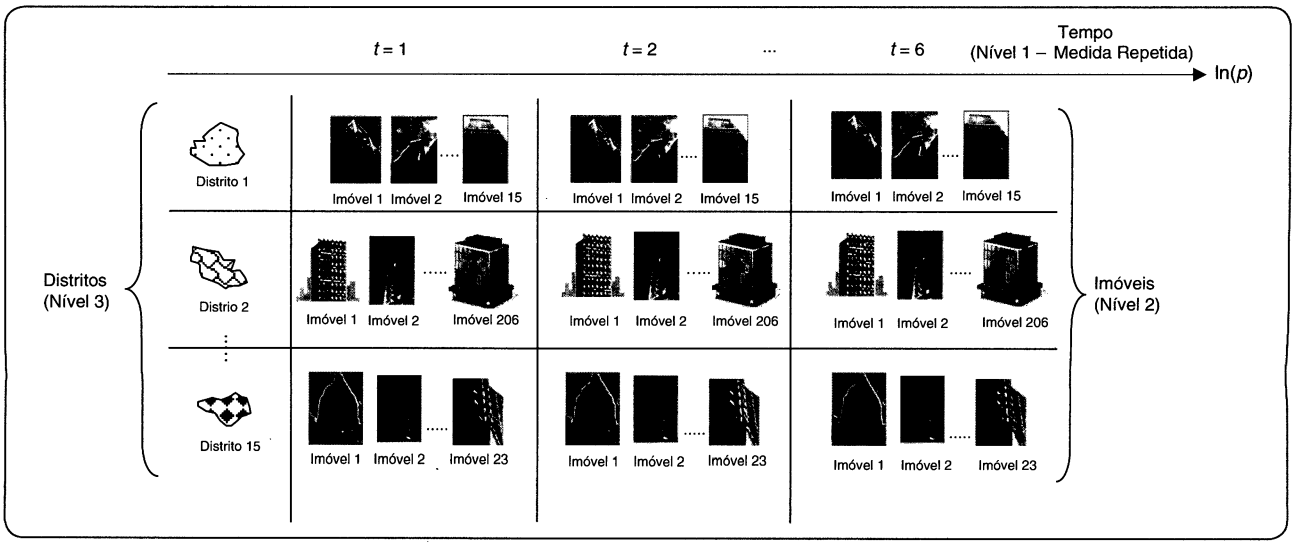
$$nota_{ij} = \gamma_{00} + \gamma_{10}.renda_{ij} + \gamma_{11}.pesqdes_j.renda_{ij} + u_{0j} + r_{ij}$$

n) Apresente a expressão do último modelo estimado, com interceptos aleatórios e variáveis de níveis 1 e 2.
o) Elabore um gráfico em que seja possível comparar os valores previstos da nota obtida na competição de ciências gerados por esta modelagem hierárquica de dois níveis (HLM2) com os valores reais obtidos (valores observados) pelos estudantes da amostra.

2) Uma empresa de locação de escritórios comerciais possui uma carteira de 277 imóveis em determinado município, e sua diretoria deseja saber se existem diferenças nos preços de aluguel por metro quadrado entre imóveis e também nos preços médios de aluguel dos imóveis entre diferentes distritos, ao longo do tempo. Para tanto, a equipe de marketing estruturou a base de dados, que se encontra no arquivo **Imóveis Comerciais.dta**, com características desses 277 escritórios já locados ($j = 1, ..., 277$), cujos preços firmados de locação foram monitorados ao longo dos últimos seis anos ($t = 1, ..., 6$), e dos 15 distritos municipais ($k = 1, ..., 15$) em que se localizam os imóveis. As variáveis presentes nesta base são:

Variável	Descrição
distrito	Código do distrito *k*.
imóvel	Código do imóvel *j*.
lnp	Logaritmo natural do preço de aluguel por metro quadrado (ajustado pela inflação, base ano 1).
ano	Variável temporal (medida repetida) correspondente ao período de monitoramento (ano 1 a 6).
alim	Existência de restaurante ou praça de alimentação no empreendimento em que se encontra o imóvel (Não = 0; Sim = 1).
vaga4	Existência de uma quantidade de vagas de estacionamento maior ou igual a quatro (Não = 0; Sim = 1).
valet	Existência de *valet park* no edifício do escritório (Não = 0; Sim = 1).
metrô	Existência de estação de metrô no distrito onde está localizado o imóvel (Não = 0; Sim = 1).
violência	Taxa média de mortalidade por causas externas no distrito onde está localizado o imóvel (por cem mil habitantes).

Essa base de dados, em que períodos (nível 1) estão aninhados em imóveis (nível 2), e esses em distritos (nível 3), está estruturada conforme a lógica apresentada na figura a seguir:



Pede-se:

a) Elabore uma tabela que comprove a existência de uma estrutura desequilibrada de dados agrupados de imóveis em distritos.
b) Elabore uma tabela que comprove a existência de um painel desbalanceado de dados em relação aos períodos de monitoramento dos imóveis.

c) Elabore um gráfico que permita que seja visualizada a evolução temporal do logaritmo natural do preço de aluguel por metro quadrado dos imóveis em análise.

d) Elabore um gráfico que permita que se verifique a existência de um comportamento aproximadamente linear da média do logaritmo natural do preço de aluguel por metro quadrado dos imóveis ao longo dos períodos de tempo.

e) Elabore um gráfico que apresente, por distrito municipal, as evoluções temporais das médias dos logaritmos naturais dos preços de aluguel por metro quadrado dos imóveis (ajustes lineares por MQO).

f) Dada existência de três níveis de análise, com medidas repetidas (nível 1) aninhadas a imóveis (nível 2), e estes aninhados a distritos municipais (nível 3), estime o seguinte modelo nulo:

$$\ln(p)_{tjk} = \pi_{0jk} + e_{tjk}$$
$$\pi_{0jk} = b_{00k} + r_{0jk}$$
$$b_{00k} = \gamma_{000} + u_{00k}$$

que resulta em:

$$\ln(p)_{tjk} = \gamma_{000} + u_{00k} + r_{0jk} + e_{tjk}$$

g) Com base na estimação do modelo nulo, calcule as correlações intraclasse de níveis 2 e 3 e discuta os resultados.

h) Ainda por meio da estimação do modelo nulo, é possível afirmar que há variabilidade no preço de aluguel dos imóveis comerciais ao longo do período analisado e que há variabilidade no preço de aluguel, ao longo do tempo, entre imóveis de um mesmo distrito e entre imóveis localizados em distritos diferentes?

i) A partir do resultado do teste de razão de verossimilhança gerado, é possível rejeitar a hipótese nula de que os interceptos aleatórios sejam iguais a zero, ou seja, é possível descartar a estimação de um modelo tradicional de regressão linear para estes dados?

j) Estime o seguinte modelo de tendência linear com interceptos aleatórios:

$$\ln(p)_{tjk} = \pi_{0jk} + \pi_{1jk}.ano_{jk} + e_{tjk}$$
$$\pi_{0jk} = b_{00k} + r_{0jk}$$
$$\pi_{1jk} = b_{10k}$$
$$b_{00k} = \gamma_{000} + u_{00k}$$
$$b_{10k} = \gamma_{100}$$

que resulta na seguinte expressão:

$$\ln(p)_{tjk} = \gamma_{000} + \gamma_{100}.ano_{jk} + u_{00k} + r_{0jk} + e_{tjk}$$

k) Discuta a significância estatística, ao nível de 5% de significância, das estimações dos parâmetros de efeitos fixos e aleatórios.

l) Elabore dois gráficos de barras que permitam a visualização dos interceptos aleatórios por distrito e por imóvel.

m) Estime o seguinte modelo de tendência linear com interceptos e inclinações aleatórias:

$$\ln(p)_{tjk} = \pi_{0jk} + \pi_{1jk}.ano_{jk} + e_{tjk}$$
$$\pi_{0jk} = b_{00k} + r_{0jk}$$
$$\pi_{1jk} = b_{10k} + r_{1jk}$$
$$b_{00k} = \gamma_{000} + u_{00k}$$
$$b_{10k} = \gamma_{100} + u_{10k}$$

que resulta em:

$$\ln(p)_{tjk} = \gamma_{000} + \gamma_{100}.ano_{jk} + u_{00k} + u_{10k}.ano_{jk} + r_{0jk} + r_{1jk}.ano_{jk} + e_{tjk}$$

n) Calcule as novas correlações intraclasse de níveis 2 e 3 e discuta os resultados.

o) Elabore um teste de razão de verossimilhança para comparar as estimações dos modelos de tendência linear com interceptos aleatórios e com interceptos e inclinações aleatórias.

p) Estime o seguinte modelo de tendência linear com interceptos e inclinações aleatórias e variáveis de nível 2:

$$\begin{aligned}
&\ln(p)_{tjk} = \pi_{0jk} + \pi_{1jk}.ano_{jk} + e_{tjk}\\
&\pi_{0jk} = b_{00k} + b_{01k}.alim_{jk} + b_{02k}.vaga4_{jk} + r_{0jk}\\
&\pi_{1jk} = b_{10k} + b_{11k}.valet_{jk} + r_{1jk}\\
&b_{00k} = \gamma_{000} + u_{00k}\\
&b_{01k} = \gamma_{010}\\
&b_{02k} = \gamma_{020}\\
&b_{10k} = \gamma_{100} + u_{10k}\\
&b_{11k} = \gamma_{110}
\end{aligned}$$

que resulta na seguinte expressão:

$$\begin{aligned}
\ln(p)_{tjk} = {} & \gamma_{000} + \gamma_{100}.ano_{jk} + \gamma_{010}.alim_{jk} + \gamma_{020}.vaga4_{jk} + \gamma_{110}.valet_{jk}.ano_{jk}\\
& + u_{00k} + u_{10k}.ano_{jk} + r_{0jk} + r_{1jk}.ano_{jk} + e_{tjk}
\end{aligned}$$

q) Apresente a expressão do último modelo estimado, com medidas repetidas, interceptos e inclinações aleatórias e variáveis de nível 2.

r) Por meio deste modelo, é possível afirmar que o logaritmo natural do preço de aluguel por metro quadrado dos imóveis segue uma tendência linear ao longo do tempo, existindo variância significativa de interceptos e de inclinações entre aqueles localizados no mesmo distrito e entre aqueles localizados em distritos distintos? Em caso afirmativo, a existência de restaurante ou praça de alimentação no empreendimento, a existência de uma quantidade de vagas de estacionamento maior ou igual a quatro e a existência de *valet park* no edifício onde está o imóvel explicam parte dessa variabilidade?

s) Estime o seguinte modelo de tendência linear com interceptos e inclinações aleatórias e variáveis de níveis 2 e 3:

$$\begin{aligned}
&\ln(p)_{tjk} = \pi_{0jk} + \pi_{1jk}.ano_{jk} + e_{tjk}\\
&\pi_{0jk} = b_{00k} + b_{01k}.alim_{jk} + b_{02k}.vaga4_{jk} + r_{0jk}\\
&\pi_{1jk} = b_{10k} + b_{11k}.valet_{jk} + r_{1jk}\\
&b_{00k} = \gamma_{000} + \gamma_{001}.metrô_{k} + u_{00k}\\
&b_{01k} = \gamma_{010}\\
&b_{02k} = \gamma_{020}\\
&b_{10k} = \gamma_{100} + \gamma_{101}.metrô_{k} + \gamma_{102}.violência_{k} + u_{10k}\\
&b_{11k} = \gamma_{110}
\end{aligned}$$

que resulta na seguinte expressão:

$$\begin{aligned}
\ln(p)_{tjk} = {} & \gamma_{000} + \gamma_{100}.ano_{jk} + \gamma_{010}.alim_{jk} + \gamma_{020}.vaga4_{jk} + \gamma_{001}.metrô_{k}\\
& + \gamma_{110}.valet_{jk}.ano_{jk} + \gamma_{101}.metrô_{k}.ano_{jk} + \gamma_{102}.violência_{k}.ano_{jk}\\
& + u_{00k} + u_{10k}.ano_{jk} + r_{0jk} + r_{1jk}.ano_{jk} + e_{tjk}
\end{aligned}$$

t) Apresente as matrizes de variância-covariância dos efeitos aleatórios para os níveis *distrito* e *imóvel*.

u) Estime o mesmo modelo de tendência linear com interceptos e inclinações aleatórias e variáveis de níveis 2 e 3, porém agora considerando termos aleatórios (u_{00k}, u_{10k}) e (r_{0jk}, r_{1jk}) correlacionados.

v) Apresente as novas matrizes de variância-covariância dos efeitos aleatórios para os níveis *distrito* e *imóvel*.

w) Elabore um teste de razão de verossimilhança para comparar as estimações dos modelos com termos aleatórios (u_{00k}, u_{10k}) e (r_{0jk}, r_{1jk}) independentes e correlacionados. O que se pode concluir com base no resultado do teste?

x) Qual a expressão final do modelo multinível estimado?

y) É possível afirmar que a existência de metrô e o indicador de violência no distrito explicam parte da variabilidade da evolução do logaritmo natural do preço de aluguel por metro quadrado entre imóveis localizados em diferentes distritos?

z) Elabore um gráfico em que seja possível comparar os valores previstos do logaritmo natural do preço de aluguel por metro quadrado gerados por esta modelagem hierárquica de três níveis (HLM3) com aqueles gerados por meio de uma estimação por MQO que faz uso das mesmas variáveis explicativas do modelo do item (x) inseridas no componente de efeitos fixos (*ano*, *alim*, *vaga4*, *metrô*, *valet***ano*, *metrô***ano* e *violência***ano*), e com os valores reais observados do logaritmo natural do preço de aluguel por metro quadrado dos imóveis.

APÊNDICE

Modelos hierárquicos não lineares

Conforme discutimos, os modelos lineares generalizados multinível (*generalized linear latent and mixed models - GLLAMM*), analogamente aos modelos lineares generalizados (*GLM*), comportam os modelos hierárquicos lineares (*HLM*), estudados ao longo do capítulo, e os **modelos hierárquicos não lineares** (***hierarchical non linear models - HNM***). Estes últimos, por sua vez, referem-se a situações em que, existindo uma estrutura aninhada de dados, a variável dependente apresenta-se de maneira categórica ou com dados de contagem, razão pela qual optamos por apresentar, no presente apêndice, exemplos de modelos hierárquicos não lineares dos tipos logístico, Poisson e binomial negativo. A Figura 16.53 apresenta a lógica dos modelos lineares generalizados multinível, com destaque para os modelos que serão estudados a partir de agora.

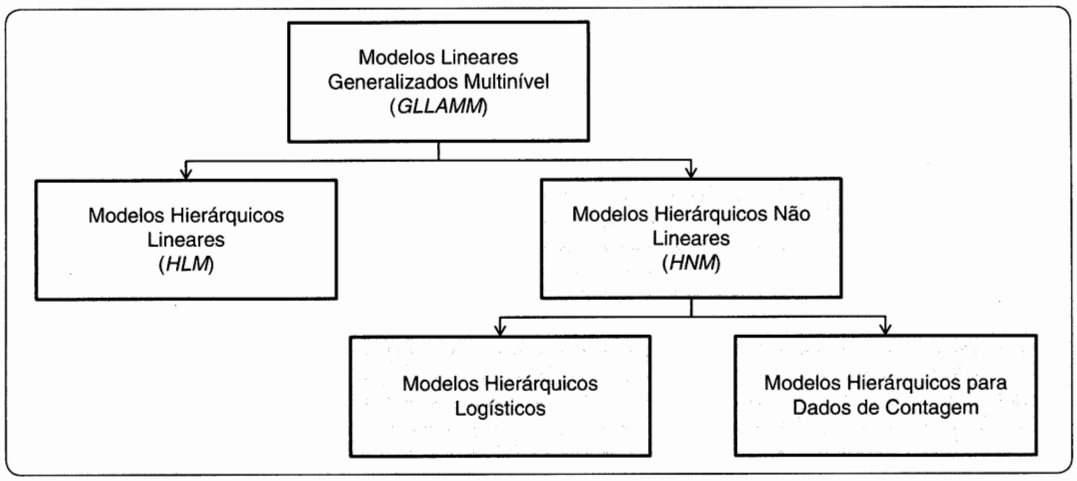


Figura 16.53 Modelos lineares generalizados multinível, com destaque para os modelos hierárquicos não lineares.

A) Modelos Hierárquicos Logísticos

De maneira análoga ao estudado no Capítulo 13 e na seção 15.4.1 do Capítulo 15, os **modelos de regressão logística com efeitos mistos** podem ser utilizados quando a variável dependente apresentar-se de maneira qualitativa e dicotômica e os dados estiverem dispostos em determinada estrutura aninhada (em níveis), podendo haver dados agrupados ou com medidas repetidas. Nessas situações, o pesquisador pode estimar um modelo com o intuito de capturar a relação entre o comportamento de variáveis explicativas e a ocorrência do fenômeno em estudo, representado por uma variável dicotômica (*dummy*), bem como estudar a decomposição de variância dos componentes de efeitos aleatórios decorrentes da presença de uma estrutura multinível.

Nesta seção, apresentaremos um modelo hierárquico logístico de dois níveis com dados agrupados. De maneira geral, e partindo das expressões (13.10) e (16.23), podemos definir, da seguinte maneira, este modelo com dois níveis de análise, em que o primeiro nível oferece as variáveis explicativas $X_1, ..., X_Q$ referentes a cada indivíduo i ($i = 1, ..., n$), e o segundo nível, as variáveis explicativas $W_1, ..., W_S$ referentes a cada grupo j ($j = 1, ..., J$), invariantes para as observações pertencentes a um mesmo grupo:

Nível 1:
$$p_{ij} = \frac{1}{1+e^{-\left(b_{0j}+b_{1j}.X_{1ij}+b_{2j}.X_{2ij}+...+b_{Qj}.X_{Qij}\right)}} \quad (16.45)$$

em que p_{ij} representa a probabilidade de ocorrência do evento de interesse para cada observação i pertencente a determinado grupo j e b_{qj} ($q = 0, 1, ..., Q$) referem-se aos coeficientes de nível 1.

Nível 2:
$$b_{qj} = \gamma_{q0} + \sum_{s=1}^{S_q} \gamma_{qs}.W_{sj} + u_{qj} \quad (16.46)$$

em que γ_{qs} ($s = 0, 1, ..., S_q$) referem-se aos coeficientes de nível 2 e u_{qj} são os efeitos aleatórios de nível 2, normalmente distribuídos, com média igual a zero e variância τ_{qq}. Além disso, eventuais termos de erro independentes de u_{qj} apresentam média igual a zero e variância $\pi^2/3$.

Vamos, neste momento, apresentar um exemplo. Uma pesquisa foi elaborada em nível global com o intuito de investigar se existem diferenças na realização de viagens internacionais de turismo entre casais residentes em diferentes países. Para tanto, coletaram-se dados de 1.622 casais localizados em 50 países, como a idade média do casal e a quantidade de filhos. Parte do banco de dados elaborado é apresentada na Tabela 16.5, porém a base de dados completa pode ser acessada por meio do arquivo **Turismo.dta**.

Tabela 16.5 Exemplo: realização de viagens internacionais de casais (nível 1) residentes em diferentes países (nível 2).

Observação (Casal i – Nível 1)	País j em que o casal mora (Nível 2)	Realizou viagem internacional de turismo no último ano (Y_{ij})	Idade média do casal (X_{1ij})	Quantidade de filhos (X_{2ij})
1	França	Sim	68	2
2	França	Sim	37	0
...				
117	França	Sim	54	3
...				
1.604	Egito	Não	55	2
1.605	Egito	Não	51	2
...				
1.622	Egito	Sim	39	0

Após abrirmos esse arquivo, podemos digitar o comando **desc**, que faz com que seja possível analisarmos as características do banco de dados, como a quantidade de observações, a quantidade de variáveis e a descrição de cada uma delas. A Figura 16.54 apresenta este *output* do Stata.

```
. desc

  obs:         1,622
 vars:             4
 size:        42,172
-------------------------------------------------------------------------------
              storage   display    value
variable name   type    format     label      variable label
-------------------------------------------------------------------------------
país            str14   %14s                  país j em que o casal mora (nível 2)
turismo         float   %9.0g      turismo    realizou viagem internacional de turismo
                                              no último ano?
idade           float   %9.0g                 idade média do casal (anos)
filhos          float   %9.0g                 quantidade de filhos
-------------------------------------------------------------------------------
Sorted by:
```

Figura 16.54 Descrição do banco de dados **Turismo.dta**.

Como o intuito neste apêndice não é o de discutir novamente os conceitos abordados ao longo do capítulo, vamos partir para a estimação seguinte:

$$p(turismo)_{ij} = \frac{1}{1+e^{-(b_{0j}+b_{1j}.idade_{ij}+b_{2j}.filhos_{ij})}}$$

$$b_{0j} = \gamma_{00} + u_{0j}$$
$$b_{1j} = \gamma_{10}$$
$$b_{2j} = \gamma_{20}$$

que resulta no modelo com interceptos aleatórios:

$$p(turismo)_{ij} = \frac{1}{1+e^{-(\gamma_{00}+\gamma_{10}.idade_{ij}+\gamma_{20}.filhos_{ij}+u_{0j})}}$$

sendo a variável *turismo* dicotômica (*dummy*), em que valores iguais a 1 correspondem a casais que realizaram viagens internacionais de turismo no último ano e valores iguais a 0, caso contrário.

Para a estimação deste modelo no Stata, devemos digitar o seguinte comando:

```
melogit turismo idade filhos || país: , nolog
```
[8]

cujos *outputs* são apresentados na Figura 16.55.

```
. melogit turismo idade filhos || país: , nolog

Mixed-effects logistic regression               Number of obs      =      1622
Group variable: país                            Number of groups   =        50

                                                Obs per group: min =         2
                                                               avg =      32.4
                                                               max =       118

Integration points =   7                        Wald chi2(2)       =     52.18
Log likelihood = -1038.1176                     Prob > chi2        =    0.0000

------------------------------------------------------------------------------
     turismo |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       idade |   .0150543   .0066673     2.26   0.024     .0019866    .0281221
      filhos |  -.4239421   .0598524    -7.08   0.000    -.5412507   -.3066335
       _cons |   .4393716   .2954913     1.49   0.137    -.1397806    1.018524
------------------------------------------------------------------------------

------------------------------------------------------------------------------
  Random-effects Parameters  |   Estimate   Std. Err.     [95% Conf. Interval]
-----------------------------+------------------------------------------------
país: Identity               |
                  var(_cons) |   .2551956   .0880873      .1297356    .5019808
------------------------------------------------------------------------------
LR test vs. logistic regression: chibar2(01) =    52.82 Prob>=chibar2 = 0.0000
```

Figura 16.55 *Outputs* do modelo hierárquico logístico com interceptos aleatórios no Stata.

Com base nessa figura, podemos inicialmente verificar que temos 1.622 observações (casais) aninhadas em 50 grupos (países), o que caracteriza a estrutura de dados agrupados em dois níveis.

Um pesquisador mais curioso poderá verificar que as estimações dos parâmetros dos componentes de efeitos fixos e aleatórios são idênticas às que seriam obtidas por meio do seguinte comando:

```
meglm turismo idade filhos || país: , family(bernoulli) link(logit) nolog
```

[8] Para versões anteriores à versão 13 do Stata, o comando deverá ser **`xtmelogit turismo idade filhos || país: , var nolog`**.

em que o termo **meglm** significa *multilevel mixed-effects generalized linear model* e que, portanto, torna necessária a definição da família de distribuições da variável dependente que, neste caso, é a Bernoulli, e da função de ligação canônica que, nesta situação, é a logística[9].

Além disso, também podem ser diretamente obtidas as *odds ratios* dos parâmetros de efeitos fixos, digitando-se o termo **or** (*odds ratio*) ao final dos comandos apresentados.

Dado que os termos de erro independentes de u_{qj} apresentam variância igual $\pi^2/3$, podemos definir a seguinte correlação intraclasse:

$$rho = \frac{\tau_{00}}{\tau_{00} + \frac{\pi^2}{3}} = \frac{0,255}{0,255 + \frac{\pi^2}{3}} = 0,072$$

que indica que aproximadamente 7% da variância total dos termos de erro é devido à alteração do comportamento da variável dependente entre países. A partir da versão 13 do Stata, é possível obter diretamente esta correlação intraclasse, digitando-se o comando **estat icc** logo após a estimação do correspondente modelo.

Embora o Stata não mostre, de maneira direta, o resultado dos testes z com os respectivos níveis de significância para os parâmetros de efeitos aleatórios, o fato de a estimação do componente de variância τ_{00}, correspondente ao intercepto aleatório u_{0j}, ser consideravelmente superior ao seu erro-padrão indica haver alteração significante no comportamento de casais residentes em diferentes países em relação à realização de viagens internacionais de turismo. Estatisticamente, podemos verificar que $z = 0,255 / 0,088 = 2,90 > 1,96$, sendo 1,96 o valor crítico da distribuição normal padrão que resulta em um nível de significância de 5%.

Mesmo que não tenham sido consideradas variáveis de países que podem eventualmente explicar tal comportamento, como características culturais, econômicas ou sociais, temos condições de verificar que, enquanto o incremento de idade aumenta a probabilidade esperada de que casais passem a realizar viagens internacionais de turismo, *ceteris paribus*, a realização dessas viagens diminui com o incremento da quantidade de filhos, também *ceteris paribus*. O modelo estimado apresenta a seguinte expressão:

$$p(turismo)_{ij} = \frac{1}{1 + e^{-(0,439 + 0,015.idade_{ij} - 0,424.filhos_{ij} + u_{0j})}}$$

Na parte inferior da Figura 16.55, podemos verificar, pelo resultado do teste de razão de verossimilhança, que a estimação deste modelo multinível é mais adequada do que a estimação de um modelo tradicional de regressão logística binária para os dados do nosso exemplo.

Portanto, podemos obter os valores das probabilidades esperadas de ocorrência do evento em estudo (realização de viagem internacional de turismo) para cada um dos casais da amostra. Para tanto, devemos digitar o seguinte comando, que gera uma nova variável (*phat*) no banco de dados:

```
predict phat
```

Além disso, também podemos obter os termos de erro u_{0j}, invariantes para casais de um mesmo país. Para tanto, devemos digitar o seguinte comando:

```
predict u0, remeans
```

que faz com que nova variável, *u0*, também seja gerada no banco de dados.

O comando a seguir, que gera os *outputs* da Figura 16.56, mostra os valores de *phat* e os termos de erro *u0* apenas para os casais residentes no Brasil:

```
list país turismo phat u0 if país == "Brasil"
```

---

[9] Se o pesquisador optar por estimar um modelo hierárquico não linear do tipo probit, cuja distribuição da variável dependente também é a Bernoulli, conforme estudamos no apêndice do Capítulo 13, poderá usar um dos dois comandos a seguir:

```
meprobit turismo idade filhos || país: , nolog
meglm turismo idade filhos || país: , family(bernoulli) link(probit) nolog
```

```
. list país turismo phat u0 if país == "Brasil"

     +-----------------------------------------+
     |   país   turismo       phat         u0 |
     |-----------------------------------------|
1198. | Brasil       Sim   .6316937   .1049601 |
1199. | Brasil       Não    .491252   .1049601 |
1200. | Brasil       Sim   .7533196   .1049601 |
1201. | Brasil       Sim    .747682   .1049601 |
1202. | Brasil       Sim   .4950149   .1049601 |
     |-----------------------------------------|
1203. | Brasil       Não    .491252   .1049601 |
1204. | Brasil       Não   .4874901   .1049601 |
1205. | Brasil       Sim    .717749   .1049601 |
1206. | Brasil       Sim   .6659743   .1049601 |
1207. | Brasil       Sim   .6068546   .1049601 |
     |-----------------------------------------|
1208. | Brasil       Não   .6068546   .1049601 |
1209. | Brasil       Sim   .6032571   .1049601 |
1210. | Brasil       Sim   .6175761   .1049601 |
1211. | Brasil       Sim   .6495774   .1049601 |
1212. | Brasil       Sim   .6731711   .1049601 |
     |-----------------------------------------|
1213. | Brasil       Não   .7207888   .1049601 |
1214. | Brasil       Não   .6862789   .1049601 |
     +-----------------------------------------+
```

Figura 16.56 Probabilidades esperadas de realização de viagem internacional de turismo e termos de erro u_{0j} para casais residentes no Brasil (j = Brasil).

Apenas para fins didáticos, o pesquisador poderá verificar que a variável *phat* também pode ser gerada por meio da seguinte expressão:

```
gen phat = (1) / (1 + exp(-(0.4393717 + 0.0150543*idade -
0.4239421*filhos + u0)))
```

Por fim, podemos elaborar um gráfico que mostra, em função da variável *filhos*, os ajustes das curvas *S* (funções sigmoides) das probabilidades esperadas de que casais residentes em cinco específicos países, escolhidos em função de suas localizações distintas no globo, realizem viagens internacionais de turismo. Este gráfico, apresentado na Figura 16.57, é obtido por meio da digitação do seguinte comando:

```
graph twoway scatter phat filhos || mspline phat filhos if
país=="França" || mspline phat filhos if país=="Estados Unidos" || mspline
phat filhos if país=="Japão" || mspline phat filhos if país=="África
do Sul" || mspline phat filhos if país=="Venezuela" ||, legend(label(2
"França") label(3 "Estados Unidos") label(4 "Japão") label(5 "África do
Sul") label(6 "Venezuela"))
```

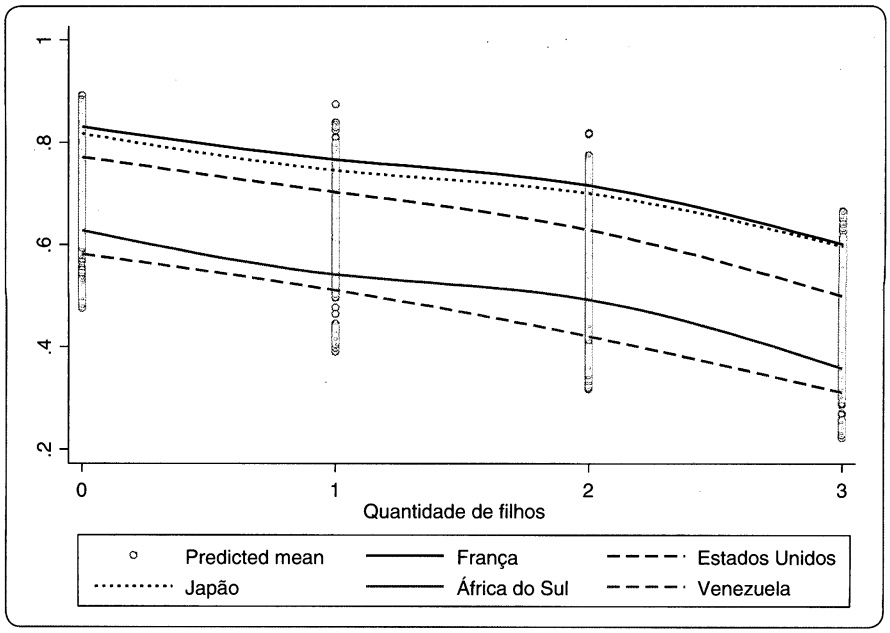


Figura 16.57 Ajustes das probabilidades esperadas de que casais residentes em cinco países realizem viagens internacionais de turismo, em função da quantidade de filhos.

Por meio deste gráfico, temos condições, de fato, de visualizar os comportamentos distintos entre casais provenientes de países diferentes em relação à realização de viagens internacionais de turismo.

B) Modelos Hierárquicos para Dados de Contagem

Analogamente ao estudado no Capítulo 14 e na seção 15.4.2 do Capítulo 15, os **modelos de regressão para dados de contagem com efeitos mistos** podem ser utilizados quando a variável dependente apresentar-se na forma quantitativa, porém com valores discretos e não negativos, e os dados estiverem dispostos em determinada estrutura aninhada (em níveis), podendo haver dados agrupados ou com medidas repetidas.

Nesta seção, apresentaremos um modelo hierárquico para dados de contagem com três níveis e dados agrupados. De maneira geral, e partindo-se das expressões (14.4), (16.30) e (16.31), podemos definir, da seguinte maneira, este modelo de três níveis, em que o primeiro nível apresenta as variáveis explicativas $Z_1, ..., Z_P$ referentes às unidades i ($i = 1, ..., n$) de nível 1, o segundo nível, as variáveis explicativas $X_1, ..., X_Q$ referentes às unidades j ($j = 1, ..., J$) de nível 2 e invariantes paras as unidades pertencentes a um mesmo grupo j, e o terceiro nível, as variáveis explicativas $W_1, ..., W_S$ referentes às unidades k ($k = 1, ..., K$) de nível 3 e invariantes para as unidades pertencentes a um mesmo grupo k:

Nível 1:
$$\ln\left(\lambda_{ijk}\right) = \pi_{0jk} + \pi_{1jk}.Z_{1jk} + \pi_{2jk}.Z_{2jk} + ... + \pi_{Pjk}.Z_{Pjk} \tag{16.47}$$

em que λ é o número esperado de ocorrências ou a taxa média estimada de incidência do fenômeno em estudo para dada exposição, π_{pjk} ($p = 0, 1, ..., P$) referem-se aos coeficientes de nível 1 e Z_{pjk} é uma p-ésima variável explicativa de nível 1 para a observação i na unidade de nível 2 j e na unidade de nível 3 k.

Nível 2:
$$\pi_{pjk} = b_{p0k} + \sum_{q=1}^{Q_p} b_{pqk}.X_{qjk} + r_{pjk} \tag{16.48}$$

em que b_{pqk} ($q = 0, 1, ..., Q_p$) referem-se aos coeficientes de nível 2, X_{qjk} é uma q-ésima variável explicativa de nível 2 para a unidade j na unidade de nível 3 k, e r_{pjk} são os efeitos aleatórios do nível 2, assumindo-se, para cada unidade j, que o vetor $(r_{0jk}, r_{1jk}, ..., r_{Pjk})'$ apresenta distribuição normal multivariada com cada elemento possuindo média zero e variância $\tau_{r\pi pp}$.

Nível 3:
$$b_{pqk} = \gamma_{pq0} + \sum_{s=1}^{S_{pq}} \gamma_{pqs}.W_{sk} + u_{pqk} \tag{16.49}$$

em que γ_{pqs} ($s = 0, 1, ..., S_{pq}$) referem-se aos coeficientes de nível 3, W_{sk} é uma s-ésima variável explicativa de nível 3 para a unidade k, e u_{pqk} são os efeitos aleatórios do nível 3, assumindo-se que para cada unidade k, o vetor composto pelos termos u_{pqk} apresenta distribuição normal multivariada com cada elemento possuindo média zero e variância $\tau_{u\pi pp}$.

Imagine que tenha sido realizada uma pesquisa nacional com o objetivo de estudar, no último ano, a relação entre a quantidade de acidentes de trânsito e a quantidade média de álcool ingerida por habitante/dia (em gramas) em diversos distritos municipais localizados em todo o território nacional, bem como se existem diferenças nessa relação entre distritos situados em diferentes municípios e diferentes estados da federação. Para tanto, foram pesquisados dados de 1.062 distritos municipais localizados em 234 municípios das 27 unidades federativas (26 estados e Distrito Federal). Parte do banco de dados elaborado é apresentada na Tabela 16.6, porém a base de dados completa pode ser acessada por meio do arquivo **Acidentes de Trânsito.dta**.

Tabela 16.6 Exemplo: acidentes de trânsito em distritos municipais (nível 1) de diferentes municípios (nível 2) e diferentes estados (nível 3).

Estado *k* (Nível 3)	Município *j* (Nível 2)	Distrito municipal *i* (Nível 1)	Quantidade de acidentes de trânsito no último ano (Y_{ijk})	Quantidade média de álcool ingerida por habitante/dia, em gramas (Z_{jk})
AC	1	1	9	12,57
AC	2	2	10	13,36
		...		
AC	3	11	2	12,33
		...		
TO	231	1.052	2	11,94
TO	231	1.053	3	10,54
		...		
TO	234	1.062	5	11,74

A Figura 16.58 apresenta o *output* do Stata gerado ao digitarmos o comando **desc**.

```
. desc

  obs:         1,062
 vars:             5
 size:        11,682
-------------------------------------------------------------------------------
              storage   display    value
variable name   type    format     label      variable label
-------------------------------------------------------------------------------
estado          str2    %2s                   estado k (nivel 3)
município       int     %8.0g                 município j (nível 2)
distrito        int     %8.0g                 distrito municipal i (nível 1)
acidentes       byte    %8.0g                 quantidade de acidentes de trânsito no
                                                distrito no último ano
alcool          float   %9.2f                 quantidade média de álcool ingerida por
                                                habitante/dia no distrito (em gramas)
-------------------------------------------------------------------------------
Sorted by:
```

Figura 16.58 Descrição do banco de dados **Acidentes de Trânsito.dta**.

Seguindo a lógica apresentada no Capítulo 14, vamos inicialmente elaborar o histograma da variável *acidentes*, que será a variável dependente do modelo a ser proposto. Para tanto, devemos digitar o seguinte comando, que gera o histograma da Figura 16.59.

```
hist acidentes, discrete freq
```

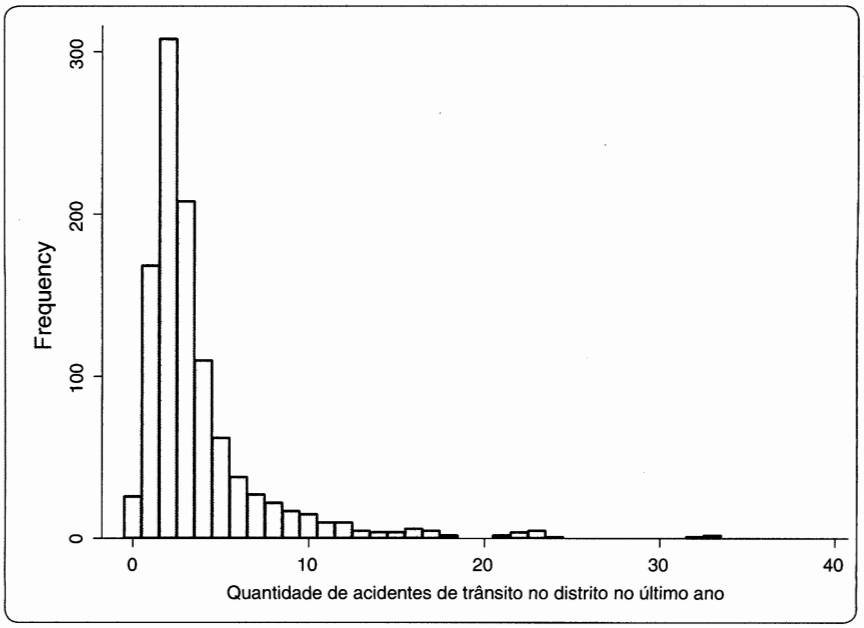


Figura 16.59 Histograma da variável dependente *acidentes*.

Conforme estudamos no Capítulo 14, é interessante que o pesquisador avalie se a média e a variância da variável dependente são iguais, ou ao menos próximas, antes da elaboração de qualquer estimação que envolva dados de contagem, a fim de que seja possível ter uma ideia acerca da adequação da estimação do modelo Poisson ou se será necessária a estimação de um modelo binomial negativo. A digitação do seguinte comando permitirá que este diagnóstico preliminar seja elaborado, cujos resultados são apresentados na Figura 16.60:

```
tabstat acidentes, stats(mean var)
```

```
. tabstat acidentes, stats(mean var)

    variable |      mean  variance
-------------+--------------------
   acidentes |  3.812618  15.24007
----------------------------------
```

Figura 16.60 Média e variância da variável dependente *acidentes*.

Mesmo que a variância da variável *acidentes* seja bem maior do que sua média, o que indica a existência de **superdispersão nos dados**, vamos inicialmente, para fins didáticos, estimar um modelo Poisson. Na modelagem da quantidade de acidentes de trânsito, embora uma possibilidade seja a inclusão, no componente de efeitos fixos, de variáveis *dummy* que representem municípios e estados, vamos tratá-los como efeitos aleatórios e estimar um **modelo de regressão multinível do tipo Poisson** com três níveis e interceptos aleatórios. Além disso, a definição da existência de superdispersão nos dados, que indica uma melhor adequação do **modelo de regressão multinível do tipo binomial negativo** em relação ao modelo Poisson, será elaborada na sequência, por meio de um teste de razão de verossimilhança.

Vamos, portanto, partir para a seguinte estimação:

$$\ln\left(acidentes_{ijk}\right) = \pi_{0jk} + \pi_{1jk}.alcool_{jk}$$
$$\pi_{0jk} = b_{00k} + r_{0jk}$$
$$\pi_{1jk} = b_{10k}$$
$$b_{00k} = \gamma_{000} + u_{00k}$$
$$b_{10k} = \gamma_{100}$$

que resulta no modelo com interceptos aleatórios:

$$\ln\left(acidentes_{ijk}\right) = \gamma_{000} + \gamma_{100}.alcool_{jk} + u_{00k} + r_{0jk}$$

em que a variável *acidentes* representa o fenômeno em estudo, apresentando-se na forma quantitativa e apenas com valores não negativos e discretos (dados de contagem), indicando a incidência de acidentes de trânsito no último ano no distrito municipal *i* localizado no município *j* do estado *k*.

Para a estimação no Stata do modelo proposto, devemos digitar o seguinte comando:

```
mepoisson acidentes alcool || estado: || município: , nolog
```[10]

em que a lógica de inserção dos diferentes níveis obedece ao mesmo critério de aninhamento discutido ao longo do capítulo, ou seja, do maior para o menor nível, sendo os níveis separados pelos termos `||`. Os *outputs* gerados são apresentados na Figura 16.61.

[10] Para versões anteriores à versão 13 do Stata, o comando deverá ser `xtmepoisson acidentes alcool || estado: || município: , var nolog`.

```
. mepoisson acidentes alcool || estado: || município: , nolog

Mixed-effects Poisson regression                Number of obs      =      1062

------------------------------------------------------------
                |   No. of       Observations per Group
 Group Variable |   Groups    Minimum    Average    Maximum
----------------+-------------------------------------------
         estado |       27          1       39.3         95
      município |      235          1        4.5         13
------------------------------------------------------------

Integration method: mvaghermite                 Integration points =       7

                                                Wald chi2(1)       =    5.60
Log likelihood = -2295.9047                     Prob > chi2        =  0.0180
------------------------------------------------------------------------------
       acidentes |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-----------------+------------------------------------------------------------
         alcool |   .0478279    .020216     2.37   0.018     .0082053    .0874506
          _cons |   .7293659   .2638594     2.76   0.006     .2122111    1.246521
-----------------+------------------------------------------------------------
estado           |
       var(_cons)|   .3857761     .12319                      .2063103    .7213563
-----------------+------------------------------------------------------------
estado>município |
       var(_cons)|   .0829691   .0142976                       .059188    .1163053
------------------------------------------------------------------------------
LR test vs. Poisson regression:      chi2(2) =  1279.65   Prob > chi2 = 0.0000

Note: LR test is conservative and provided only for reference.
```

Figura 16.61 *Outputs* do modelo hierárquico Poisson com interceptos aleatórios no Stata.

Com base nesta figura, podemos verificar inicialmente a existência de uma estrutura desequilibrada de dados agrupados em três níveis. Além disso, o resultado do teste de razão de verossimilhança mostra que existe variabilidade significativa entre distritos localizados em diferentes municípios e estados, o que acaba por favorecer o uso do modelo multinível Poisson em relação a um modelo tradicional de regressão Poisson sem efeitos aleatórios.

Antes de prosseguirmos, podemos digitar o comando **estimates store mepoisson**, que faz com que os resultados desta estimação sejam arquivados para posterior comparação com os que serão obtidos pela estimação do modelo binomial negativo. Além disso, também podemos digitar **predict lambda**, que gera uma nova variável no banco de dados (*lambda*) correspondente aos valores estimados de incidência de acidentes de trânsito no último ano em cada um dos 1.062 distritos municipais. Por fim, o pesquisador ainda pode digitar o termo **irr** (*incidence rate ratio*) ao final do comando apresentado, conforme estudamos no Capítulo 14, a fim de que sejam estimadas as taxas de incidência de acidentes de trânsito por ano correspondentes à alteração em cada parâmetro do componente de efeitos fixos.

Um pesquisador ainda mais curioso poderá verificar que as estimações dos parâmetros dos componentes de efeitos fixos e aleatórios são idênticas às que seriam obtidas por meio do seguinte comando:

```
meglm acidentes alcool || estado: || município: , family(poisson)
link(log) nolog
```

que explicita, para o modelo linear generalizado multinível (termo **meglm**), que a distribuição considerada da variável dependente é a Poisson e a função de ligação canônica é a logarítmica.

É possível que, após a estimação dos parâmetros do componente de efeitos aleatórios, as contagens de acidentes de trânsito apresentem superdispersão. Neste sentido, devemos reexaminar os dados estimando um modelo binomial negativo, a fim de que seus resultados possam ser comparados com os obtidos pela estimação do modelo Poisson. Para tanto, devemos digitar o seguinte comando:

```
menbreg acidentes alcool || estado: || município: , nolog
```
[11]

Os resultados obtidos são apresentados na Figura 16.62.

[11] A estimação de modelos multinível do tipo binomial negativo (comando **menbreg**) passou a estar disponível no Stata a partir da versão 13.

```
. menbreg acidentes alcool || estado: || município: , nolog

Mixed-effects nbinomial regression              Number of obs      =      1062
Overdispersion:                 mean

------------------------------------------------------------
                |   No. of       Observations per Group
 Group Variable |   Groups    Minimum    Average    Maximum
----------------+-------------------------------------------
         estado |       27          1       39.3         95
      município |      235          1        4.5         13
------------------------------------------------------------

Integration method: mvaghermite                 Integration points =       7

                                                Wald chi2(1)       =    4.38
Log likelihood = -2234.3721                     Prob > chi2        =  0.0363
------------------------------------------------------------------------------
       acidentes |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-----------------+------------------------------------------------------------
          alcool |   .0466768   .0222975     2.09   0.036     .0029746    .0903791
           _cons |   .7538477   .2843403     2.65   0.008      .196551    1.311144
-----------------+------------------------------------------------------------
         /lnalpha |  -2.258241   .1355339   -16.66   0.000    -2.523883    -1.9926
-----------------+------------------------------------------------------------
estado           |
       var(_cons)|   .3775391   .1205934                      .2018698    .7060775
-----------------+------------------------------------------------------------
estado>município |
       var(_cons)|   .0613878   .0138809                      .0394104    .0956212
------------------------------------------------------------------------------
LR test vs. nbinomial regression:     chi2(2) =   508.99   Prob > chi2 = 0.0000

Note: LR test is conservative and provided only for reference.
```

Figura 16.62 *Outputs* do modelo hierárquico binomial negativo com interceptos aleatórios no Stata.

Na parte inferior desta figura, podemos verificar, pelo resultado do teste de razão de verossimilhança, que a estimação deste modelo multinível é mais adequada do que a estimação de um modelo tradicional de regressão binomial negativo sem efeitos aleatórios para os dados do nosso exemplo. Além disso, todos os parâmetros dos componentes de efeitos fixos e aleatórios são estatisticamente diferentes de zero, ao nível de significância de 5%.

A estimação das variâncias de u_{00k} e r_{0jk} apresentaram valores menores do que os respectivos valores obtidos quando da estimação do modelo multinível Poisson (de 0,386 para 0,377 para u_{00k} e de 0,083 para 0,061 para r_{0jk}), fato que se justifica pela adição de um parâmetro de superdispersão que controla a variabilidade dos dados.

Na Figura 16.62, podemos verificar que é apresentada a estimação de **lnalpha**. Lembremos, conforme estudamos no Capítulo 14, que *alpha* (ou ϕ), que é a superdispersão condicional dos dados, representa o inverso do parâmetro de forma da distribuição binomial negativa. Para os dados do nosso exemplo, temos que $\hat{alpha} = e^{-2,258} = 0,105$.

Analogamente, os parâmetros dos componentes de efeitos fixos e aleatórios também podem ser obtidos por meio do seguinte comando:

```
meglm acidentes alcool || estado: || município: , family(nbinomial)
link(log) nolog
```

A fim de compararmos as estimações dos modelos multinível dos tipos Poisson e binomial negativo, devemos elaborar um teste de razão de verossimilhança, digitando o seguinte comando:

```
lrtest mepoisson ., force
```

em que o termo **mepoisson** refere-se à estimação do modelo Poisson. Como estamos comparando dois diferentes estimadores (**mepoisson e menbreg**), devemos utilizar o termo **force** quando da elaboração deste teste de razão de verossimilhança. O resultado do teste é apresentado na Figura 16.63 e, por meio do qual, podemos verificar que o modelo binomial negativo é mais adequado, comprovando a existência de superdispersão nos dados.

```
. lrtest mepoisson ., force

Likelihood-ratio test                                 LR chi2(1)  =    123.07
(Assumption: mepoisson nested in .)                   Prob > chi2 =    0.0000

Note: The reported degrees of freedom assumes the null hypothesis is not on the boundary
of the parameter space.  If this is not true, then the reported test is conservative.
```

Figura 16.63 Teste de razão de verossimilhança para verificação da adequação do modelo hierárquico binomial negativo.

Portanto, a expressão da quantidade média estimada de acidentes de trânsito por ano, para determinado distrito municipal *i* em determinado município *j* num estado *k*, é dada por:

$$u_{ijk} = e^{\left(0{,}754+0{,}047.alcool_{jk}+u_{00k}+r_{0jk}\right)}$$

em que *u* representa o número esperado de ocorrências ou a taxa média estimada de incidência de acidentes de trânsito para a exposição de um ano. A fim de que essas quantidades estimadas sejam geradas no banco de dados (nova variável *u*), podemos digitar o seguinte comando:

```
predict u
```

Além disso, também podemos obter os termos de erro u_{00k} (invariantes para distritos localizados em um mesmo estado) e r_{0jk} (invariantes para distritos localizados no mesmo município). Para tanto, devemos digitar o seguinte comando:

```
predict u00 r0, remeans
```

que faz com que duas novas variáveis, *u00* e *r0*, também sejam geradas no banco de dados.

O comando a seguir, que gera os *outputs* da Figura 16.64, mostra os valores de *u*, *u00* e *r0* apenas para os distritos dos municípios de Mato Grosso:

```
list estado município acidentes u u00 r0 if estado=="MT",
sepby(município)
```

```
. list estado município acidentes u u00 r0 if estado=="MT", sepby(município)

     +-------------------------------------------------------------+
     | estado   municí~o   aciden~s          u        u00         r0 |
     |-------------------------------------------------------------|
669. |     MT        150          2   1.600369   -.815816   -.0064477 |
670. |     MT        150          2    1.63053   -.815816   -.0064477 |
671. |     MT        150          1    1.63053   -.815816   -.0064477 |
672. |     MT        150          1   1.585499   -.815816   -.0064477 |
673. |     MT        150          2   1.499133   -.815816   -.0064477 |
     |-------------------------------------------------------------|
674. |     MT        151          0   1.415119   -.815816   -.1107979 |
675. |     MT        151          3   1.441788   -.815816   -.1107979 |
676. |     MT        151          1   1.428391   -.815816   -.1107979 |
677. |     MT        151          1   1.441788   -.815816   -.1107979 |
678. |     MT        151          1   1.338034   -.815816   -.1107979 |
679. |     MT        151          1   1.388943   -.815816   -.1107979 |
680. |     MT        151          2   1.415119   -.815816   -.1107979 |
681. |     MT        151          1   1.350584   -.815816   -.1107979 |
682. |     MT        151          1   1.350584   -.815816   -.1107979 |
683. |     MT        151          2    1.40197   -.815816   -.1107979 |
684. |     MT        151          1   1.376037   -.815816   -.1107979 |
685. |     MT        151          1   1.441788   -.815816   -.1107979 |
     |-------------------------------------------------------------|
686. |     MT        152          2   1.667662   -.815816      .01607 |
687. |     MT        152          2   1.576821   -.815816      .01607 |
688. |     MT        152          1   1.621606   -.815816      .01607 |
689. |     MT        152          2   1.547654   -.815816      .01607 |
690. |     MT        152          1   1.547654   -.815816      .01607 |
691. |     MT        152          2   1.533273   -.815816      .01607 |
     |-------------------------------------------------------------|
692. |     MT        153          1   1.462078   -.815816    -.031476 |
693. |     MT        153          2   1.489632   -.815816    -.031476 |
694. |     MT        153          1   1.517706   -.815816    -.031476 |
     +-------------------------------------------------------------+
```

Figura 16.64 Quantidades reais e estimadas de acidentes de trânsito e termos de erro u_{00k} e r_{0jk} para distritos municipais em Mato Grosso (*k* = Mato Grosso).

Por meio desta figura, podemos verificar que, enquanto os valores de *u00* são invariantes para todos os distritos municipais de Mato Grosso, os valores de *r0* são invariantes por município.

Apenas para fins didáticos, o pesquisador poderá verificar que a variável *u* também pode ser gerada por meio da seguinte expressão:

```
gen u = exp(0.7538477 + 0.0466768*alcool + u00 + r0)
```

Por fim, podemos elaborar um gráfico que compara os ajustes das estimações dos modelos tradicional e multinível do tipo binomial negativo. Este gráfico, apresentado na Figura 16.65, é obtido por meio da digitação dos seguintes comandos:

```
quietly nbreg acidentes alcool
predict utrad
graph twoway scatter acidentes alcool || mspline utrad alcool ||
mspline u alcool ||, legend(label(2 "Binomial Negativo Tradicional")
label(3 "Binomial Negativo Multinível"))
```

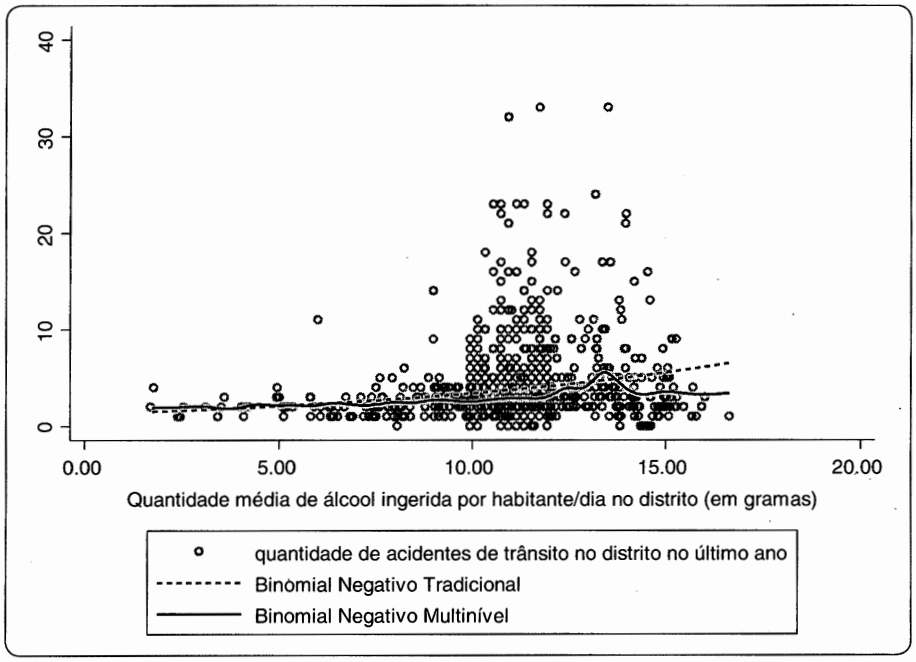


Figura 16.65 Ajustes das quantidades estimadas de acidentes de trânsito pelos modelos tradicional e multinível do tipo binomial negativo, em função da quantidade média de álcool ingerida por habitante/dia no distrito.