MODELO LINEAR HIERÁRQUICO: UM MÉTODO ALTERNATIVO PARA ANÁLISE DE DESEMPENHO ESCOLAR

SANDRA MARIA CONCEIÇÃO PINHEIRO

Orientadora: Profª Drª Maria Cristina Falcão Raposo

Co-orientadora: Profª Drª Claudia Regina Oliveira de Paiva Lima

Área de Concentração: Estatística Aplicada

Dissertação apresentada ao Departamento de Estatística da Universidade Federal de Pernambuco para obtenção do grau de Mestre em Estatística

Recife, janeiro de 2005

Universidade Federal de Pernambuco
Mestrado em Estatística

20 de janeiro de 2005
(data)

Nós recomendamos que a dissertação de mestrado de autoria de

Sandra Maria Conceição Pinheiro

intitulada

Modelo linear hierárquico: um método alternativo para análise de desempenho escolar

seja aceita como cumprimento parcial dos requerimentos para o grau de Mestre em Estatística.

Coordenador da Pós-Graduação em Estatística

Banca Examinadora:

Prof. Francisco Cribari Neto
Coordenador de Mestrado
em Estatística da UFPE

Maria Cristina Falcão Raposo
Maria Cristina Falcão Raposo — orientador

Dalton Francisco de Andrade (UFSC)

Mozart Neves Ramos

Este documento será anexado à versão final da dissertação.

Dedico este trabalho a Deus e aos meus
pais, Roque e Maria das Neves (em memória).

AGRADECIMENTOS

Ao verdadeiro Deus, eterno e imortal, por mais esta vitória e por me manter firme em todas as etapas da minha vida.

Aos meus pais, Roque e Maria das Neves (em memória), pelo incentivo, pelo amor e total apoio em todos os momentos da minha vida.

Aos meus irmãos (Rosângela, Júlio, Roque Filho, Carlos Eduardo, Luís Carlos, Virgínia) e a todos os queridos sobrinhos pelo carinho, incentivo e apoio.

Às minhas orientadoras, Maria Cristina e Claudia Regina, pela confiança, dedicação, paciência e competência durante todo o período de desenvolvimento deste trabalho.

À Drª Ana Marlúcia (ENUFBA) por ter me ajudado a crescer profissionalmente, pela força, confiança e apoio durante todo período do mestrado.

Aos professores do curso de mestrado em estatística da UFPE pela oportunidade de aprimorar meus conhecimentos, em especial à Viviana Giampaoli pelas sugestões dadas.

Aos professores do curso de graduação da UFBA, em especial à professora Deborah Medeiros pelo incentivo.

Aos meus amigos da turma do mestrado: Andréa (pelo carinho e confiança), André (pela simpatia e atenção), Gecynalda (pela força, incentivo e amizade), Júnior (pelos momentos de descontração e alegria), Lenaldo (pela simplicidade e cuidado), Renata (pela atenção, carinho e sinceridade), Sandra Rêgo (pelo cuidado, modéstia e gentileza) e Tatiane (pela simplicidade e contagiante alegria). Agradeço a todos pela ajuda, apoio e pelos bons momentos de convivência durante este período.

À família Soares Gomes (D. Geci, Gecynete, Gecynalda, Ariane e Larissa) por terem me acolhido como um membro da família com tanto carinho, no primeiro ano do mestrado e por ampliarem meu leque de amizade.

Às minhas amigas da ENUFBA (Ana, Conceição, Mônica, Nadja, Nedja, Valterlinda e Lucivalda) e aos bolsistas pelo carinho e atenção.

A Gilson, Keila, Moisés, Patrícia Leal, Silvia Patrícia, e Tatiene pela atenção e carinho.

A turma do mestrado de 2004 (Artur, Camilo, Carlos Gadelha, Carlos Tomé, Daniela, Denis, Francisco, Hernando, Kátia, Milena, Polyane, Themis e Tiago), pelos poucos mas agradáveis momentos.

À Valéria Bittencourt pela competência e eficiência.

Aos participantes da banca examinadora, pelas sugestões.

À Secretaria de Educação do Estado de Pernambuco por gentilmente ceder os dados do SAEPE (2002).

A todos que de alguma forma contribuiram para a execução deste trabalho.

Resumo

Os modelos multiníveis ou modelos lineares hierárquicos foram desenvolvidos para análise de dados que possuem uma estrutura de grupo, ou seja, uma estrutura de hierarquia, por levarem em consideração a dependência dos dados existente dentro de cada nível hierárquico e entre os níveis hierárquicos. As estimativas dos parâmetros dos modelos hierárquicos são apresentadas separando os efeitos fixos dos aleatórios. Através da análise dos dados do SAEPE, fornecidos pela Secretaria de Educação e Cultura do Estado de Pernambuco são apresentados modelos hierárquicos com dois níveis para avaliação do desempenho em matemática e português dos alunos da $4^{\underline{a}}$ e $8^{\underline{a}}$ série do ensino fundamental e para os alunos da $3^{\underline{a}}$ série do ensino médio. O desempenho desses alunos foi avaliado através das notas obtidas pelos mesmos nas duas disciplinas utilizando os programas HLM e R. Os resultados mostraram que a utilização deste tipo de modelagem é o mais apropriado, quando se possui dados com estrutura de grupos. Uma comparação dos coeficientes de determinação (R^2) dos modelos de regressão múltipla e dos modelos hierárquicos mostra o desempenho superior dos modelos hierárquicos.

Palavras-chave: Modelo linear hierárquico, níveis, efeitos fixos, efeitos aleatórios e desempenho escolar.

Abstract

The multilevels models or hierarchical linear models were developed for analysis of data that possess a group structure, in other words, a structure of hierarchy, for they take inside into account the variability of the existent data of each hierarchical level and among the hierarchical levels. The estimates of the parameters of the hierarchical models are presented separating the fixed effects from the random effects. Through the analysis of the data of SAEPE supplied by the General office of Education and Culture of the State of Pernambuco, hierarchical models are presented with two levels evaluation of the acting in mathematics and portuguese the students' of the 4th and 8th series of the fundamental teaching and for the students of the 3rd series of the medium teaching. Those students' acting was evaluated through the grades obtained by the same ones in the two disciplines using the programs HLM and R. The results showed that the use of this modelling type it is the most appropriate, when it is possessed data with structure of groups. A comparison of the determination coefficients (R^2) of the models of multiple regression and of the hierarchical models it shows the superior acting of the hierarchical models.

Keywords: hierarchical lineal models, level, fixed effects, random effects, school acting.

SUMÁRIO

Capítulo 1

Introdução

Vários tipos de dados observacionais, incluindo aqueles coletados em ciências humanas, biológicas bem como em outras áreas do conhecimento, possuem uma estrutura de grupo. Por exemplo, estudos animais e humanos têm uma ordem natural dada pela relação de herança, onde a descendência se agrupa dentro de famílias. Descendentes tendem a ser mais semelhantes em suas características físicas e mentais que indivíduos escolhidos de forma aleatória em uma grande população. Crianças de uma mesma família podem ter tendência a ter baixa estatura, talvez porque seus pais tenham baixa estatura ou porque vivam em um ambiente comum caracterizado pela baixa renda (Goldstein, 1999).

Os indivíduos interagem com o contexto social no qual eles vivem significando que são influenciados pelos grupos sociais aos quais eles pertencem, e por sua vez esses indivíduos imprimem características e significados ao grupo. Dados com esta estrutura, onde indivíduos e grupos em estudo estão agrupados em unidades de diferentes níveis de hierarquia devem ser analisados usando modelo multinível. Nesta condição geralmente os indivíduos e os grupos sociais são conceitualizados como um sistema hierárquico de indivíduos e grupos, incorporando diferentes níveis hierárquicos, e como resultado pode ter variáveis definidas em cada nível.

Os modelos multiníveis foram desenvolvidos para analisar dados que possuem uma estrutura hierárquica, eles levam em conta a variabilidade dos dados existente dentro de cada nível hierárquico e entre os níveis hierárquicos.

Em muitas investigações, em especial aquelas da área de saúde, o modelo multinível ou hierárquico recebe enfoque diferente onde a hierarquia é definida com base na ordem de importância das variáveis no estudo. Este tipo de enfoque pode ser visto em Abbad e Torres (2002) que afirmam que a regressão hierárquica é utilizada em estudos confirmatórios, uma vez que esse tipo de análise busca a explicação sobre relacionamento entre variáveis descrita em modelos teóricos consistentes, ou seja, em modelos que apresentam um conjunto de proposições empíricas que já indicam a magnitude e direção da relação entre variáveis. Neste caso, a ordem de entrada dos preditores na equação de regressão é definida pelo pesquisador, que baseia sua decisão em teorias ou resultados de outras pesquisas de consenso no mundo acadêmico (Torres e Curtin, 1999).

1.1. Exemplos de Aplicação

Como já referido, os modelos hierárquicos podem ser aplicados em várias áreas do conhecimento cujos dados possuam uma estrutura de grupo. Uma das áreas onde este tipo de modelagem é bastante utilizada é a área de educação. Vários autores como Aitkin *et al* (1981) analisaram dados na área de educação, onde os alunos compõem o nível do indivíduo e as classes escolares, onde os alunos estão agrupados, formam o nível do grupo. Ferrão *et al* (2002) e Aitkin e Longford (1986) também utilizaram o modelo hierárquico quando analisaram dados de avaliação educacional considerando aluno como a unidade do nível do indivíduo e escola como unidade do nível do grupo.

Conforme referido por Bryk e Raudenbush (1992, 2002) em pesquisa social, Mason *et al* (1983) examinaram em 15 países os efeitos que a educação materna e a localização da residência, área rural ou urbana, exercem sobre a fertilidade. Sabe-se que em muitos países, altos níveis de educação e residências em áreas urbanas predizem baixa fertilidade. Os pesquisadores provaram que tais efeitos podem depender de características dos países como o nível de desenvolvimento econômico nacional, indicado pelo produto interno bruto, e também a intensidade do planejamento familiar. Eles encontraram que altos níveis de educação materna realmente estavam associados com baixas taxas de fertilidade nos países pesquisados. Verificaram também que, diferenças de taxas de fertilidade em áreas rural e urbana eram maiores em países com um alto produto interno bruto e baixa organização de planejamento familiar.

A utilização dos modelos hierárquicos tem crescido bastante nos últimos anos na área de saúde, este crescimento se deve ao interesse em determinar variáveis que, ao nível do grupo e/ou do indivíduo, influenciam potencialmente no desenvolvimento de determinada doença (Diez-Roux, 2000). Por exemplo, pesquisadores da área de saúde podem estar interessados em avaliar, em uma determinada empresa, se a idade, que pode ser uma característica do nível do indivíduo, e o setor de trabalho, que pode ser uma característica do nível do grupo, influenciam nos níveis pressóricos dos mesmos.

Em Kreft e Leeuw (1998) encontra-se um exemplo de dados de trabalhadores que foram coletados em 12 indústrias diferentes. As variáveis ao nível do indivíduo foram: o nível educacional (como variável explicativa) e a renda do indivíduo (como variável resposta). O tipo de indústria, bem como a distinção entre indústria pública e privada, foram variáveis do nível do grupo. A análise dos dados ao nível do indivíduo (trabalhador) mostrou uma relação positiva entre o nível de educação e a renda: o nível de educação mais elevado leva à uma renda pessoal mais alta. A análise executada ao nível do grupo (indústria), com as 12 indústrias como observações, mostrou surpreendentemente um resultado oposto. Uma relação negativa foi encontrada entre educação e renda. A mais alta média do nível educacional dos fun-

cionários de uma indústria está relacionada a mais baixa média de renda dos trabalhadores da mesma.

Também na área criminalista os modelos multiníveis são usados como mostra Tseloni (1999), que avalia o número de crimes pessoais (raptos, agressão sexual, ameaças, roubos, assaltos, etc) cometidos por cada membro das residências pesquisadas durante um período de 18 meses. Neste caso, os membros das residências são as unidades do nível do indivíduo e as residências são as unidades do nível do grupo.

Várias outras aplicações podem ser encontradas em áreas como demografia, sociologia, economia, dentre outras.

1.2. Revisão da Literatura

De acordo com Bryk e Raudenbush (1992) o termo modelo linear hierárquico foi introduzido por Lindley e Smith (1972) como parte de suas contribuições em estimação bayesiana de modelos lineares, e dentro desse contexto eles elaboraram um suporte geral para dados agrupados com estrutura de erro complexa. Mas, esta contribuição não foi devidamente aproveitada por um período de tempo, em decorrência do mau uso dos métodos de estimação apropriados para estimar os componentes de covariância, na presença de dados não balanceados.

Os modelos multiníveis estão também relacionados a um estudo feito por Bennett (1976) na Inglaterra, que observou que as crianças do ensino fundamental expostas à uma maneira formal de aprendizado da leitura exibiam um aprendizado maior do que as não expostas. Os dados foram analisados utilizando técnicas comuns de regressão múltipla, ignorando os agrupamentos para professores e classes escolares, e considerando as crianças como sendo as unidades de análise.

Como relata Goldstein (1995, 2003) os resultados, estatisticamente significantes identificados por Bennett, contribuiram para outras investigações com estrutura hierárquica e, subsequëntemente, Aitkin *et al* (1981) demonstraram que, ao conduzir a análise com as crianças agrupadas em classes escolares, as diferenças significantes desapareciam, e as crianças que se submeteram ao ensino formal não apresentaram qualquer diferença em relação às demais.

A grande maioria da modelagem realizada até o final da década de 80 não levava em consideração como os dados estavam estruturados e, portanto, as análises apropriadas para os dados. Em parte, isto devia-se à falta de métodos e softwares que viabilizassem a generalização da abordagem multinível ou hierárquica. Assim, o pesquisador tinha que definir uma unidade sobre a qual o seu estudo iria ser aplicado.

A partir de 1986, com o artigo de Aitkin e Longford foi iniciada uma série de procedimentos, que resultaram nas idéias centrais dos modelos hierárquicos. Os modelos lineares hierárquicos com 2 e 3 níveis para aplicações específicas a dados educacionais e experimentos com medidas repetidas, foram discutidos por Bryk e Raudenbush (1992). Uma orientação mais teórica dos modelos hierárquicos incluindo a discussão de um modelo hierárquico para análise fatorial, modelos com respostas categorizadas e outros modelos multivariados é fornecida por Longford (1993). Até meados da década de 90, o desenvolvimento desses modelos voltou-se para as áreas de dados com respostas discretas, modelos de séries temporais, classificações cruzadas, dados perdidos e modelos não lineares, com aplicação em diversas áreas do conhecimento (Goldstein, 1995).

De acordo com Bergamo (2002) a aplicação dos modelos tradicionais de regressão pressupõe a independência entre os indivíduos. No entanto, a maioria dos eventos resulta de fatores que sofrem influências de diferentes níveis, neste sentido os dados são estruturados em hierarquias, e as unidades de um mesmo nível pertencem a uma unidade de nível mais alto, deste modo raramente são independentes porque compartilham de um mesmo ambiente ou apresentam características semelhantes. Desta forma, a suposição de independência é violada passando a existir correlação entre essas unidades. Assim, os modelos multiníveis atendem adequadamente a este nível de complexidade por considerarem todas as correlações existentes entre as observações nos diferentes níveis de hierarquia.

Ainda sob a ótica de Bergamo (2002), nos modelos hierárquicos, as unidades de cada nível são vistas como tendo um efeito aleatório, ou seja, são amostras aleatórias de uma população dessas unidades. Estes efeitos aleatórios tornam os coeficientes aleatórios por levar em conta a variabilidade entre essas unidades, seja de forma simples, através da variabilidade apenas nos interceptos, ou de forma mais complexa através de variabilidade também nas inclinações. Além disso, os modelos multiníveis com efeitos aleatórios possuem a vantagem de acomodarem a estrutura hierárquica aos dados quando comparados com os modelos tradicionais.

A idéia de separar a análise de regressão dentro de cada grupo, seguido pelo ajuste de modelos para os coeficientes de regressão do nível do indivíduo adicionando variáveis explicativas para o nível do grupo, não é suficiente para especificar um modelo hierárquico. É essencial perceber que modelo hierárquico envolve uma integração estatística dos diferentes modelos especificados nos níveis de interesse. A simples integração leva a acrescentar no modelo coeficientes aleatórios, onde os coeficientes de regressão do nível do indivíduo são tratados como variáveis aleatórias no nível do grupo, significando que o coeficiente de regressão do nível do indivíduo é visto como originário de uma distribuição de probabilidade. Os parâmetros mais importantes desta distribuição, a média e a variância, estão entre o conjunto dos parâmetros que serão estimados no modelo hierárquico. O modelo torna-se

mais geral quando são adicionadas variáveis explicativas no nível 2 (do grupo) (Kreft e Leeuw, 1998).

Quando variáveis de níveis diferentes são analisadas como variáveis de um único nível, originam dois diferentes tipos de problemas. O primeiro tipo é estatístico. Quando os dados de um nível mais baixo como, por exemplo, alunos de uma escola, são agrupados em unidades de um nível mais elevado como, por exemplo, as classes escolares, resulta em perda de informação em decorrência desse agrupamento. Por outro lado, se os dados são desagrupados, ou seja, as informações de um nível mais elevado são desagregadas para um nível mais baixo, um nível individual; então há uma repetição da mesma informação entre os indivíduos pertencentes aquele mesmo grupo. Os testes estatísticos tradicionais tratam todas estas informações como se fossem independentes. Quando o tamanho de amostra tem o mesmo número de informações que as unidades do nível individual, observa-se que os testes de significância rejeitam com muito mais freqüência a hipótese nula quando comparados ao nível de significância normalmente sugerido.

O outro tipo de problema encontrado é conceitual, que pode ser de dois tipos diferentes. Se o analista não for cuidadoso na interpretação dos resultados pode cometer o erro do nível incorreto, que consiste em analisar os dados de um nível, e emitir as conclusões com base em outro nível. Provavelmente o erro mais conhecido é o erro ecológico, que é a interpretação de dados agregados, em nível individual. Conclusões de análises produzidas em um nível mais baixo, que são retiradas em um nível mais alto, também são errôneas e este erro é conhecido como erro atomístico. Destaca-se ainda um outro tipo de erro conhecido como paradoxo de Simpson, que expressa que conclusões completamente erradas podem ser retiradas se os dados forem agrupados, retirados de populações heterogêneas, e forem apresentados e analisados como se fossem originários de uma população homogênea (Hox, 1995).

Como o problema multinível está ligado a estrutura de um conjunto de dados que é composta por vários níveis hierárquicos, a questão a ser respondida é como os vários indivíduos e grupos de variáveis influenciam uma única variável resposta. Tipicamente, pode-se obter algumas variáveis explicativas do nível mais alto agregando informações das variáveis individuais de nível mais baixo. O objetivo da análise é determinar o efeito direto das variáveis relacionadas aos indivíduos e variáveis explicativas a nível de grupo sobre o evento estudado, e determinar se as variáveis explicativas do nível do grupo se adequam como moderadoras das relações do nível individual. Se as variáveis do nível de grupo moderam a relação do nível mais baixo, está indicando o aumento da interação estatística entre variáveis explicativas de níveis diferentes. No passado, tais dados eram comumente analisados usando análise de regressão múltipla convencional, com uma variável dependente no nível (individual) mais baixo e uma coleção de variáveis explicativas de todos os níveis disponíveis

(Roberts e Burstein, 1980; Van den Eeden e Hüttner, 1982). Neste sentido, esta análise inclui todos os dados disponíveis em um único nível incorrendo nos problemas conceituais e estatísticos mencionados acima. Muitas pesquisas foram direcionadas na busca de métodos de análise mais apropriados para este modelo de regressão hierárquico, e assim resolver os problemas conceituais e estatísticos associados.

1.3. Definição dos Objetivos

O objetivo principal da presente dissertação é estudar o modelo de regressão multinível básico, que é o de dois níveis, apresentando os casos particulares onde pode-se considerar variáveis explicativas associadas à resposta no nível do indivíduo (nível 1), e variáveis explicativas no nível do grupo (nível 2).

Os dados do SAEPE (Sistema de Avaliação Educacional de Pernambuco) de 2002 foram cedidos pela Secretaria de Educação e Cultura do Estado de Pernambuco. Com esse sistema de avaliação a Secretaria de Educação e Cultura do Estado pretende desenvolver uma estratégia de monitoramento e de incentivos centrados na melhoria da qualidade e do desempenho do Ensino Básico no Estado. Desta forma, foram aplicadas avaliações de português e matemática aos alunos da $2^{\underline{a}}$, $4^{\underline{a}}$ e da $8^{\underline{a}}$ série do Ensino Fundamental, e da $3^{\underline{a}}$ série do Ensino Médio das redes Municipais e Estadual. Os resultados apresentados neste trabalho são das avaliações aplicadas a 75717 alunos da $4^{\underline{a}}$ e $8^{\underline{a}}$ série do Ensino Fundamental e 28371 da $3^{\underline{a}}$ série do Ensino Médio.

As análises dos dados foram realizadas em dois programas computacionais, o sistema computacional **R**, através de um pacote denominado *nlme*, que foi desenvolvido por Pinheiro e Bates (2000), que provê métodos para ajuste linear (função *lme*) e não linear (função *nlme*) de modelos de efeitos mistos, assumindo que tanto os efeitos aleatórios quanto os erros seguem uma distribuição gaussiana; o HLM, que é um programa que foi desenvolvido especificamente para análise de modelos com estrutura hierárquica, por Bryk *et al* (1996, 2001). Ambos utilizando o método de máxima verossimilhança restrita.

1.4. Resumo do Conteúdo

Os seis capítulos seguintes desta dissertação estão organizados como descrito abaixo.

No capítulo 2 são apresentados alguns submodelos do modelo multinível básico mais completo (dois níveis), que considera a presença de mais de uma variável explicativa tanto ao nível do indivíduo, quanto ao nível do grupo, com interceptos e inclinações como coeficientes aleatórios. Os métodos de estimação dos parâmetros como o de máxima verossimilhança

completo e restrito, o algoritmo EM (esperança e maximização) e o BLUP (melhor preditor linear não tendencioso) são apresentados no capítulo 3 e os intervalos de confiança e testes de hipóteses dos parâmetros, para modelos com efeitos fixos e aleatórios, podem ser encontrados no capítulo 4. Já o capítulo 5 é dedicado às técnicas de diagnósticos, como estimação dos resíduos, verificação das suposições do modelo e verificação dos pontos influentes. A aplicação feita com o banco de dados da Secretaria de Educação do Estado de Pernambuco é apresentada no capítulo 6, e as conclusões encontram-se no capítulo 7.

Capítulo 2

Modelo Linear Hierárquico

Segundo Diez-Roux (2000) o termo análise multinível (ou modelagem hierárquica) tem sido utilizado em campos como educação, demografia e sociologia, para descrever uma abordagem analítica que permite examinar simultaneamente os efeitos que as variáveis do nível do grupo e do nível do indivíduo exercem sobre a variável resposta.

O modelo multinível ou hierárquico é abordado no presente trabalho em sua forma básica, ou seja, com dois níveis de hierarquia mas, existem modelos com mais de dois níveis hierárquicos que devem ser considerados de acordo com a estrutura dos dados trabalhados.

2.1. Modelo de Regressão Linear Simples

No modelo de regressão linear simples, a relação entre a variável resposta e a variável explicativa é escrita como:

$$Y_i = \beta_0 + \beta_1 X_i + e_i, \text{ com } i = 1, \ldots, n,$$

onde

Y_i representa a resposta do i-ésimo indivíduo;

β_0 representa o valor esperado da variável resposta Y_i para X_i igual a zero;

β_1 representa a mudança esperada em Y_i quando X_i aumenta de uma unidade;

X_i é a variável explicativa do i-ésimo indivíduo;

e_i é o erro associado ao i-ésimo indivíduo, com as suposições:
$e_i \sim N(0, \sigma^2)$ e os e_i's independentes, $i = 1, \ldots, n$.

A variabilidade de um experimento pode ser explicada por um conjunto de retas de regressão que podem ser originadas da realização do experimento várias vezes sob as mesmas condições. Quando o objetivo é estudar não apenas um determinado evento, mas tudo o que o envolve, é necessário ajustar um modelo que leve em conta toda a variabilidade entre os experimentos e incorpore os diferentes aspectos de cada um deles. Considerar os dados de acordo com uma estrutura hierárquica leva em conta tal tipo de análise (Bergamo, 2002).

2.2. Modelo Linear Hierárquico de Dois Níveis

O modelo de regressão multinível como é denominado por Hox (1995) e Goldstein (1999), assume uma variedade de nomes tal como modelo linear hierárquico chamado por Bryk e Raudenbush (1992) e permite que diferentes níveis sejam especificados em modelos separados e depois combinados em um único modelo.

Este tipo de modelagem assume que há um conjunto de dados hierárquicos, que possui uma variável resposta (Y) que é medida no nível individual, e variáveis explicativas que podem residir no nível do indivíduo (X) e/ou do grupo (W), que é um nível mais elevado. Conceitualmente o modelo pode ser visto como um sistema hierárquico de equações de regressão.

O modelo de regressão hierárquico definido a seguir é considerado como básico, já que possui apenas dois níveis de hierarquia. Neste modelo o subscrito i denota o i-ésimo indivíduo ($i = 1, \ldots, n_j$) do grupo j ($j = 1, \ldots, J$).

De acordo com Bryk e Raudenbush (1992) o modelo linear hierárquico mais simples possível é aquele onde assume-se que não há variável explicativa em nenhum dos dois níveis. É o modelo com apenas o intercepto aleatório ou análise de variância (ANOVA) com efeitos aleatórios. Este modelo é o primeiro passo na análise de dados hierárquicos e sua equação é dada por:

$$Y_{ij} = \beta_{0j} + e_{ij},\ i = 1, \ldots, n_j,\ j = 1, \ldots, J \tag{2.1}$$

onde cada erro do nível 1, e_{ij}, segue uma distribuição normal com média zero e variância σ^2, e o modelo do nível 2 é:

$$\beta_{0j} = \gamma_{00} + u_{0j},\ j = 1, \ldots, J \tag{2.2}$$

onde

γ_{00} representa a média da variável resposta na população;

u_{0j} é o efeito aleatório associado com o grupo j, $u_{0j} \sim N(0, \tau_{00}),\ j = 1, \ldots, J$, e $u'_{0j}s$ são independentes dos $e'_{ij}s$.

Substituindo a equação (2.2) em (2.1) tem-se o modelo combinado:

$$Y_{ij} = \gamma_{00} + u_{0j} + e_{ij},\ i = 1, \ldots, n_j,\ j = 1, \ldots, J. \tag{2.3}$$

A equação (2.3) é um modelo de efeitos aleatórios porque os efeitos dos grupos (u_{0j}) são interpretados como aleatórios. Este modelo não explica nenhuma variância, apenas a decompõe em duas componentes independentes: σ^2, que é a variância dos erros do nível 1 (do indivíduo), aqui denominado e_{ij}; e τ_{00}, que é a variância dos erros do nível 2 (do grupo),

definidos por u_{0j}. Usando este modelo o coeficiente de correlação intra classe (ρ) pode ser obtido pela equação:

$$\rho = \frac{\tau_{00}}{Var(Y_{ij})}, \tag{2.4}$$

onde

$$\begin{aligned} Var(Y_{ij}) &= Var(\gamma_{00} + u_{0j} + e_{ij}) \\ &= \tau_{00} + \sigma^2. \end{aligned}$$

A correlação intra classe (ρ) mede o grau de homogeneidade dentro do grupo, ou seja, os indivíduos de um mesmo grupo tendem a ser mais parecidos nas características pesquisadas. Portanto, ρ mede a proporção da variação que há entre a resposta e os grupos do nível 2.

Um outro modelo hierárquico que pode ser considerado é quando há uma variável explicativa ao nível do grupo, conhecido como modelo de regressão de médias como respostas. Uma vez que as médias dos J grupos podem variar de acordo com determinada característica de grupo (W), então o modelo para o nível 2 é dado por:

$$\beta_{0j} = \gamma_{00} + \gamma_{01} W_j + u_{0j},\ j = 1, \ldots, J, \tag{2.5}$$

onde

β_{0j} é a variável resposta de um modelo de regressão linear onde as variáveis explicativas correspondem a características do grupo j.

Neste caso temos uma variável explicativa (W) para o nível 2, e substituindo a equação (2.5) na equação (2.1) temos o modelo combinado:

$$Y_{ij} = \gamma_{00} + \gamma_{01} W_j + u_{0j} + e_{ij},\ i = 1, \ldots, n_j,\ j = 1, \ldots, J \tag{2.6}$$

onde

γ_{00} é o intercepto médio dos grupos para W_j igual a zero;

γ_{01} é a diferença média entre os J grupos;

u_{0j} é o efeito aleatório do j-ésimo grupo sobre o intercepto para W_j igual a zero;

e e_{ij} é definido como em (2.1).

No modelo de regressão multinível mais completo pode-se considerar o intercepto (β_{0j}) e o coeficiente de inclinação (β_{1j}) variando por grupo, ou seja, podem ser considerados como coeficientes aleatórios, o qual é conhecido como modelo de regressão de coeficientes aleatórios. Uma regressão para cada grupo separadamente pode ser montada para predizer a variável dependente Y pela variável explicativa X através da equação:

$$Y_{ij} = \beta_{0j} + \beta_{1j} X_{ij} + e_{ij},\ i = 1, \ldots, n_j,\ j = 1, \ldots, J \tag{2.7}$$

tendo os modelos para o nível 2 como:

$$\beta_{0j} = \gamma_{00} + u_{0j}, \tag{2.8a}$$

e

$$\beta_{1j} = \gamma_{10} + u_{1j}, \tag{2.8b}$$

onde

γ_{10} é a inclinação média dos J grupos;

u_{1j} é o efeito do j-ésimo grupo sobre a inclinação, $j = 1, \ldots, J$; e γ_{00} e u_{0j} como definidos em (2.2).

A dispersão dos efeitos aleatórios do nível 2 pode ser apresentada como uma matriz de variância e covariância:

$$Var \begin{bmatrix} u_{0j} \\ u_{1j} \end{bmatrix} = \begin{bmatrix} \tau_{00} & \tau_{01} \\ \tau_{10} & \tau_{11} \end{bmatrix} = T, \tag{2.9}$$

onde

$Var(u_{0j}) = \tau_{00}$ é a variância incondicional nos interceptos do nível 1, $j = 1, \ldots, J$;

$Var(u_{1j}) = \tau_{11}$ é a variância incondicional nas inclinações do nível 1, $j = 1, \ldots, J$;

$Cov(u_{0j}, u_{1j}) = \tau_{01}$ é a covariância incondicional entre interceptos e inclinações do nível 1, $j = 1, \ldots, J$.

Ao substituir as equações (2.8a) e (2.8b) na equação (2.7) tem-se o modelo combinado:

$$Y_{ij} = \gamma_{00} + \gamma_{10}X_{ij} + u_{0j} + u_{1j}X_{ij} + e_{ij},\ i = 1, \ldots, n_j,\ j = 1, \ldots, J. \tag{2.10}$$

Este modelo mostra que a resposta Y_{ij} é função da equação de regressão média $\gamma_{00} + \gamma_{10}X_{ij}$ mais um erro aleatório com os seguintes componentes:

u_{0j} é o efeito aleatório do j-ésimo grupo sobre a média, $j = 1, \ldots, J$;

$u_{1j}X_{ij}$ onde u_{1j} é o efeito aleatório do j-ésimo grupo sobre a inclinação β_{1j}, $j = 1, \ldots, J$;

e e_{ij} que é o erro aleatório do nível 1, $i = 1, \ldots, n_j$ e $j = 1, \ldots, J$.

O modelo de regressão multinível também pode levar em consideração a característica W_j ao nível do grupo, que ajuda a predizer a variabilidade dos dois coeficientes, isto é, do intercepto e da inclinação. Este tipo de modelagem é conhecido como modelo com interceptos e inclinações como respostas. Desta forma, as equações (2.8a) e (2.8b) serão substituidas por:

$$\beta_{0j} = \gamma_{00} + \gamma_{01}W_j + u_{0j}, \tag{2.11a}$$

e

$$\beta_{1j} = \gamma_{10} + \gamma_{11}W_j + u_{1j},\ j = 1, \ldots, J. \tag{2.11b}$$

A equação (2.11a) pressupõe que o intercepto (β_{0j}) possa ser predito pela variável do nível 2, W. Já a equação (2.11b) indica que a relação entre a variável resposta (Y) e a variável explicativa do nível 1 (X) depende da variável explicativa do nível 2 (W).

Os coeficientes γ não variam segundo o grupo, por esta razão são chamados coeficientes fixos. Toda variação encontrada entre os grupos, depois de predizer-se os coeficientes β na presença da variável do nível de grupo (W), é assumida como sendo variação residual que é capturada pelos termos de erro residual u_j.

Substituindo as equações (2.11a) e (2.11b) na equação (2.7) temos o modelo de regressão multinível:

$$Y_{ij} = \gamma_{00}+\gamma_{10}X_{ij}+\gamma_{01}W_j+\gamma_{11}W_jX_{ij}+u_{0j}+u_{1j}X_{ij}+e_{ij},\ i = 1, \ldots, n_j,\ j = 1, \ldots, J. \quad (2.12)$$

O segmento $\gamma_{00} + \gamma_{10}X_{ij} + \gamma_{01}W_j + \gamma_{11}W_jX_{ij}$ define a parte fixa ou determinística do modelo, e o segmento $u_{0j} + u_{1j}X_{ij} + e_{ij}$ contém todos os termos aleatórios, sendo chamado de parte aleatória ou estocástica do modelo.

O termo W_jX_{ij} é um termo de interação que aparece no modelo como conseqüência da variação da inclinação β_{1j} da regressão modelada, considerando a variável ao nível do indivíduo (X) com a variável ao nível do grupo (W). Uma vez que o termo de erro aleatório u_{1j} é multiplicado pela variável explicativa X_{ij}, o erro total resultante será diferente para diferentes valores de X_{ij}, uma situação que em regressão múltipla comum é chamada heteroscedástica. Mas, o modelo de regressão múltipla comum assume homoscedasticidade, significando que todos os erros são independentes de todas as variáveis explicativas, e quando a suposição de que os erros são homoscedásticos não é verdadeira, a análise de regressão múltipla comum não é apropriada (Hox, 1995).

Quando as variáveis que compõem o termo de interação são expressas como desvios de suas respectivas médias tem-se uma interpretação dos coeficientes do modelo mais fácil. Centrar as variáveis explicativas na média amostral global pode ser mais adequado para a interpretação do intercepto da regressão β_{0j}, quando, por exemplo, o valor zero não for adequado para as variáveis explicativas do nível 1 incluídas no modelo.

Outros submodelos decorrentes de mudanças na equação (2.11b) que podem ser mencionados são:

1) A análise de covariância (ANCOVA) com efeitos aleatórios, que é obtido quando considera-se que as inclinações não variam aleatoriamente e não são afetadas pelo efeito de W_j, que é uma característica de grupo. A equação (2.11b) torna-se:

$$\beta_{1j} = \gamma_{10},\ j = 1, \ldots, J.$$

2) O modelo onde as inclinações variam não aleatoriamente, que é considerado quando a variância residual (τ_{11}) é bem próxima de zero. A equação (2.11b) é dada por:

$$\beta_{1j} = \gamma_{10} + \gamma_{11} W_j,\ j = 1, \ldots, J.$$

O método dos mínimos quadrados ordinários não é indicado para estimar os parâmetros do modelo como o (2.12), quando se considera que os tamanhos das amostras diferem entre os grupos. Isto porque, nesse tipo de estrutura, são usados vários tipos diferentes de parâmetros. Especificamente, os coeficientes β_j's no nível 1 podem ser fixos ou aleatórios e no nível 2 , os coeficientes γ's são considerados fixos, as variâncias e covariâncias entre os níveis são chamadas componentes de variância. Assim, para estimar os parâmetros envolvidos no modelo, são necessários processos iterativos (Hox, 1995; Goldstein, 1999).

De um modo geral no modelo há mais de uma variável explicativa no nível 1, do indivíduo, e mais de uma variável explicativa no nível 2, do grupo. Assumindo que existem q variáveis explicativas no nível 1 ($q = 1, \ldots, Q$) e p variáveis explicativas no nível 2 ($p = 1, \ldots, P$) tem-se o modelo:

$$Y_{ij} = \beta_{0j} + \beta_{1j} X_{1ij} + \ldots + \beta_{Qj} X_{Qij} + e_{ij},\ i = 1, \ldots, n_j,\ j = 1, \ldots, J \tag{2.13}$$

onde os parâmetros desconhecidos são dados por:

$$\begin{aligned}
\beta_{0j} &= \gamma_{00} + \gamma_{01} W_{1j} + \ldots + \gamma_{0P} W_{Pj} + u_{0j} \\
\beta_{1j} &= \gamma_{10} + \gamma_{11} W_{1j} + \ldots + \gamma_{1P} W_{Pj} + u_{1j} \\
&\vdots \\
\beta_{Qj} &= \gamma_{Q0} + \gamma_{Q1} W_{1j} + \ldots + \gamma_{QP} W_{Pj} + u_{Qj},\ j = 1, \ldots, J
\end{aligned}$$

Como já mencionado, assume-se que os erros do nível 1 (e_{ij}) são normalmente distribuídos com média zero e variância comum σ^2, em todos os grupos. Os termos de erro do nível 2, u_{qj}, são independentes dos erros e_{ij} e têm distribuição normal multivariada com médias iguais a zero. A variância do erro residual u_{0j} é a variância dos interceptos entre os grupos (τ_{00}) e as variâncias dos erros residuais u_{qj} são variâncias das inclinações entre os grupos especificado como τ_{qq}. As covariâncias entre os termos de erro residual $\tau_{q'q''}$ ($q' \neq q''$) são diferentes de zero.

Capítulo 3

Métodos de Estimação

Em uma análise hierárquica há três tipos de parâmetros que podem ser estimados: os efeitos fixos, os efeitos aleatórios do nível 1 e as componentes de variância e covariância (Bryk e Raudenbush, 1992).

Existem métodos de estimação como o método de máxima verossimilhança, que pode ser completo ou restrito, o algoritmo EM e o melhor preditor linear não viesado (BLUP) que serão descritos neste capítulo.

Hox (1995) considera que os estimadores mais usados na análise de modelos lineares hierárquicos são os estimadores de máxima verossimilhança que estimam os parâmetros do modelo ao prover estimativas para os valores populacionais que maximizam a função de verossimilhança.

Inicialmente serão apresentados os estimadores de mínimos quadrados para os efeitos fixos tendo em vista que sob a suposição de normalidade os estimadores de máxima verossimilhança são iguais aos estimadores de mínimos quadrados.

3.1. Estimação dos Parâmetros Fixos do MLH

Para o modelo (2.1) que envolve apenas um efeito fixo (γ_{00}), estima-se a média geral como uma média ponderada das médias amostrais dos J grupos do nível 2. Relembrando a equação (2.1), que é o modelo do nível 1, tem-se:

$$Y_{ij} = \beta_{0j} + e_{ij} \tag{3.1}$$

com $e_{ij} \sim N(0, \sigma^2)$, $i = 1, \ldots, n_j,\ j = 1, \ldots, J$.

A média para as n_j observações dentro do grupo j, resulta em um modelo do nível 1 com média amostral como resposta:

$$\overline{Y}_{.j} = \beta_{0j} + \overline{e}_{.j}, \tag{3.2}$$

onde

$\overline{Y}_{.j} = \frac{\sum Y_{ij}}{n_j}$ é a média do grupo j. Este é um estimador para β_{0j}, $j = 1, \ldots, J$;

$\overline{e}_{.j} = \sum \frac{e_{ij}}{n_j}$, é o erro médio de estimação que tem variância dada por:

$$Var(\overline{e}_{.j}) = \frac{\sigma^2}{n_j} = V_j, \tag{3.3}$$

onde V_j é a variância do erro médio.

Lembrando o modelo do nível 2, tem-se:

$$\beta_{0j} = \gamma_{00} + u_{0j}, \tag{3.4}$$

onde $u_{0j} \sim N(0, \tau_{00})$, sendo τ_{00} a variância dos erros do nível 2, sendo a média geral γ_{00}. Logo o parâmetro β_{0j} tem variância τ_{00}.

Da substituição de (3.4) em (3.2) temos o modelo combinado para $\overline{Y}_{.j}$:

$$\overline{Y}_{.j} = \gamma_{00} + u_{0j} + \overline{e}_{.j}, \tag{3.5}$$

onde $\overline{Y}_{.j}$ possui variância com duas componentes:

$$\begin{aligned} Var(\overline{Y}_{.j}) &= Var(u_{0j}) + Var(\overline{e}_{.j}) \\ &= \tau_{00} + V_j \\ &= \Delta_j,\ j = 1, \ldots, J. \end{aligned} \tag{3.6}$$

Embora a variância do parâmetro, τ_{00}, seja constante por grupo do nível 2, a variância do erro médio, V_j, varia dependendo do tamanho da amostra (n_j) de cada grupo do nível 2. Caso os grupos do nível 2 tenham o mesmo tamanho de amostra, então $V_j = V$ e $\Delta_j = \Delta$, portanto $\Delta = \tau_{00} + V$.

Fazendo,

$$z_j = \overline{Y}_{.j} - \mathrm{E}(\overline{Y}_{.j}) = \overline{Y}_{.j} - \gamma_{00}$$

Minimizando a soma dos quadrados dos desvios:

$$M = \sum {z_j}^2 = \sum (\overline{Y}_{.j} - \mathrm{E}(\overline{Y}_{.j}))^2 = \sum (\overline{Y}_{.j} - \gamma_{00})^2$$

$$\frac{dM}{d\gamma_{00}} = -2 \sum (\overline{Y}_{.j} - \gamma_{00})$$

$$\sum (\overline{Y}_{.j} - \widetilde{\gamma}_{00}) = 0 \Longrightarrow \sum \overline{Y}_{.j} - \sum \widetilde{\gamma}_{00} = 0 \Longrightarrow J\widetilde{\gamma}_{00} = \sum \overline{Y}_{.j}$$

Desta forma o estimador de γ_{00} é somente a média dos valores de $\overline{Y}_{.j}$:

$$\widetilde{\gamma}_{00} = \frac{\sum \overline{Y}_{.j}}{J}. \tag{3.7}$$

Este estimador é não viesado de variância mínima considerando que todos os J grupos têm o mesmo tamanho de amostra (Bryk e Raudenbush, 1992 e 2002).

Porém, se os tamanhos de amostras forem diferentes, as estatísticas $\overline{Y}_{.j}$, terão variâncias diferentes, $\Delta_j = \tau_{00} + V_j$. Considerando cada $\overline{Y}_{.j}$ como um estimador independente e não viesado de γ_{00} com variância Δ_j, definimos a precisão (C) de $\overline{Y}_{.j}$ como:

$$C(\overline{Y}_{.j}) = \Delta_j^{-1},\ j = 1, \ldots, J. \tag{3.8}$$

Então, assumindo-se que Δ_j é conhecido, o estimador não viesado de variância mínima de γ_{00} é a média ponderada pela precisão:

$$\widehat{\gamma}_{00} = \frac{\sum \Delta_j^{-1} \overline{Y}_{.j}}{\sum \Delta_j^{-1}} \,. \tag{3.9}$$

$\widehat{\gamma}_{00}$ é chamado de estimador de mínimos quadrados ponderados de γ_{00}. Os valores de Δ_j devem ser conhecidos ou estimados para que se possa calcular $\widehat{\gamma}_{00}$.

Considerando ainda o modelo (2.1) levando em consideração que o modelo do nível 2 é obtido como em (2.5), a média do nível 1 será predita por uma variável do nível 2 e o modelo combinado para a média amostral é:

$$\overline{Y}_{.j} = \gamma_{00} + \gamma_{01} W_j + u_{0j} + \overline{e}_{.j},\ j = 1, \ldots, J, \tag{3.10}$$

e a variância de $\overline{Y}_{.j}$ dado W_j é obtida como em (3.6), sendo que agora Δ_j é a variância residual de $\overline{Y}_{.j}$, que é a variância condicional de $\overline{Y}_{.j}$ dado W_j, e os erros u_{0j} e e_{ij} continuam sendo normalmente distribuídos.

Quando todos os J grupos têm o mesmo tamanho de amostra, a variância residual Δ_j é idêntica em todos os grupos, o único estimador não viesado de mínima variância de γ_{01} é o estimador de mínimos quadrados ordinários:

$$\widetilde{\gamma}_{01} = \frac{\sum (W_j - \overline{W}_{.})(\overline{Y}_{.j} - \overline{Y}_{..})}{\sum (W_j - \overline{W}_{.})^2}, \tag{3.11}$$

onde

$\overline{W}_{.} = \frac{\sum W_j}{J}$ é a média da variável explicativa do nível 2;

$\overline{Y}_{.j}$ é a resposta média de cada grupo dado W_j, $j = 1, \ldots, J$;

$\overline{Y}_{..} = \frac{\sum \overline{Y}_{.j}}{J}$ é a média das estimativas médias dos grupos.

Considerando o modelo (3.10), o estimador de mínimos quadrados ordinários de γ_{00} é:

$$\widetilde{\gamma}_{00} = \overline{Y}_{..} - \widetilde{\gamma}_{01} \overline{W}_{.} \ . \tag{3.12}$$

Quando os tamanhos das amostras n_j são diferentes, os $\overline{Y}_{.j}$ têm variâncias diferentes dadas por $\Delta_j = \tau_{00} + V_j$. Assumindo-se que estas variâncias (Δ_j) são conhecidas, o único estimador não viesado de variância mínima de γ_{01} é o estimador de mínimos quadrados ponderados onde cada conjunto de dados é proporcionalmente ponderado pela sua precisão Δ_j^{-1}, dado por:

$$\widehat{\gamma}_{01} = \frac{\sum \Delta_j^{-1} (W_j - \overline{W}.^*)(\overline{Y}_{.j} - \overline{Y}_{..}^*)}{\sum \Delta_j^{-1} (W_j - \overline{W}.^*)^2}, \tag{3.13}$$

onde $\overline{W}_{\cdot}^{\,*}$ e $\overline{Y}_{\cdot\cdot}^{\,*}$ são precisões médias ponderadas, sendo dadas por:

$$\overline{W}_{\cdot}^{*} = \frac{\sum \Delta_j^{-1} W_j}{\sum \Delta_j^{-1}},$$

e

$$\overline{Y}_{\cdot\cdot}^{\,*} = \frac{\sum \Delta_j^{-1} \overline{Y}_{\cdot j}}{\sum \Delta_j^{-1}}.$$

O estimador de mínimos quadrados ponderado de γ_{00} é:

$$\widehat{\gamma}_{00} = \overline{Y}_{\cdot\cdot}^{\,*} - \widehat{\gamma}_{01} \overline{W}_{\cdot}^{\,*} \quad . \tag{3.14}$$

Analisando o modelo (2.10), no qual há uma variável explicativa para o nível 1 e nenhuma variável explicativa para o nível 2, o modelo para a média amostral é dado por:

$$\overline{Y}_{\cdot j} = \gamma_{00} + \gamma_{10} \overline{X}_{\cdot j} + u_{0j} + u_{1j} \overline{X}_{\cdot j} + \overline{e}_{\cdot j},\ j = 1, \ldots, J, \tag{3.15}$$

e a variância de $\overline{Y}_{\cdot j}$ dado $\overline{X}_{\cdot j}$ é:

$$\begin{aligned} Var(\overline{Y}_{\cdot j}) &= Var(u_{0j}) + Var(u_{1j} \overline{X}_{\cdot j}) + Var(\overline{e}_{\cdot j}) \\ &= \tau_{00} + \overline{X}_{\cdot j}^{\,2} \tau_{11} + V_j \\ &= \Delta_j, \end{aligned} \tag{3.16}$$

onde Δ_j é a variância condicional de $\overline{Y}_{\cdot j}$ dado $\overline{X}_{\cdot j}$.

O estimador não viesado de γ_{10} é o estimador de mínimos quadrados ordinários, quando considera-se que todos os grupos têm o mesmo tamanho de amostra e é dado por:

$$\widetilde{\gamma}_{10} = \frac{\sum (\overline{X}_{\cdot j} - \overline{X}_{\cdot\cdot})(\overline{Y}_{\cdot j} - \overline{Y}_{\cdot\cdot})}{\sum (\overline{X}_{\cdot j} - \overline{X}_{\cdot\cdot})^2}, \tag{3.17}$$

onde

$\overline{X}_{\cdot j} = \frac{\sum X_{ij}}{n_j}$ é a média da variável explicativa do nível 1 em cada grupo, $j = 1, \ldots, J$;

$\overline{X}_{\cdot\cdot} = \frac{\sum \overline{X}_{\cdot j}}{J}$ é a média das estimativas médias de cada grupo da variável explicativa do nível 1;

$\overline{Y}_{\cdot j}$ é a resposta média de cada grupo dado $\overline{X}_{\cdot j}$, $j = 1, \ldots, J$;

$\overline{Y}_{\cdot\cdot}$ é a média das estimativas da média de cada grupo dada a variável explicativa do nível 1.

O estimador de mínimos quadrados ordinário de γ_{00} é:

$$\widetilde{\gamma}_{00} = \overline{Y}_{\cdot\cdot} - \widetilde{\gamma}_{10} \overline{X}_{\cdot\cdot} \quad . \tag{3.18}$$

Levando-se em conta que os tamanhos de amostras n_j são diferentes em cada grupo, as variâncias Δ_j também diferem, e o estimador não viesado de mínima variância de $\widetilde{\gamma}_{10}$, considerando Δ_j conhecido, é o estimador de mínimos quadrados ponderado, sendo cada conjunto de dados proporcional a sua precisão Δ_j^{-1}:

$$\widehat{\gamma}_{10} = \frac{\sum \Delta_j^{-1}(\overline{X}_{.j} - \overline{X}^*_{..})(\overline{Y}_{.j} - \overline{Y}^*_{..})}{\sum \Delta_j^{-1}(\overline{X}_{.j} - \overline{X}^*_{..})^2}, \tag{3.19}$$

onde $\quad \overline{X}^*_{..} = \frac{\sum \Delta_j^{-1}\overline{X}_{.j}}{\Delta_j^{-1}} \quad$ e $\quad \overline{Y}^*_{..} = \frac{\sum \Delta_j^{-1}\overline{Y}_{.j}}{\Delta_j^{-1}}$.

O estimador de mínimos quadrados ponderado de γ_{00} é:

$$\widehat{\gamma}_{00} = \overline{Y}^*_{..} - \widehat{\gamma}_{10}\overline{X}^*_{..} \quad . \tag{3.20}$$

Considerando agora o modelo mais geral obtido pela equação (2.13), onde existem $q\,(q = 1, \ldots, Q)$ variáveis explicativas no nível 1 e $p\,(p = 1, \ldots, P)$ variáveis explicativas no nível 2, a extensão dos princípios básicos de estimação é feita de forma direta. O modelo geral do nível 1 com P variáveis explicativas é expresso logo abaixo em notação matricial:

$$\underset{\sim}{Y}_j = \mathbf{X}_j \underset{\sim}{\beta}_j + \underset{\sim}{e}_j,\ \underset{\sim}{e}_j \sim N(0, \sigma^2\mathbf{I}),\ j = 1, \ldots, J, \tag{3.21}$$

onde

$\underset{\sim}{Y}_j{}' = [\,Y_{1j} \quad Y_{2j} \quad \ldots \quad Y_{n_jj}\,]$, é o vetor da variável resposta do grupo j, $j = 1, \ldots, J$;

$\mathbf{X}_j = \begin{bmatrix} 1 & X_{11j} & \ldots & X_{Q1j} \\ 1 & X_{12j} & \ldots & X_{Q2j} \\ \vdots & \vdots & \ldots & \vdots \\ 1 & X_{1n_jj} & \ldots & X_{Qn_jj} \end{bmatrix}$, é a matriz de variáveis preditoras do nível 1, do grupo j, $j = 1, \ldots, J$;

$\underset{\sim}{\beta}_j{}' = [\,\beta_{0j} \quad \beta_{1j} \quad \ldots \quad \beta_{Qj}\,]$, é o vetor de parâmetros desconhecidos, $j = 1, \ldots, J$;

$\mathbf{I} = \begin{bmatrix} 1 & 0 & \ldots & 0 \\ 0 & 1 & \ldots & 0 \\ \vdots & \vdots & \ldots & \vdots \\ 0 & 0 & \ldots & 1 \end{bmatrix}$, é a matriz identidade de dimensão $n_j,\ j = 1, \ldots, J$;

$\underset{\sim}{e}_j{}' = [\,e_{1j} \quad e_{2j} \quad \ldots \quad e_{n_jj}\,]$, é o vetor de erros aleatórios, $j = 1, \ldots, J$.

Assumindo-se que $\mathbf{X}_j$ é posto coluna completo Q, o estimador de mínimos quadrados ordinários de $\underset{\sim}{\beta}_j$ é:

$$\widehat{\underset{\sim}{\beta}}_j = (\mathbf{X}_j'\mathbf{X}_j)^{-1}\mathbf{X}_j'\underset{\sim}{Y}_j, \tag{3.22}$$

e sua matriz de variância é dada por:

$$\begin{aligned} Var(\widehat{\underset{\sim}{\beta}}_{\,j}) &= \mathbf{V}_j \\ &= \sigma^2(\mathbf{X}_j'\mathbf{X}_j)^{-1}. \end{aligned} \tag{3.23}$$

Substituindo $\underset{\sim}{Y}_{\,j}$ da equação (3.21) em (3.22), o modelo para $\widehat{\underset{\sim}{\beta}}_{\,j}$ é dado por:

$$\widehat{\underset{\sim}{\beta}}_{\,j} = \underset{\sim}{\beta}_{\,j} + \underset{\sim}{r}_{\,j},\ \underset{\sim}{r}_{\,j} \sim\ N(0, \mathbf{V}_j), \tag{3.24}$$

onde

$\underset{\sim}{r}_{\,j} = (\mathbf{X}_j'\mathbf{X}_j)^{-1}\mathbf{X}_j'\,\underset{\sim}{e}_{\,j}$, é o vetor de erros;

$\mathbf{V}_j$ é dada na expressão (3.23).

O modelo geral para $\underset{\sim}{\beta}_{\,j}$ no nível 2 é:

$$\underset{\sim}{\beta}_{\,j} = \mathbf{W}_j\underset{\sim}{\gamma}\ + \underset{\sim}{u}_{\,j},\ \underset{\sim}{u}_{\,j} \sim\ N(0, \mathbf{T}_{(\underset{\sim}{u}_{\,j})}), \tag{3.25}$$

onde

$\mathbf{W}_j = \begin{bmatrix} \underset{\sim}{W}_{\,0j} & \underset{\sim}{0} & \dots & \underset{\sim}{0} \\ \underset{\sim}{0} & \underset{\sim}{W}_{\,1j} & \dots & \underset{\sim}{0} \\ \vdots & \vdots & \dots & \vdots \\ \underset{\sim}{0} & \underset{\sim}{0} & \dots & \underset{\sim}{W}_{\,Qj} \end{bmatrix}$, é a matriz de variáveis preditoras do nível 2, sendo $\underset{\sim}{W}_{\,qj} = [\,1 \quad W_{1j} \quad W_{2j} \quad \dots \quad W_{Pj}\,]$ o vetor de preditores de β_{qj}, e $\underset{\sim}{0}$ um vetor Px1 de zeros, $q = 0, 1, \dots, Q$, $p = 1, \dots, P$ e $j = 1, \dots, J$;

$\underset{\sim}{\gamma}\,' = \begin{bmatrix} \underset{\sim}{\gamma}_{\,0} & \underset{\sim}{\gamma}_{\,1} & \underset{\sim}{\gamma}_{\,2} & \dots & \underset{\sim}{\gamma}_{\,Q} \end{bmatrix}$, é o vetor de efeitos fixos, sendo $\underset{\sim}{\gamma}_{\,q} = [\,\gamma_{q0} \quad \gamma_{q1} \quad \gamma_{q2} \quad \dots \quad \gamma_{qP}\,]$;

$\underset{\sim}{u}_{\,j}{}' = [\,u_{0j} \quad u_{1j} \quad \dots \quad u_{Qj}\,]$, é o vetor de efeitos aleatórios;

$\mathbf{T}_{(\underset{\sim}{u}_{\,j})} = \begin{bmatrix} \tau_{00} & \tau_{01} & \dots & \tau_{0Q} \\ \tau_{10} & \tau_{11} & \dots & \tau_{1Q} \\ \vdots & \vdots & \dots & \vdots \\ \tau_{Q0} & \tau_{Q1} & \dots & \tau_{QQ} \end{bmatrix}$, é a matriz de variância e covariância.

Observe que $\mathbf{T}_{(\underset{\sim}{u}_{\,j})}$ indica a dispersão de $\underset{\sim}{\beta}_{\,j}$ sobre o valor esperado de $\mathbf{W}_j\underset{\sim}{\gamma}$.

Substituindo a equação (3.25) na equação (3.24), temos o modelo combinado:

$$\widehat{\underset{\sim}{\beta}}_{\,j} = \mathbf{W}_j\underset{\sim}{\gamma}\ + \underset{\sim}{u}_{\,j} + \underset{\sim}{r}_{\,j}, \tag{3.26}$$

onde a variância de $\widehat{\underset{\sim}{\beta}}_{\ j}$ dado $\mathbf{W}_j$ é:

$$\begin{aligned} Var(\widehat{\underset{\sim}{\beta}}_{\ j}) &= Var(\underset{\sim}{u}_{\ j} + \underset{\sim}{r}_{\ j}) \\ &= \mathbf{T}_{(\underset{\sim}{u}_{\ j})} + \mathbf{V}_j \\ &= \mathbf{\Delta}_j. \end{aligned} \tag{3.27}$$

Caso os J grupos tenham o mesmo tamanho de amostra, então cada $\widehat{\underset{\sim}{\beta}}_{\ j}$ terá a mesma dispersão $\mathbf{\Delta}$, onde:

$$\begin{aligned} \mathbf{\Delta} &= \mathbf{T} + \mathbf{V} \\ &= \mathbf{T} + \sigma^2(\mathbf{X}'\mathbf{X})^{-1}. \end{aligned} \tag{3.28}$$

Desta forma, o único estimador não viesado de mínima variância de $\underset{\sim}{\gamma}$ é o estimador de mínimos quadrados ordinários:

$$\widetilde{\underset{\sim}{\gamma}} = \left(\sum \mathbf{W}_j'\mathbf{W}_j\right)^{-1}\sum \mathbf{W}_j'\,\widehat{\underset{\sim}{\beta}}_{\ j}. \tag{3.29}$$

Porém, se o tamanho de amostra difere em cada grupo, os valores da dispersão $\mathbf{\Delta}_j$ também serão diferentes, e o único estimador não viesado de mínima variância para $\underset{\sim}{\gamma}$, assumindo que cada $\mathbf{\Delta}_j$ é conhecido, é o estimador de mínimos quadrados generalizado:

$$\widehat{\underset{\sim}{\gamma}} = \left(\sum \mathbf{W}_j'\mathbf{\Delta}_j^{-1}\mathbf{W}_j\right)^{-1}\sum \mathbf{W}_j'\mathbf{\Delta}_j^{-1}\,\widehat{\underset{\sim}{\beta}}_{\ j}, \tag{3.30}$$

que pondera cada conjunto de dados pela respectiva matriz de precisão $\mathbf{\Delta}_j^{-1}$, que é o inverso da matriz de variância e covariância.

3.2. Método de Máxima Verossimilhança

O método de Máxima Verossimilhança adota como estimativas dos parâmetros, os valores que maximizam a probabilidade (variável discreta) ou a densidade de probabilidade (variável contínua) de ser obtida a amostra observada. É necessário conhecer a distribuição da variável em estudo para obter estimadores de Máxima Verossimilhança.

Os softwares estatísticos normalmente usam dois procedimentos diferentes de estimação por máxima verossimilhança na análise de modelo linear hierárquico. Um é o procedimento de máxima verossimilhança completo, que inclui na função de verossimilhança os efeitos fixos e as componentes de variância. O outro procedimento é o de máxima verossimilhança restrito, que inclui apenas as componentes de variância na função de verossimilhança. A diferença entre os métodos de estimação é que o procedimento de máxima verossimi-

lhança completo trata as estimativas dos coeficientes de regressão como quantidades que são conhecidas quando estimam as componentes de variância, e o procedimento de máxima verossimilhança restrito trata estas estimativas como quantidades que carregam uma certa incerteza. Na prática, a diferença entre os dois métodos de estimação não é muito grande (Hox, 1995).

O cálculo das estimativas por máxima verossimilhança requer um procedimento iterativo onde são fornecidos valores iniciais para os vários parâmetros que estão sendo estimados, em seguida estes valores vão sendo aperfeiçoados a medida em que o processo é repetido uma série de vezes. Depois de cada iteração, o critério de convergência verifica se a diferença entre os dois últimos valores das estimativas obtidas é muito pequena, caso esta diferença seja próxima de zero conclui-se que houve convergência, finalizando o procedimento.

Nos casos onde os tamanhos de amostra são iguais, existem fórmulas matemáticas de forma fechada para estimar as componentes de variância e covariância. Quando os tamanhos de amostras são diferentes, os procedimentos iterativos são usados para obter estimadores eficientes, em geral, pelo método de máxima verossimilhança (completo ou restrito) (Hox, 1995).

3.2.1. Método de Máxima Verossimilhança Completo

O modelo linear hierárquico descreve a variável dependente $\underset{\sim}{Y}$ e aplica o princípio de máxima verossimilhança para este modelo. A distribuição de $\underset{\sim}{Y}$ é assumida como sendo normal, com média dependendo dos coeficientes de regressão e o parâmetro de dispersão dependendo das componentes de variância. Estes parâmetros são estimados pela técnica correspondente, que é simplesmente chamada de máxima verossimilhança, mas às vezes também chamada de máxima verossimilhança completa (Kreft e Leeuw, 1998).

De acordo com Bryk e Raudenbush (1992) para um conjunto de possíveis valores dos parâmetros $\underset{\sim}{\gamma}$,$\mathbf{T}_{(\underset{\sim}{u}_j)}$, e σ^2 nas equações (3.21) e (3.25), existe alguma probabilidade de observar uma amostra de valores de $\underset{\sim}{Y}$, onde $\underset{\sim}{Y}$ é um vetor Nx1 de valores observados da variável resposta estudada, das N unidades do nível 1. Veja que $N = \sum n_j$, onde n_j é o tamanho da amostra do j-ésimo grupo. A idéia central do método de máxima verossimilhança é escolher estimativas de $\underset{\sim}{\gamma}$,$\mathbf{T}_{(\underset{\sim}{u}_j)}$, e σ^2 para as quais a probabilidade de observar os valores reais de $\underset{\sim}{Y}$ é máxima.

Estimadores baseados no método de máxima verossimilhança têm certas propriedades úteis. Esses estimadores, sob certas suposições, são consistentes (isto é, eles podem ser bastante próximos do verdadeiro valor do parâmetro com alta probabilidade, se muitos dados forem coletados) e assintoticamente eficiente (isto é, para um tamanho de amostra grande,

o estimador de máxima verossimilhança é aproximadamente não viesado com variância mínima).

Anteriormente foi mencionado que, para tamanhos de amostras dos J grupos diferentes, o estimador de γ_{00} em um modelo com apenas o intercepto aleatório não balanceado é dado como em (3.9), sob a suposição que Δ_j é conhecido. Caso os Δ_j fossem desconhecidos, mas fossem substituídos pelos respectivos valores estimados por máxima verossimilhança, o estimador $\widehat{\gamma}_{00}$ resultante seria, pela propriedade de invariância, um estimador de máxima verossimilhança com suas propriedades estatísticas úteis já mencionadas.

Outras características do estimador de máxima verossimilhança é que sua distribuição amostral converge assintoticamente para uma normal com uma variância que pode ser estimada facilmente. Desta forma, se até mesmo o método para obter o estimador de máxima verossimilhança completo é iterativo (porque não possui uma expressão analítica de forma fechada disponível), a distribuição para grandes amostras do estimador está bem definida.

3.2.2. Método de Máxima Verossimilhança Restrito

O princípio de máxima verossimilhança aplicado aos resíduos de mínimos quadrados é conhecido como máxima verossimilhança restrito. Significa que primeiro é removido o efeito das variáveis fixas (lembrando que os resíduos, que possuem distribuição normal, são não correlacionados com todas as variáveis fixas no modelo). Como a distribuição dos resíduos não depende das estimativas dos efeitos fixos, só depende das componentes de variância, então, aplicar o princípio de máxima verossimilhança aos resíduos implica que não se pode estimar os coeficientes de regressão. Desta forma, é aplicado um outro princípio para estimar os coeficientes de regressão que é o de mínimos quadrados ponderados, usado para estimar as componentes de variância para construir a matriz ponderada (Kreft e Leeuw, 1998).

O método de máxima verossimilhança restrito requer a integração do método de máxima verossimilhança completo com respeito a γ . Esta integração da verossimilhança é uma idéia bayesiana que faz sentido apenas se a verossimilhança for observada como a densidade conjunta à posteriori de γ , $\mathbf{T}_{(\underset{\sim}{u}\ j)}$ e σ^2 (para maiores detalhes ver Dempster *et al*, 1981), onde σ^2 é a variância dos erros do nível 1, γ é o vetor de efeitos fixos e $\mathbf{T}_{(\underset{\sim}{u}\ j)}$ é a matriz de variância e covariância do nível 2.

Uma deficiência do método de máxima verossimilhança completo apontada por Bryk e Raudenbush (1992) é que os estimadores das variâncias e covariâncias são condicionais aos estimadores pontuais dos efeitos fixos. Considerando o modelo de regressão múltipla abaixo:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_Q X_{Qi} + e_i,\ q = 1, \ldots, Q, \tag{3.31}$$

onde os erros e_i $(i = 1, \ldots, n)$ são normalmente distribuídos com média zero e variância constante, σ^2; imaginando que os $Q+1$ coeficientes de regressão $(\beta_0, \beta_1, \ldots, \beta_Q)$ são conhecidos, então, o estimador de máxima verossimilhança de σ^2 é:

$$\widehat{\sigma}^2 = \frac{\sum e_i^2}{n}. \tag{3.32}$$

Supondo agora que os $Q+1$ coeficientes de regressão são desconhecidos, eles devem então ser estimados. Deste modo, o resíduo é dado por:

$$\widehat{e}_i = Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 X_{1i} - \widehat{\beta}_2 X_{2i} - \ldots - \widehat{\beta}_Q X_{Qi}, \tag{3.33}$$

onde cada $\widehat{\beta}_q$ é estimado pelo método de mínimos quadrados ordinários. Neste caso, o estimador não viesado de σ^2 é dado por:

$$\widehat{\sigma}^2 = \frac{\sum \widehat{e}_i^2}{(n - Q - 1)}. \tag{3.34}$$

Pode-se observar que o denominador $(n - Q - 1)$, corrigido pelos graus de liberdade usa os $Q + 1$ parâmetros de regressão estimados. Como freqüentemente Q é pequeno, a correção tem efeito pequeno, uma correção melhor é $(n - Q + 1)$ em lugar de $(n - Q - 1)$ (Bryk e Raudenbush, 1992 e 2002). Porém a medida em que o número de variáveis explicativas (Q) aumenta, o uso da equação (3.32) pode conduzir a um grande viés na estimação de σ^2. O viés pode ser negativo, assim a estimativa de σ^2 pode ser pequena, conduzindo à intervalos de confiança artificialmente curtos e a testes de hipóteses excessivamente amplos, ou seja, com pequena região de aceitação da hipótese. Considerando então o modelo de regressão múltipla (3.31), a diferença entre as equações (3.32) e (3.34) é precisamente a diferença entre o método de máxima verossimilhança completo (3.32) e o método de máxima verossimilhança restrito (3.34).

Tratando-se do modelo linear hierárquico, a diferença entre estimativas de variância e covariância baseada no método de máxima verossimilhança completo e máxima verossimilhança restrito, não é expressa em forma algébrica simples. Porém, as estimativas de máxima verossimilhança restrita das componentes de variância são ajustadas por conta da incerteza sobre os efeitos fixos e os resultados obtidos pelo método de máxima verossimilhança completo não sofrem este ajuste.

No caso do modelo linear hierárquico de dois níveis, o método de máxima verossimilhança completo e o método de máxima verossimilhança restrito geralmente produzem resultados muito parecidos para σ^2, mas a diferença visível pode ocorrer na estimação de $\mathbf{T}_{(\underset{\sim}{u}\ j)}$. Nos casos onde o número de grupos do nível 2, J, é grande, os dois métodos de estimação podem produzir resultados muito similares. Entretanto, se J é pequeno o estimador da variância pelo método de máxima verossimilhança completo, $\widehat{\tau}_{qq}$, pode produzir resultado

menor que o encontrado pelo método de máxima verossimilhança restrito, por um fator de aproximadamente $(J-Q)/J$, onde Q é o número total de elementos do vetor de efeitos fixos, $\underset{\sim}{\gamma}$.

3.3. Algoritmo EM

Como relata Sousa (2002) em muitos casos não é possível obter uma forma explícita para o estimador de máxima verossimilhança (EMV), desta forma, uma alternativa para se obter as estimativas dos parâmetros em questão com dados incompletos, é utilizar métodos iterativos como o algoritmo EM. Este algoritmo foi apresentado em 1977 por Dempster *et al* sendo uma alternativa computacional aos algoritmos de Newton-Raphson e Scoring de Fisher, pois não requer a obtenção da segunda derivada do logaritmo da função de verossimilhança. O fato de cada iteração consistir de dois passos: passo E (obtenção das esperanças sobre a distribuição condicional) e passo M (estimação de máxima verossimilhança de dados completos) resultou no termo EM.

De acordo com Dempster *et al* (1977) o algoritmo EM foi introduzido por Hartley (1958) como um procedimento para calcular estimativas de máxima verossimilhança de uma amostra aleatória de tamanho n de uma população discreta, onde alguns dos valores não são atribuídos aos indivíduos, mas às observações agregadas. Carter e Myers (1973) propuseram o algoritmo EM para estimação de máxima verossimilhança de combinação linear de funções de probabilidade discreta, usando combinação linear de variáveis aleatórias Poisson como um exemplo. Brown (1974) também sugeriu o algoritmo para calcular estimativas de máxima verossimilhança de frequências esperadas de observações sob um modelo de independência em uma tabela de dupla entrada, com algumas observações *missing*.

Este algoritmo tem sido usado para estimação de mínimos quadrados em análise de variância ou de forma equivalente, para estimação de máxima verossimilhança sob o modelo linear normal com variância residual qualquer, sendo uma referência básica Healy e Westmacott (1956).

A idéia chave é que cálculos de mínimos quadrados exatos são facilmente executados para matrizes de desenho especial, que incorporam o balanceamento necessário e propriedades de ortogonalidade, enquanto o cálculo de mínimos quadrados para desenhos não balanceados requer a inversão de uma matriz de dimensão grande.

Vários métodos precursores na linha do algoritmo EM são considerados como algoritmos EM num contexto especial. Blight (1970) foi quem criou um método iterativo para solucionar o problema que era encontrar o EMV dos parâmetros da família exponencial para uma amostra censurada, obtendo a equação de verossimilhança.

O princípio dos "dados faltantes" ou informações *missing* que está relacionado às idéias básicas do algoritmo EM, foi introduzido por Orchard e Woodbury (1972) que estabeleceram a relação entre a função do logaritmo da verossimilhança de dados completos e incompletos, conduzindo ao fato que o EMV é um ponto fixo. Uma distribuição empírica de dados arbitrariamente agrupados, censurados e truncados foi apresentada por Turnbull (1976) onde ele obteve uma versão do algoritmo EM e observou que problemas com dados truncados podem ser tratados como um problema com dados incompletos. Propriedades da equação de verossimilhança no contexto geral de dados incompletos da família exponencial foram verificados por Sundberg (1974), enquanto Beale e Little (1975) desenvolveram um algoritmo e a teoria associada ao caso da normal multivariada com dados incompletos. Mas, foram Dempster *et al* (1977) que formularam o problema de um modo geral ao provarem resultados gerais sobre os procedimentos utilizados na formulação do algoritmo, estabelecendo sua convergência e outras propriedades básicas em sua generalidade.

O algoritmo EM propõe a substituição dos valores incompletos por valores estimados, em seguida a estimação dos parâmetros. Logo após, a re-estimação dos valores incompletos, assumindo que as novas estimativas dos parâmetros são corretas, e finalmente a re-estimação dos parâmetros de forma iterativa, até a convergência.

Este algoritmo é numericamente estável, pois haverá convergência da sucessão de iterações quase sempre para um máximo local do logaritmo da função de verossimilhança. Por se basear em resultados de dados completos, o algoritmo EM é facilmente implementado e pode ser efetuado em computadores com pouca memória, pois requer pequeno espaço de armazenamento já que cada iteração é rápida e simples. As estimativas estão dentro do espaço paramétrico e as soluções são altamente robustas para valores iniciais que estão distantes do verdadeiro valor do parâmetro. Ele tem a desvantagem de não produzir estimativa da matriz de covariância do parâmetro estimado e, converge lentamente se muitos dos dados forem *missing*. Em alguns casos não é possível tratar analiticamente o passo E.

Considerando o modelo geral dado em (3.21) sendo o modelo do nível 2 como apresentado em (3.25), então o modelo combinado é dado por:

$$\underset{\sim}{Y}_{\,j} = \mathbf{X}_j \mathbf{W}_j \underset{\sim}{\gamma} + \mathbf{X}_j \underset{\sim}{u}_{\,j} + \underset{\sim}{e}_{\,j}\,. \tag{3.35}$$

Levando em consideração os modelos provenientes de cada unidade $j = 1, \ldots, J$, a equação (3.35) pode ser escrita como:

$$\underset{\sim}{Y} = \mathbf{X}\mathbf{W}\underset{\sim}{\gamma} + \mathbf{X}\underset{\sim}{u} + \underset{\sim}{e}\ , \tag{3.36}$$

que é um caso especial do modelo linear geral misto:

$$\underset{\sim}{Y} = \mathbf{A}_1 \underset{\sim}{\theta}_{\,1} + \mathbf{A}_2 \underset{\sim}{\theta}_{\,2} + \underset{\sim}{e}\ , \tag{3.37}$$

com $\mathbf{A}_1 = \mathbf{XW}$, $\mathbf{A}_2 = \mathbf{X}$, $\underset{\sim}{\theta}_1 = \gamma$ e $\underset{\sim}{\theta}_2 = \underset{\sim}{u}$, sendo $\mathbf{A}_1$ e $\mathbf{A}_2$ matrizes de preditores conhecidos;

$\underset{\sim}{Y}$ o vetor de respostas;

$\underset{\sim}{\theta}_1 \sim N(\underset{\sim}{0}, \mathbf{\Gamma})$ o vetor de efeitos fixos desconhecidos com média a priori zero e matriz de dispersão a priori $\mathbf{\Gamma}$;

$\underset{\sim}{\theta}_2 \sim N(\underset{\sim}{0}, \mathbf{T})$ é o vetor de efeitos aleatórios desconhecidos do nível 2, com

$$\mathbf{T} = \begin{bmatrix} \mathbf{T}_{(\underset{\sim}{u}_1)} & \underset{\sim}{0} & \dots & \underset{\sim}{0} \\ \underset{\sim}{0} & \mathbf{T}_{(\underset{\sim}{u}_2)} & \dots & \underset{\sim}{0} \\ \vdots & \vdots & \dots & \vdots \\ \underset{\sim}{0} & \underset{\sim}{0} & \dots & \mathbf{T}_{(\underset{\sim}{u}_j)} \end{bmatrix};$$

$\underset{\sim}{e} \sim N(\underset{\sim}{0}, \sigma^2\mathbf{I})$ é o vetor de erros aleatórios do nível 1.

De acordo com Dempster *et al* (1981), o modelo (3.37) pode ser re-escrito como:

$$d = \underset{\sim}{Y} - \mathbf{A}_1\underset{\sim}{\theta}_1 = \mathbf{A}_2\underset{\sim}{\theta}_2 + \underset{\sim}{e} \ . \tag{3.38}$$

A log-verossimilhança para o modelo geral é proporcional a:

$$log[f(\underset{\sim}{Y} \mid \sigma^2, \mathbf{T}_{(\underset{\sim}{u}_j)}, \underset{\sim}{\theta}_1)] \propto -(N - JR)log(\sigma^2) - Jlog \mid \mathbf{T}_{(\underset{\sim}{u}_j)} \mid +$$

$$+ \sum log \mid \mathbf{C}_j^{-1} \mid - \frac{\sum d_j'(d_j - A_{2j}\theta_{2j}^*)}{\sigma^2}, \tag{3.39}$$

onde

$\mathbf{C}_j^{-1} = (A_{2j}'A_{2j} + \sigma^2\mathbf{T}_{(\underset{\sim}{u}_j)}^{-1})^{-1}$ e $\theta_{2j}^* = C_j^{-1}A_{2j}'d_j$;

J é o total de unidades do nível 2 (grupos) e R é o número de efeitos aleatórios por unidade do nível 2.

Usando o teorema de Bayes e sabendo-se que $\underset{\sim}{\theta}_2$ e $\underset{\sim}{Y}$ são conhecidos, a verossimilhança da equação (3.39) pode ser maximizada através do algoritmo EM. A função de log-verossimilhanca dos dados completos é:

$$log[f(\underset{\sim}{Y}, \underset{\sim}{\theta}_2 \mid \underset{\sim}{\theta}_1, \sigma^2, \mathbf{T}_{(\underset{\sim}{u}_j)})] \propto -Nlog(\sigma^2) - Jlog \mid \mathbf{T}_{(\underset{\sim}{u}_j)} \mid -$$

$$\frac{\sum(d_j - A_{2j}\theta_{2j})'(d_j - A_{2j}\theta_{2j})}{\sigma^2} - \sum \theta_{2j}'\mathbf{T}_{(\underset{\sim}{u}_j)}{}^{-1}\theta_{2j}. \tag{3.40}$$

Sabendo que d é como dado em (3.38), e desenvolvendo o numerador da terceira parcela da função de log-verossimilhança (3.40) para maximizá-la com respeito a $\underset{\sim}{\theta}_1$, σ^2, $\mathbf{T}_{(\underset{\sim}{u}_j)}$

tem-se:

$$\begin{aligned}\sum(d_j - A_{2j}\theta_{2j})'(d_j - A_{2j}\theta_{2j}) &= \sum(Y_j - A_{1j}\theta_1 - A_{2j}\theta_{2j})'(Y_j - A_{1j}\theta_1 - A_{2j}\theta_{2j}) \\ &= \sum(Y_j'Y_j - Y_j'A_{1j}\theta_1 - Y_j'A_{2j}\theta_{2j} - \theta_1'A_{1j}'Y_j + \theta_1'A_{1j}'A_{1j}\theta_1 \\ &\quad +\theta_1'A_{1j}'A_{2j}\theta_{2j} - \theta_{2j}'A_{2j}'Y_j + \theta_{2j}'A_{2j}'A_{1j}\theta_1 + \theta_{2j}'A_{2j}'A_{2j}\theta_{2j}).\end{aligned}$$

Sabendo-se que $Y_j'A_{1j}\theta_1 = (\theta_1'A_{1j}'Y_j)'$ e $\theta_1'A_{1j}'A_{2j}\theta_{2j} = (\theta_{2j}'A_{2j}'A_{1j}\theta_1)'$,

temos que

$$\begin{aligned}\sum(d_j - A_{2j}\theta_{2j})'(d_j - A_{2j}\theta_{2j}) &= \sum Y_j'Y_j - 2\sum\theta_1'A_{1j}'Y_j - \sum Y_j'A_{2j}\theta_{2j} + \sum\theta_1'A_{1j}'A_{1j}\theta_1 \\ &\quad +2\sum\theta_1'A_{1j}'A_{2j}\theta_{2j} - \sum\theta_{2j}'A_{2j}'Y_j + \sum\theta_{2j}'A_{2j}'A_{2j}\theta_{2j}.\end{aligned}$$

Sendo $B = \sum(d_j - A_{2j}\theta_{2j})'(d_j - A_{2j}\theta_{2j})$, tem-se que a função de log-verossimilhança pode ser reescrita como:

$$log[f(\underset{\sim}{Y}, \underset{\sim}{\theta}_2 \mid \underset{\sim}{\theta}_1, \sigma^2, \mathbf{T}_{(\underset{\sim}{u}_j)})] \propto -Nlog(\sigma^2) - Jlog \mid \mathbf{T}_{(\underset{\sim}{u}_j)} \mid -$$

$$\frac{B}{\sigma^2} - \sum\theta_{2j}'\mathbf{T}_{(\underset{\sim}{u}_j)}^{-1}\theta_{2j}$$

Maximizando com respeito a $\underset{\sim}{\theta}_1$, tem-se:

$$\frac{d}{d\underset{\sim}{\theta}_1}log[f(\underset{\sim}{Y}, \underset{\sim}{\theta}_2 \mid \underset{\sim}{\theta}_1, \sigma^2\mathbf{T}_{(\underset{\sim}{u}_j)})] = \frac{1}{\sigma^2}[-2\sum(d\widehat{\theta}_1')A_{1j}'Y_j + \sum(d\widehat{\theta}_1')A_{1j}'A_{1j}\widehat{\theta}_1 +$$

$$\sum\widehat{\theta}_1'A_{1j}'A_{1j}(d\widehat{\theta}_1) + 2\sum(d\widehat{\theta}_1')A_{1j}'A_{2j}\theta_{2j}] = 0.$$

Como $\sum(d\widehat{\theta}_1')A_{1j}'A_{1j}\widehat{\theta}_1 = \sum\widehat{\theta}_1'A_{1j}'A_{1j}(d\widehat{\theta}_1)$, então:

$$\begin{aligned}\frac{1}{\sigma^2}[-2\sum(d\widehat{\theta}_1')A_{1j}'Y_j + 2\sum(d\widehat{\theta}_1')A_{1j}'A_{1j}\widehat{\theta}_1 + 2\sum(d\widehat{\theta}_1')A_{1j}'A_{2j}\theta_{2j}] &= 0. \\ -\sum A_{1j}'Y_j + \sum A_{1j}'A_{1j}\widehat{\theta}_1 + \sum A_{1j}'A_{2j}\theta_{2j} &= 0. \\ \sum A_{1j}'A_{1j}\widehat{\theta}_1 - \sum A_{1j}'(Y_j - A_{2j}\theta_{2j}) &= 0.\end{aligned}$$

$$\sum A_{1j}'A_{1j}\widehat{\theta}_1 = \sum A_{1j}'(Y_j - A_{2j}\theta_{2j}).$$

Logo

$$\widehat{\theta}_1 = \left(\sum A_{1j}'A_{1j}\right)^{-1}\sum A_{1j}'(Y_j - A_{2j}\theta_{2j}).$$

Maximizando agora com respeito a σ^2 tem-se:

$$\frac{d}{d\sigma^2} log[f(\underset{\sim}{Y}, \underset{\sim}{\theta}_2 \mid \underset{\sim}{\theta}_1, \sigma^2, \mathbf{T}_{(\underset{\sim}{u}\ j)})] = -\frac{N}{\widehat{\sigma}^2} + \frac{\sum (d_j - A_{2j}\theta_{2j})'(d_j - A_{2j}\theta_{2j})}{(\widehat{\sigma}^2)^2} = 0.$$

$$-\frac{N}{\widehat{\sigma}^2} = -\frac{\sum (d_j - A_{2j}\theta_{2j})'(d_j - A_{2j}\theta_{2j})}{(\widehat{\sigma}^2)^2}.$$

Desta forma:

$$\widehat{\sigma}^2 = \frac{\sum (d_j - A_{2j}\theta_{2j})'(d_j - A_{2j}\theta_{2j})}{N}.$$

Maximizando em relação a $\mathbf{T}_{(\underset{\sim}{u}\ j)}$:

$$\frac{d}{d\mathbf{T}_{(\underset{\sim}{u}\ j)}} log[f(\underset{\sim}{Y}, \underset{\sim}{\theta}_2 \mid \underset{\sim}{\theta}_1, \sigma^2, \mathbf{T}_{(\underset{\sim}{u}\ j)})] = -(2J\mathbf{T}^{-1}_{(\underset{\sim}{u}\ j)} - JD_{\mathbf{T}^{-1}_{(\underset{\sim}{u}\ j)}}$$

$$-\sum(-2\mathbf{T}^{-1}_{(\underset{\sim}{u}\ j)}\theta_{2j}\theta'_{2j}\mathbf{T}^{-1}_{(\underset{\sim}{u}\ j)} + D_{\mathbf{T}^{-1}_{(\underset{\sim}{u}\ j)}\theta_{2j}\theta'_{2j}\mathbf{T}^{-1}_{(\underset{\sim}{u}\ j)}}) = 0,$$

onde $D_{\mathbf{T}^{-1}_{(\underset{\sim}{u}\ j)}}$ é uma matriz diagonal cujos elementos são constituídos pela diagonal de $\mathbf{T}^{-1}_{(\underset{\sim}{u}\ j)}$, e $D_{\mathbf{T}^{-1}_{(\underset{\sim}{u}\ j)}\theta_{2j}\theta'_{2j}\mathbf{T}^{-1}_{(\underset{\sim}{u}\ j)}}$ é uma matriz diagonal cujos elementos são constituídos pela diagonal de $\mathbf{T}^{-1}_{(\underset{\sim}{u}\ j)}\theta_{2j}\theta'_{2j}\mathbf{T}^{-1}_{(\underset{\sim}{u}\ j)}$.

Depois de algumas manipulações algébricas, temos:

$$\widehat{\mathbf{T}}_{(\underset{\sim}{u}\ j)} = J^{-1}\sum \theta_{2j}\theta'_{2j}.$$

Durante o passo E do algoritmo EM, são calculadas as esperanças condicionais das estatísticas suficientes dos dados completos que são dadas por:

$$\mathrm{E}\Big[\Big(\sum A'_{1j}A_{1j}\Big)^{-1}\sum(Y_j - A_{2j}\theta_{2j}) \mid Y, \theta_1, \sigma^2, \mathbf{T}_{(\underset{\sim}{u}\ j)}\Big] =$$

$$\Big(\sum A'_{1j}A_{1j}\Big)^{-1}\sum(Y_j - A_{2j}\theta^*_{2j}), \tag{3.41}$$

$$\mathrm{E}\Big[\sum(d_j - A_{2j}\theta_{2j})'(d_j - A_{2j}\theta_{2j}) \mid Y, \theta_1, \sigma^2, \mathbf{T}_{(\underset{\sim}{u}\ j)}\Big] =$$

$$\sum(d_j - A_{2j}\theta^*_{2j})'(d_j - A_{2j}\theta^*_{2j}) + \sigma^2\, tr\Big[\sum A'_{2j}A_{2j}C_j^{-1}\Big] \tag{3.42}$$

e

$$\mathrm{E}\Big(\sum \theta_{2j}\theta'_{2j} \mid Y, \theta_1, \sigma^2, \mathbf{T}_{(\underset{\sim}{u}\ j)}\Big) = \sum \theta^*_{2j}\theta'^*_{2j} + \sigma^2\sum C_j^{-1}, \tag{3.43}$$

onde $\theta^*_{2j} = C_j^{-1} A'_{2j} d_j$.

O cálculo dessas esperanças condicionais depende de algumas estatísticas suficientes usadas por máxima verossimilhança restrita (Bryk e Raudenbush, 1992 e 2002).

3.4. Melhor Preditor Linear Não Viesado (BLUP)

O melhor preditor linear não viesado (BLUP), que prediz os efeitos aleatórios do modelo, foi obtido por Henderson na década de 50 quando descreveu a metodologia dos modelos mistos para obtenção do BLUP dos valores genéticos dos suínos. O BLUP é uma combinação de duas técnicas propostas de avaliação genética de suínos, uma foi desenvolvida na década de 30 denominada mínimos quadrados de Yates, e a outra na década de 40 denominada de índices de seleção de Smith e Hazel (Antunes, 2002).

Trabalhando com simulação de populações com características semelhantes as de suínos, Sorensen e Kennedy (1984) compararam dois métodos de seleção para avaliar a eficiência dos mesmos, um dos métodos usando mínimos quadrados ordinários e outro usando o modelo linear misto obtendo o BLUP. Eles avaliaram os ganhos obtidos com a seleção de cada um dos métodos durante três gerações consecutivas e observaram que as estimativas obtidas com o BLUP foram sempre melhores que as obtidas com o método de mínimos quadrados ordinários. Isto foi confirmado quando eles usaram as estimativas de ganho dos valores genéticos verdadeiros e observaram que eram próximos aos ganhos estimados pelo BLUP.

A eficiência do método de mínimos quadrados ordinários e do BLUP também foi estudada por Panter e Allen (1995), na identificação de linhagens superiores em cruzamentos de variedades de soja. Eles fizeram simulações com dados balanceados (igual número de dados em cada sub-classe) e desbalanceados, e em todos os casos estudados o BLUP apresentou menor erro padrão, maiores valores de correlação entre os valores preditos e o desempenho dos cruzamentos avaliados em campo, e ainda um percentual maior de identificação de cruzamentos superiores dentre os realizados.

Os suecos usaram o modelo do BLUP durante quase 20 anos para predizer valores de procriação dos seus cavalos reprodutores (garanhões) e foram os primeiros do mundo a aplicar esta técnica em cavalos. O Dr. Jan Philipsson, professor de procriação de animais e o presidente anterior da Associação de Warmblood sueca, trabalharam extensivamente para desenvolver o BLUP para os garanhões suecos. O índice é atualizado anualmente e é uma valiosa ferramenta para os criadores comprarem os cavalos pelo seu desempenho.

Henderson *et al* (1959) propuseram uma metodologia de modelos mistos, para obter predições dos valores genéticos, tratados como efeito aleatório, corrigidos para os demais

efeitos fixos contidos no modelo. O BLUP dos valores genéticos de cada indivíduo, pode ser obtido pela metodologia dos modelos mistos.

Considerando o modelo combinado apresentado em (3.36):

$$\underset{\sim}{Y}_{\ j} = \mathbf{X}_j\mathbf{W}_j\underset{\sim}{\gamma}\ + \mathbf{X}_j\underset{\sim}{u}_{\ j} + \underset{\sim}{e}_{\ j}$$

Chamando $\mathbf{X}_j\mathbf{W}_j = \mathbf{Z}_j$, tem-se o modelo misto abaixo:

$$\underset{\sim}{Y}_{\ j} = \mathbf{Z}_j\underset{\sim}{\gamma}\ + \mathbf{X}_j\underset{\sim}{u}_{\ j} + \underset{\sim}{e}_{\ j}$$

onde,

$\underset{\sim}{Y}_{\ j}$ é o vetor de observações;

$\mathbf{Z}_j$ é a matriz de incidência dos efeitos fixos;

$\mathbf{X}_j$ é a matriz de incidência dos efeitos aleatórios;

$\underset{\sim}{\gamma}$ é o vetor de efeitos fixos a serem estimados;

$\underset{\sim}{u}_{\ j}$ é o vetor de efeitos aleatórios a serem preditos com $N(\underset{\sim}{0}\ , \mathbf{T}_{(\underset{\sim}{u}_{\ j})})$;

e $\underset{\sim}{e}_{\ j}$ é o vetor de erros aleatórios associados a cada observação tal que $\underset{\sim}{e}_{\ j} \sim N(\underset{\sim}{0}\ , \sigma^2\mathbf{I})$.

Sabendo-se que $\mathbf{T}_{(\underset{\sim}{u}_{\ j})}$ e $\sigma^2\mathbf{I}$ são conhecidos, a variância de $\underset{\sim}{Y}_{\ j}$ é:

$$\begin{aligned}\mathbf{V} = Var(\underset{\sim}{Y}_{\ j}) &= Var(\mathbf{Z}_j\underset{\sim}{\gamma}\ ) + Var(\mathbf{X}_j\underset{\sim}{u}_{\ j}) + Var(\underset{\sim}{e}_{\ j}) \\ &= \mathbf{X}_j'\mathbf{T}_{(\underset{\sim}{u}_{\ j})}\mathbf{X}_j + \sigma^2\mathbf{I}.\end{aligned}$$

Não tendo conhecimento do vetor de efeitos aleatórios que está associado com uma determinada observação, é necessário estimá-lo mas, em modelos de componentes de variância estima-se a variância do vetor de efeitos aleatórios $\underset{\sim}{u}_{\ j}$.

Fazendo o processo de estimação por máxima verossimilhança, tem-se que a densidade conjunta de $\underset{\sim}{Y}_{\ j}$ e $\underset{\sim}{u}_{\ j}$ é dada por:

$$\begin{aligned}f(\underset{\sim}{Y}_{\ j}, \underset{\sim}{u}_{\ j}) &= g(\underset{\sim}{Y}_{\ j}, \underset{\sim}{u}_{\ j})\, h(\underset{\sim}{u}_{\ j}) \\ &= c\, exp\Big[-\frac{1}{2}(\underset{\sim}{Y}_{\ j} - \mathbf{Z}_j\underset{\sim}{\gamma}\ - \mathbf{X}_j\underset{\sim}{u}_{\ j})'(\sigma^2\mathbf{I})^{-1}(\underset{\sim}{Y}_{\ j} - \mathbf{Z}_j\underset{\sim}{\gamma}\ - \mathbf{X}_j\underset{\sim}{u}_{\ j})\Big] \\ &\quad exp\Big[-\frac{1}{2}(\underset{\sim}{u}_{\ j})'\mathbf{T}_{(\underset{\sim}{u}_{\ j})}\underset{\sim}{u}_{\ j})\Big] \\ &= c\, exp(G),\end{aligned}$$

onde c é uma constante e

$$G = -\frac{1}{2}\Big[\underset{\sim}{Y}'_j(\sigma^2\mathbf{I})^{-1}\underset{\sim}{Y}_j - 2\underset{\sim}{Y}'_j(\sigma^2\mathbf{I})^{-1}\mathbf{Z}_j\underset{\sim}{\gamma} - 2\underset{\sim}{Y}'_j(\sigma^2\mathbf{I})^{-1}\mathbf{X}_j\underset{\sim}{u}_j + \underset{\sim}{\gamma}'\mathbf{Z}'_j(\sigma^2\mathbf{I})^{-1}\mathbf{Z}_j\underset{\sim}{\gamma}$$

$$+2\underset{\sim}{\gamma}'\mathbf{Z}'_j(\sigma^2\mathbf{I})^{-1}\mathbf{X}_j\underset{\sim}{u}_j + \underset{\sim}{u}'_j\mathbf{X}'_j(\sigma^2\mathbf{I})^{-1}\mathbf{X}_j\underset{\sim}{u}_j + \underset{\sim}{u}'_j\mathbf{T}_{(\underset{\sim}{u}_j)}{}^{-1}\underset{\sim}{u}_j\Big]$$

Igualando a zero as derivadas de G com respeito a $\underset{\sim}{\gamma}$ e $\underset{\sim}{u}_j$, podem ser obtidas as seguintes equações:

$$\mathbf{Z}'_j(\sigma^2\mathbf{I})^{-1}\underset{\sim}{Y}_j = \mathbf{Z}'_j(\sigma^2\mathbf{I})^{-1}\mathbf{Z}_j\underset{\sim}{\gamma} + \mathbf{Z}'_j(\sigma^2\mathbf{I})^{-1}\mathbf{X}_j\underset{\sim}{u}_j. \tag{3.44}$$

$$\mathbf{X}'_j(\sigma^2\mathbf{I})^{-1}(\underset{\sim}{Y}_j - \mathbf{Z}_j\underset{\sim}{\gamma}) = (\mathbf{X}'_j(\sigma^2\mathbf{I})^{-1}\mathbf{X}_j + \mathbf{T}_{(\underset{\sim}{u}_j)}{}^{-1})\underset{\sim}{u}_j. \tag{3.45}$$

Diferenciando pela segunda vez verifica-se que pode ser obtido um ponto de máximo, resolvendo as equações (3.44) e (3.45). Solucionando (3.45) em função de $\underset{\sim}{u}_j$ e substituindo em (3.44) tem-se que

$$\mathbf{Z}'_j[(\sigma^2\mathbf{I})^{-1} - (\sigma^2\mathbf{I})^{-1}\mathbf{X}_j(\mathbf{X}'_j(\sigma^2\mathbf{I})^{-1}\mathbf{X}_j + \mathbf{T}_{(\underset{\sim}{u}_j)}{}^{-1})^{-1}\mathbf{X}'_j(\sigma^2\mathbf{I})^{-1}]\underset{\sim}{Y}_j =$$

$$\mathbf{Z}'_j[(\sigma^2\mathbf{I})^{-1} - (\sigma^2\mathbf{I})^{-1}\mathbf{X}_j(\mathbf{X}'_j(\sigma^2\mathbf{I})^{-1}\mathbf{X}_j + \mathbf{T}_{(\underset{\sim}{u}_j)}{}^{-1})^{-1}\mathbf{X}'_j(\sigma^2\mathbf{I})^{-1}]\mathbf{Z}_j\underset{\sim}{\gamma}\ .$$

Fazendo $\mathbf{W}_j = [(\sigma^2\mathbf{I})^{-1} - (\sigma^2\mathbf{I})^{-1}\mathbf{X}_j(\mathbf{X}'_j(\sigma^2\mathbf{I})^{-1}\mathbf{X}_j + \mathbf{T}_{(\underset{\sim}{u}_j)}{}^{-1})^{-1}\mathbf{X}'_j(\sigma^2\mathbf{I})^{-1}]$,

segue que:

$$\mathbf{Z}'_j\mathbf{W}_j\underset{\sim}{Y}_j = \mathbf{Z}'_j\mathbf{W}_j\mathbf{Z}_j\underset{\sim}{\gamma}\ , \tag{3.46}$$

mas,

$\mathbf{W}_j\mathbf{V} = \mathbf{I}$, sendo $\mathbf{V} = \mathbf{X}'_j\mathbf{T}_{(\underset{\sim}{u}_j)}\mathbf{X}_j + \sigma^2\mathbf{I}$.

Logo,

$$\mathbf{W}_j = \mathbf{V}^{-1}.$$

Dessa forma (3.46) pode ser escrita como: $\mathbf{Z}'_j\mathbf{V}^{-1}\underset{\sim}{Y}_j = \mathbf{Z}'_j\mathbf{V}^{-1}\mathbf{Z}_j\underset{\sim}{\gamma}$, então

$$\widehat{\underset{\sim}{\gamma}} = (\mathbf{Z}'_j\mathbf{V}^{-1}\mathbf{Z}_j)^{-1}\mathbf{Z}'_j\mathbf{V}^{-1}\underset{\sim}{Y}_j, \tag{3.47}$$

onde se tem as estimativas de mínimos quadrados generalizados para os efeitos fixos $\underset{\sim}{\gamma}$ (equações de Aitken).

Substituindo-se (3.47) em (3.45), tem-se que:

$$\begin{aligned}
\widehat{\underset{\sim}{u}}_j &= (\mathbf{X'}_j(\sigma^2\mathbf{I})^{-1}\mathbf{X}_j + \mathbf{T}_{(\underset{\sim}{u}_j)}{}^{-1})^{-1}\mathbf{X'}_j(\sigma^2\mathbf{I})^{-1}(\underset{\sim}{Y}_j - \mathbf{Z}_j\widehat{\underset{\sim}{\gamma}}) \\
&= (\mathbf{X'}_j(\sigma^2\mathbf{I})^{-1}\mathbf{X}_j + \mathbf{T}_{(\underset{\sim}{u}_j)}{}^{-1})^{-1}(\mathbf{X'}_j(\sigma^2\mathbf{I})^{-1}\mathbf{X}_j + \mathbf{T}_{(\underset{\sim}{u}_j)}{}^{-1})\mathbf{T}_{(\underset{\sim}{u}_j)}\mathbf{X'}_j\mathbf{V}^{-1}(\underset{\sim}{Y}_j - \mathbf{Z}_j\widehat{\underset{\sim}{\gamma}}) \\
&= \mathbf{T}_{(\underset{\sim}{u}_j)}\mathbf{X'}_j\mathbf{V}^{-1}(\underset{\sim}{Y}_j - \mathbf{Z}_j\widehat{\underset{\sim}{\gamma}}),
\end{aligned}$$

este é o BLUP do vetor de efeitos aleatórios que foi estabelecido por Henderson. Este é um estimador linear nas observações, não viesado com respeito a parte fixa do modelo, sendo ainda o melhor preditor do vetor de efeitos aleatórios $\underset{\sim}{u}_j$ (Lima, 1987).

Capítulo 4

Estimação Intervalar e Testes de Hipóteses

Este capítulo trata da estimação intervalar e dos testes de hipóteses para os efeitos fixos, as componentes de variância e covariância e os efeitos aleatórios do modelo linear hierárquico.

4.1. Estimação Intervalar

Os estimadores, de acordo com Magalhães e Lima (2002), possuem distribuição de probabilidade (por serem variáveis aleatórias), e a estimação intervalar é uma estimativa mais informativa do parâmetro de interesse pois inclui uma medida de precisão do valor obtido.

Em se tratando do modelo com apenas o intercepto aleatório ou análise de variância com efeitos aleatórios definido em (2.1), a precisão (C) de $\widehat{\gamma}_{00}$ (estimador não viesado de γ_{00}) é a soma das precisões, que é dado por:

$$C(\widehat{\gamma}_{00}) = \sum \Delta_j^{-1}. \tag{4.1}$$

A variância de $\widehat{\gamma}_{00}$ é o inverso de sua precisão:

$$Var(\widehat{\gamma}_{00}) = \Big(\sum \Delta_j^{-1} \Big)^{-1}. \tag{4.2}$$

Desta forma o intervalo com $(1-\alpha)\%$ de confiança para γ_{00} é dado por:

$$(1-\alpha)\%\text{IC}(\gamma_{00}) = \widehat{\gamma}_{00} \pm z_0 \Big(\sum \Delta_j^{-1} \Big)^{-1/2}, \tag{4.3}$$

onde z_0 é o valor crítico da distribuição normal tal que $P(z > z_0) = P(z < -z_0) = \alpha/2$, onde α é o nível de significância e (1 - α) é o nível de confiança.

Na estimação intervalar do modelo de regressão, com médias como respostas, dado em (2.5), a variância amostral do estimador $\widehat{\gamma}_{01}$, dado Δ_j, é:

$$Var(\widehat{\gamma}_{01}) = \Big[\sum \Delta_j^{-1} (W_j - \overline{W}_{\cdot}^{\,*})^2 \Big]^{-1}. \tag{4.4}$$

Assim, um intervalo com $(1-\alpha)\%$ de confiança para γ_{01} é:

$$(1-\alpha)\%\text{IC}(\gamma_{01}) = \widehat{\gamma}_{01} \pm z_0 \Big[Var(\widehat{\gamma}_{01}) \Big]^{-1/2}. \tag{4.5}$$

A estimação intervalar do modelo de regressão de coeficientes aleatórios apresentado em (2.10) é similar a do modelo de regressão com médias como respostas, sendo que a variância amostral do estimador $\widehat{\gamma}_{10}$, dado Δ_j, é da forma:

$$Var(\widehat{\gamma}_{10}) = \Big[\sum \Delta_j^{-1} (\overline{X}_{\cdot j} - \overline{X}_{\cdot\cdot}^{\,*})^2 \Big]^{-1}. \tag{4.6}$$

Então, o intervalo com $(1-\alpha)\%$ de confiança para γ_{10} é dado por:

$$(1-\alpha)\%\text{IC}(\gamma_{10}) = \widehat{\gamma}_{10} \pm z_0 \Big[Var(\widehat{\gamma}_{10}) \Big]^{-1/2}, \tag{4.7}$$

onde z_0 e α são dados como em (4.3).

Levando em consideração agora o modelo de regressão mais geral apresentado em (3.21), onde supõe-se que existem $q\,(q = 1, \ldots, Q)$ variáveis explicativas no nível 1 e $p\,(p = 1, \ldots, P)$ variáveis explicativas no nível 2, têm-se que a região de confiança para o vetor de parâmetros $\underset{\sim}{\gamma}$ (ver 3.25) é baseado nos elementos da diagonal da matriz de dispersão do estimador $\widehat{\underset{\sim}{\gamma}}$, $(\mathbf{V}_{\widehat{\underset{\sim}{\gamma}}})$, sendo esta matriz definida como abaixo:

$$\mathbf{V}_{\widehat{\underset{\sim}{\gamma}}} = Var(\widehat{\underset{\sim}{\gamma}}) = \Big(\sum \mathbf{W}_j' \boldsymbol{\Delta}_j^{-1} \mathbf{W}_j \Big)^{-1}. \tag{4.8}$$

Então um intervalo com $(1-\alpha)\%$ de confiança para um determinado elemento γ_{qp}, é dado por:

$$(1-\alpha)\%\text{IC}(\gamma_{qp}) = \widehat{\gamma}_{qp} \pm z_0 \, (V_{qq})^{-1/2}, \tag{4.9}$$

onde V_{qq} é o q-ésimo elemento diagonal de $\mathbf{V}_{\widehat{\underset{\sim}{\gamma}}}$, z_0 e α são dados em (4.3).

Estimação intervalar para os parâmetros aleatórios pode ser vista em Goldstein (1999).

4.2. Testes de Hipóteses

Respostas afirmativas ou negativas para questões importantes são necessárias em atividades nas ciências, indústrias e na vida de um modo geral. Para tentar obter essas respostas, são construídos experimentos cujos resultados têm alguma referência nas questões de interesse. O processo que determina se as respostas desses experimentos são afirmativas ou negativas é chamado de teste de hipótese (Bickel e Doksum, 1976).

De acordo com Sullivan *et al* (1999), a estatística de teste t é o procedimento normalmente usado em modelos lineares hierárquicos para testar hipóteses dos efeitos fixos e aleatórios, enquanto que as componentes de variância e covariância utilizam a estatística de Wald e a estatística χ^2. Os autores chamam a atenção para o fato dos testes serem válidos assintoticamente, em particular para componentes de variância e covariância, requerendo desta forma cuidado na interpretação, principalmente, quando o número de grupos (unidades do nível 2) for pequeno.

O Quadro 1 adiante inserido, extraído de Bryk e Raudenbush (1992) mostra que há seis tipos de hipóteses que podem ser testadas nos modelos lineares hierárquicos.

Quadro 1. Tipos de hipóteses testadas nos modelos hierárquicos e os respectivos testes

tipos de hipóteses	Efeitos fixos	Coeficientes aleatórios do nível 1	Componentes de variância
Uniparamétrico			
H_0	$\gamma_{qp} = 0$	$\beta_{qj} = 0$	$\tau_{qq} = 0$
H_1	$\gamma_{qp} \neq 0$	$\beta_{qj} \neq 0$	$\tau_{qq} > 0$
Teste	teste-t	teste-t	χ^2 ou teste-z
Multiparamétrico			
H_0	$D'\underset{\sim}{\gamma} = 0$	$D'\underset{\sim}{\beta} = 0$	$\mathbf{T} = T_0$
H_1	$D'\underset{\sim}{\gamma} \neq 0$	$D'\underset{\sim}{\beta} \neq 0$	$\mathbf{T} = T_1$
Teste	Assintoticamente χ^2	hipótese linear geral*	razão de verossimilhança(χ^2)

*Um teste de razão de verossimilhança também pode ser usado no caso do método de máxima verossimilhança completo.

Quando por exemplo, deseja-se testar se o parâmetro γ_{10} tem efeito fixo, as hipóteses dos testes serão $H_0 : \gamma_{10} = 0$ e $H_1 : \gamma_{10} \neq 0$ e será aplicado o teste $t_{\frac{\alpha}{2},(J-P-1)}$. Quando deseja-se testar se o parâmetro β_{1j} tem efeito aleatório, deve-se testar as hipóteses $H_0 : \beta_{1j} = 0$ e $H_1 : \beta_{1j} \neq 0$ com estatística de teste com distribuição $t_{\frac{\alpha}{2},(Q-1)}$.

4.2.1. Testes de Hipóteses para Efeitos Fixos

Considerando o teste para um único parâmetro, a hipótese nula de interesse é dada por:

$$H_0 : \gamma_{qp} = 0, \tag{4.10}$$

que implica que o efeito de um preditor do nível 2, W_{pj}, sobre um parâmetro particular do nível 1, β_{qj}, é nulo.

A estatística de teste é calculada tomando a razão do estimador de máxima verossimilhança completo (que pode também ser de máxima verossimilhança restrito) pelo seu erro padrão estimado, como apresentado abaixo:

$$t = \frac{\widehat{\gamma}_{qp}}{\sqrt{\widehat{V}ar(\widehat{\gamma}_{qp})}}\,. \tag{4.11}$$

A estatística de teste acima segue uma distribuição *t-Student* com $(J - P - 1)$ graus de liberdade para dados balanceados e para algumas situações de dados não balanceados. O valor estimado desta estatística de teste deve ser comparado com o ponto crítico $t_{\frac{\alpha}{2},(J-P-1)}$ da distribuição *t-Student*.

No caso do teste multiparamétrico suponha que se tem um modelo de intercepto e inclinação como resposta, no qual se quer testar se o efeito das características da variável explicativa W_j são similares, tanto no intercepto quanto na inclinação.

O modelo do nível 2 escrito em notação matricial é:

$$\begin{pmatrix} \beta_{0j} \\ \beta_{1j} \end{pmatrix} = \begin{pmatrix} 1 & W_j & 0 & 0 \\ 0 & 0 & 1 & W_j \end{pmatrix} \begin{pmatrix} \gamma_{00} \\ \gamma_{01} \\ \gamma_{10} \\ \gamma_{11} \end{pmatrix} + \begin{pmatrix} u_{0j} \\ u_{1j} \end{pmatrix} . \tag{4.12}$$

A hipótese a ser testada é que ambos γ_{01} e γ_{11} são nulos. Desta forma, pode-se escrever a hipótese nula como segue:

$$H_0 : \mathbf{D}'\underset{\sim}{\gamma} \ = 0 \, , \tag{4.13}$$

onde $\mathbf{D}' = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$, e de acordo com a hipótese nula,

$$\mathbf{D}'\underset{\sim}{\gamma} \ = \begin{pmatrix} \gamma_{01} \\ \gamma_{11} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} . \tag{4.14}$$

Sendo $\mathbf{\Delta}_j$ a variância conhecida da variável resposta, então a variância amostral de $\widehat{\underset{\sim}{\gamma}}$ é dada por:

$$Var(\widehat{\underset{\sim}{\gamma}} \,) = \left(\sum \mathbf{W}'_j \mathbf{\Delta}_j^{-1} \mathbf{W}_j \right)^{-1} = \mathbf{V}_{\widehat{\underset{\sim}{\gamma}}} \ . \tag{4.15}$$

Deste modo, a variância do vetor contraste, $\mathbf{D}'\widehat{\underset{\sim}{\gamma}}$, é obtida da seguinte forma:

$$Var(\mathbf{D}'\widehat{\underset{\sim}{\gamma}} \,) = \mathbf{D}'\mathbf{V}_{\widehat{\underset{\sim}{\gamma}}} \, \mathbf{D} = \mathbf{V}_{\mathbf{D}} \, . \tag{4.16}$$

Quando se desconhece $\mathbf{V}_{\widehat{\underset{\sim}{\gamma}}}$, pode-se estimá-la por:

$$\widehat{\mathbf{V}}_{\widehat{\underset{\sim}{\gamma}}} \ = \left(\sum \mathbf{W}'_j \widehat{\mathbf{\Delta}}_j^{-1} \mathbf{W}_j \right)^{-1} , \tag{4.17}$$

e uma estatística de teste aproximada para a hipótese nula (4.13) é do tipo:

$$U = \widehat{\underset{\sim}{\gamma}} \, ' \mathbf{D} \mathbf{V}_{\mathbf{D}}^{-1} \mathbf{D}' \widehat{\underset{\sim}{\gamma}} \ , \tag{4.18}$$

que assintoticamente tem distribuição χ^2 sob H_0, com graus de liberdade igual ao número de contrastes testados, ou seja, o número de linhas em $\mathbf{D}'$.

O teste de razão de verossimilhança é uma outra abordagem para o teste multiparamétrico que pode ser usado quando se trabalha com o método de máxima verossimilhança completo, não sendo aplicável para o método de máxima verossimilhança restrito (Bryk e Raudenbush, 1992).

4.2.2. Testes de Hipóteses para Efeitos Aleatórios

Do mesmo modo como no teste para efeitos fixos, o teste de hipótese para efeitos aleatórios pode ser considerado para um ou mais parâmetros.

No caso de um único parâmetro, a hipótese de interesse é da forma:

$$H_0 : \beta_{qj} = 0\,, \tag{4.19}$$

ou mais precisamente:

$$H_0 : u_{qj} = 0\,. \tag{4.20}$$

De maneira análoga ao teste para efeitos fixos, a estatística de teste para efeitos aleatórios é obtida tomando-se a razão do efeito aleatório estimado pela estimativa do seu erro padrão, como dado abaixo:

$$t = \frac{\widehat{u}_{qj}}{\sqrt{\widehat{V}ar(\widehat{u}_{qj})}}\,. \tag{4.21}$$

Esta estatística de teste segue uma distribuição *t-Student* para dados balanceados e para algumas situações de dados não balanceados (Sullivan *et al*, 1999). O valor amostral da estatística de teste acima deve ser comparado com o ponto crítico $t_{\frac{\alpha}{2},(Q-1)}$ da distribuição *t-Student*.

De acordo com Sullivan *et al* (1999), a estimativa do erro padrão ($\sqrt{\widehat{V}ar(\widehat{u}_{qj})}$) é maior quando se usa o método de máxima verossimilhança restrito do que quando se usa o método de máxima verossimilhança completo, especialmente se o número de unidades do nível 2, J, é pequeno.

No caso multiparamétrico, sendo $\underset{\sim}{\beta}$, o vetor de parâmetros aleatórios com dimensão $J(Q+1)$x1, então a hipótese linear geral associada à $\underset{\sim}{\beta}$ é:

$$H_0 : \mathbf{D}'\underset{\sim}{\beta} = 0\,. \tag{4.22}$$

Se o vetor de parâmetros pode ser estimado por mínimos quadrados ordinários (MQO), então a hipótese linear geral (4.22) pode ser testada através da estatística:

$$H_{MQO} = \widehat{\underset{\sim}{\beta}}' \mathbf{D}(\mathbf{D}'\widehat{\mathbf{V}}\mathbf{D})^{-1}\mathbf{D}'\widehat{\underset{\sim}{\beta}}\ , \tag{4.23}$$

onde $\widehat{\mathbf{V}}$ é uma matriz bloco diagonal com cada bloco de dimensão $(Q+1)$x$(Q+1)$ e igual a:

$$\widehat{\mathbf{V}}_j = \widehat{\sigma}^2(\mathbf{X}_j'\mathbf{X}_j)^{-1}\,. \tag{4.24}$$

$\widehat{\mathbf{V}}_j$ é a estimativa da matriz de dispersão dada em (3.23).

4.2.3. Testes de Hipóteses para Componentes de Variância e Covariância

Para testar se os coeficientes do nível 1 variam aleatoriamente, a hipótese nula a ser testada é:

$$H_0 : \tau_{qq} = 0\,, \tag{4.25}$$

sendo $\tau_{qq} = Var(\beta_{qj})\,.$

Uma possibilidade para testar a hipótese (4.25) é encontrando as estimativas dos parâmetros por mínimos quadrados ordinários. Este procedimento só é recomendável caso todos os J grupos, ou pelo menos a maioria deles, tenha número de observações (n_j) suficientes. Deste modo, considerando o modelo:

$$\beta_{qj} = \gamma_{q0} + \sum \gamma_{qp} W_{pj}\,, \tag{4.26}$$

a estatística de teste

$$\frac{\sum(\widehat{\beta}_{qj} - \widehat{\gamma}_{q0} - \sum \widehat{\gamma}_{qp} W_{pj})^2}{\widehat{V}_{qqj}} \tag{4.27}$$

tem distribuição aproximada χ^2 com $J - P - 1$ graus de liberdade.

Uma outra possibilidade de testar a hipótese (4.25) é baseada na estimativa de máxima verossimilhança do erro padrão de $\widehat{\tau}_{qq}$ calculada pelo inverso da matriz de informação, onde a estatística de teste

$$z = \frac{\widehat{\tau}_{qq}}{\sqrt{\widehat{V}ar(\widehat{\tau}_{qq})}} \tag{4.28}$$

tem distribuição aproximadamente normal, para grandes amostras.

Como citado por Natis (2000) esta aproximação normal é muito fraca, principalmente quando τ_{qq} é próximo de zero, e o teste para $\widehat{\tau}_{qq}$ baseados em intervalos de confiança simétricos pode levar a conclusões incorretas. Giampaoli (1999) traz uma discussão mais aprofundada sobre este assunto.

No caso multiparamétrico a forma mais geral de teste de hipótese para componentes de variância e covariância é o teste da razão de verossimilhança, onde a hipótese nula a ser testada é:

$$H_0 : \mathbf{T} = \mathbf{T}_0 \tag{4.29}$$

contra a alternativa:

$$H_1 : \mathbf{T} = \mathbf{T}_1\,, \tag{4.30}$$

sendo $\mathbf{T}_0$ uma forma reduzida de $\mathbf{T}_1$.

O teste da razão de verossimilhança é baseado na diferença das *deviances* do modelo, d_0 (supondo H_0) e d_1 (supondo H_1), sendo a *deviance* uma medida de ajuste do modelo que é igual a -2 vezes o valor da função de log-verossimilhança avaliada no máximo da função.

$$H = d_0 - d_1 \,. \tag{4.31}$$

Esta estatística tem, assintoticamente, uma distribuição χ^2 com M graus de liberdade, onde M é a diferença do número de componentes de variância e covariância estimados nos dois modelos.

Capítulo 5

5. Técnicas de Diagnóstico

A verificação de possíveis desvios dos presupostos feitos para o modelo é uma etapa muito importante na análise de um modelo de regressão, segundo Paula (2004). Esta etapa de análise de diagnóstico iniciou-se na análise de um único modelo de regressão com a detecção de pontos extremos, através da análise de resíduos, e também a verificação da adequação da distribuição proposta para a variável resposta.

Algumas técnicas de diagnóstico do modelo foram sendo desenvolvidas ao longo dos anos. A padronização de resíduos para o caso normal linear foi discutida inicialmente por Belsley *et al* (1980) seguida da proposta de Pregibon (1981) da componente do desvio como resíduo na classe dos modelos lineares generalizados, da proposta de Atkinson (1985) da construção de uma banda de confiança para os resíduos da regressão normal linear por simulação de Monte Carlo, dentre outras.

5.1. Estimação dos Resíduos

Em um modelo de regressão linear simples tal como apresentado na seção (2.1), tem-se um único termo de resíduo representado por e_i $(i = 1, \ldots, n)$ mas, considerando um modelo linear hierárquico existirão vários resíduos nos diferentes níveis de hierarquia (Goldstein, 1999).

Levando em consideração os parâmetros estimados de um certo modelo, para estimar um determinado resíduo, u_{0j}, em um modelo de componentes de variância no nível 2, tem-se que para cada unidade do nível 2:

$$\widehat{u}_{0j} = \mathrm{E}(u_{0j}|\underset{\sim}{Y}\ , \widehat{\underset{\sim}{\beta}}\ , \widehat{\mathbf{\Delta}}), \tag{5.1}$$

onde $\widehat{u}_{0j}$ é o resíduo predito calculado através das estimativas dos parâmetros no modelo.

Desconsiderando a variação amostral existente nas estimativas dos parâmetros da equação (5.1), tem-se:

$$\begin{aligned} Cov(\widehat{e}_{ij}, u_{0j}) &= Var(u_{0j}) = \tau_{00} \\ Cov(\widehat{e}_{ij}, e_{ij}) &= Var(e_{ij}) = \sigma^2 \\ Var(\widehat{e}_{ij}) &= \tau_{00} + \sigma^2, \end{aligned} \tag{5.2}$$

onde $\widehat{e}_{ij}$ é o resíduo da i-ésima unidade do j-ésimo grupo.

A equação (5.1) é considerada como uma regressão linear de u_{0j} no conjunto de $\widehat{\underset{\sim}{e}}_{\ j}$ (vetor

de resíduos) para a j-ésima unidade do nível 2, e (5.2) define as expressões necessárias para estimar os coeficientes de regressão e consequentemente $\widehat{u}_{0j}$.

Considerando um modelo simples de componentes de variância de dois níveis, sem variáveis explicativas como:

$$Y_{ij} = \gamma_{00} + u_{0j} + e_{ij}, \tag{5.3}$$

então, as estimativas do nível 2 de u_{0j} podem ser dadas por:

$$\widehat{u}_{0j} = \frac{n_j \tau_{00}}{(n_j \tau_{00} + \sigma^2)} \widehat{e}^*_{.j} \tag{5.4}$$

com $\widehat{e}^*_{.j} = \frac{(\sum \widehat{e}_{ij})}{n_j}$, sendo a média dos resíduos para o j-ésimo grupo do nível 2 e n_j o número de unidades do nível 1 existentes no j-ésimo grupo do nível 2 ($\widehat{e}_{ij}$ é o resíduo da i-ésima unidade do j-ésimo grupo).

Os resíduos estimados são consistentes, mas não são incondicionalmente não viesados. O termo que multiplica a média dos resíduos do grupo j é tido como um fator de diminuição, uma vez que sempre é um valor menor ou no máximo igual a um. A medida que n_j cresce este fator tende a um, porém quando se diminui o número de unidades no nível 1 em um determinado grupo j, há um decréscimo no fator de diminuição de u_{0j}, tornando o fator próximo de zero.

Os resíduos podem ter duas interpretações, a básica que os define como variáveis aleatórias com uma distribuição cujos valores dos parâmetros dizem respeito a variação existente entre as unidades do nível 2, e que provê estimativas eficientes para os coeficientes fixos. A outra interpretação é que são estimativas para cada grupo do nível 2, onde se supõe que pertencem a uma população de grupos, para predizer seus valores.

Tal como nos modelos de regressão linear simples, pode-se usar as estimativas residuais no modelo linear hierárquico para checar as suposições do modelo. As suposições mais importantes que podem ser verificadas são: a suposição de normalidade e a suposição de variância constante. Como as variâncias das estimativas residuais dependem, em geral, dos coeficientes fixos é comum a padronização dos resíduos dividindo-os pelos respectivos erros padrão. Gráficos como o QQ-*plot*, o histograma e dos resíduos contra o número de observações existentes no nível 1, ajudam na verificação de tais suposições. Há também o gráfico envelope proposto por Atkinson (1985).

Quando os resíduos em um nível mais alto têm valores elevados devem ser feitas estimativas intervalar e testes de significância, bem como estimativas pontuais, para expressar a importância dos mesmos.

5.2. Pontos Extremos

A detecção de observações extremas, ou seja, pontos que interferem de forma desproporcional nas estimativas dos parâmetros do modelo, é também um tópico importante na análise de diagnóstico do modelo. A exclusão de tais pontos é a técnica mais conhecida para verificar o impacto da retirada desses pontos nas estimativas da regressão.

A forma como estes pontos extremos se posicionam em relação aos demais pontos torna-os diferentes, pois podem afetar ou não as estimativas do modelo. Um ponto extremo afastado dos demais pontos cuja exclusão altera apenas o intercepto (isto é, os valores ajustados) é chamado de ponto aberrante. O ponto extremo que está mais afastado do subespaço gerado pelas colunas da matriz de preditores, cuja eliminação não altera muito as estimativas dos parâmetros é chamado de ponto de alavanca. Este ponto altera as variâncias dos valores ajustados dos pontos que estão próximos a ele. Já o ponto extremo chamado de influente é afastado dos demais pontos de forma que altera a estimativa da inclinação da reta ajustada.

De acordo com Paula (2004), várias propostas relacionadas com a influência dos pontos nas estimativas dos coeficientes do modelo normal linear surgiram na década de 70. Observações muito menores ou muito maiores que parecem bastante diferentes do restante do conjunto de dados, podem ser observações extremas.

A distância de Cook (Cook, 1977), que foi primeiramente criada para modelos normais lineares, é uma medida tradicional de deleção de pontos influentes como o método desenvolvido por Belsley *et al* (1980) chamado DFFITS.

A deleção individual de pontos influentes pode acarretar um problema denominado *masking effect*, que é a não detecção de pontos conjuntamente influentes. A área de diagnóstico em regressão teve uma grande inovação quando Cook em 1986, ao invés de avaliar a influência da retirada individual ou conjunta de pontos, propôs a avaliação da influência conjunta das observações sob pequenas mudanças no modelo (Paula, 2004).

A detecção destes tipos de pontos num ajuste de regressão deve ser examinada com cautela antes de decisões como a exclusão dos mesmos. Deve-se buscar razões que expliquem o comportamento desses pontos atípicos, pois pode ajudar a entender melhor a relação entre as variáveis explicativas e o fenômeno sob investigação.

Para análise de regressão comum há uma vasta literatura sobre detecção e tratamento de pontos influentes, mas segundo Langford e Lewis (1998), técnicas de exploração de dados, incluindo a detecção de observações influentes são pouco exploradas em modelagem multinível. Em modelos com estrutura hierárquica pode-se desejar saber em qual nível um determinado valor da variável resposta é considerado ponto influente e com respeito a qual variável explicativa.

5.2.1 *Deviance*

A discordância de pontos influentes em modelos linear hierárquico pode ser testada através da estatística *deviance* baseada na razão de log-verossimilhança (Goldstein, 1995). Considerando o modelo de componentes de variância de dois níveis com variação aleatória apenas no intercepto dado por:

$$Y_{ij} = \gamma_{00} + \gamma_{10} X_{ij} + u_{0j} + e_{ij}, \tag{5.5}$$

e um procedimento geralmente aplicado é excluir o grupo que está sendo examinado da parte aleatória do modelo, e incluir um parâmetro fixo separado para este ponto na parte fixa do modelo. Tomando como exemplo o modelo (5.5) com uma variável explicativa, X_{ij}, pode-se excluir um único ponto do nível 2 da parte aleatória do modelo, obtendo:

$$Y_{ij} = (1-h)\gamma_{00} + \gamma_{10} X_{ij} + hc + (1-h)u_{0j} + e_{ij}, \tag{5.6}$$

onde $h = 1$, quando for o ponto excluído e $h = 0$, caso contrário. Desta forma, a contribuição do ponto que está sendo examinado foi removida da parte aleatória do modelo no nível 2, sem excluí-lo do nível 1, isto é, foi ajustado um parâmetro para o intercepto separado (c), para o ponto investigado. De forma similar, pontos isolados podem ser omitidos da parte aleatória do modelo no nível 1 ou em ambos os níveis, por procedimento semelhante.

Considerando que a hipótese alternativa apresentada pelo modelo (5.6), que omite o ponto examinado da parte aleatória do modelo, contra a hipótese nula que o modelo adequado é o apresentado em (5.5), então a mudança na *deviance* é dada por:

$$D_{01} = -2\ln(\lambda_0/\lambda_1), \tag{5.7}$$

onde λ_0 e λ_1 são as verossimilhanças para as hipóteses nula e alternativa, respectivamente. Deste modo, o ponto omitido está no nível mais alto do modelo (nível 2), e a estatística de teste D_{01} pode ser comparada com uma distribuição χ^2 com 1 grau de liberdade (McCullagh e Nelder, 1989).

Para grandes amostras, Goldstein (1995) indica que uma alternativa é aplicar a estatística de Wald, usando uma matriz contraste, para testar que um certo número de parâmetros é diferente de zero, de forma significante.

5.2.2 Pontos de Alavanca ou de Alto *Leverage*

Os pontos de alavanca, como já mencionado, exercem um peso desproporcional no valor ajustado. Um ponto de alavanca é um ponto que possui um alto *leverage*, caso o valor *leverage* seja um valor pequeno, indica que a amostra em questão influencia pouco na construção do modelo.

Para calcular os valores *leverage* para o modelo linear hierárquico, foi considerado nesse trabalho o caso particular para o nível 2. Tomando os resíduos preditos para o q-ésimo ($q = 0, \ldots, Q$) coeficiente aleatório do nível 2 do modelo, o resíduo studentizado é definido como:

$$u'_{qj} = \frac{\widehat{u}_{qj}}{SD(\widehat{u}_{q.})},$$

onde $SD(\widehat{u}_{q.})$ é o desvio padrão dos resíduos dos J grupos do nível 2.

De acordo com Langford e Lewis (1998), como até mesmo em grandes amostras pode haver poucas observações em um ou mais níveis do modelo, é importante calcular os resíduos (e deletar os valores discrepantes) para remover o viés. Considerando que o número de parâmetros fixos existentes no modelo é muito menor que o número de observações existentes no nível 1, a deleção residual pode ser escrita como:

$$u^*_{qj} = u'_{qj}\Big(\frac{J - 1 - u'^2_{qj}}{J - 2}\Big)^{-1/2},$$

onde J é o número de grupos do nível 2.

Ao se deletar observações em um determinado nível do modelo, as estimativas encontradas em outro nível do modelo sofrem influência, por isso Langford e Lewis (1998) sugerem iniciar o processo de detecção de pontos extremos do nível mais alto retirando as unidades dentre os J grupos que apresentam resíduos muito elevados. Depois esse mesmo procedimento deve ser aplicado às unidades do nível mais baixo, dessa forma os pontos extremos serão detectados.

Pode-se desejar examinar a distribuição dos resíduos entre as observações de um determinado nível, ou dentro de um determinado grupo. Por exemplo, pode-se estar interessado em examinar a distribuição dos resíduos da escola em torno das estimativas dos parâmetros fixos, e também a distribuição dos resíduos dos estudantes dentro de uma determinada escola. Por isso Langford e Lewis (1998) sugerem observar os gráficos dos resíduos bem como medidas de assimetria e curtose da amostra, para verificar a normalidade.

Os valores *leverage* no modelo linear hierárquico podem ser calculados de modo análogo ao modelo de regressão linear simples, usando a matriz chapéu ou matriz de projeção (H), que projeta os valores ajustados do modelo nos valores observados. Particularmente, matrizes de projeção para as partes fixas e aleatórias podem ser extraídas separadamente, calculando:

$$H_{\mathbf{X}} = diag\{\mathbf{V}^{-0.5}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-0.5}\},$$

para a parte fixa e

$$H_{\mathbf{W}} = diag\{\mathbf{V}^{*-0.5}\mathbf{W}(\mathbf{W}'\mathbf{V}^{*-1}\mathbf{W})^{-1}\mathbf{W}'\mathbf{V}^{*-0.5}\},$$

para a parte aleatória, onde $\mathbf{V}^*$ é o produto quadrado de Kronecker ($\mathbf{V}^* = \mathbf{V} \otimes \mathbf{V}$), sendo $\mathbf{V} = \mathrm{E}(\underset{\sim}{e}\ \underset{\sim}{e}\ ') + \mathrm{E}(\underset{\sim}{u}\ \underset{\sim}{u}\ ')$.

Para se calcular o valor *leverage* para uma determinada variável p que varia aleatoriamente no nível 2, tem-se:

$$h_{pj} = 1 - \frac{SD(\widehat{u}_{qj})}{\sqrt{\tau_{qq}}}$$

Existe um problema na escala de h_{pj} para compará-lo com algum ponto arbitrário, cujo valor pode ser considerado incomum. Hoaglin e Welsch (1978) sugerem em modelo de regressão linear simples um ponto de corte de $2p/N$ ou $3p/N$, sendo p o número de parâmetros do modelo e $N = \sum n_j$ o número de observações. Mas, é difícil definir exatamente o número de graus de liberdade envolvidos em um modelo de coeficientes aleatórios por causa da correlação existente entre as estimativas dos parâmetros aleatórios do modelo e das covariâncias entre eles. Considerando este problema é melhor comparar os valores *leverage* para cada variável de coeficiente aleatório no modelo em uma escala comum, sendo portanto definido o valor *leverage* padronizado como:

$$h^*_{pj} = \frac{h_{pj}}{\sum_j h_{pj}}, \tag{5.9}$$

para $j = 1, \ldots, J$ grupos do nível 2.

Desta forma, cada valor *leverage* é ponderado por um fator que poderia representar o número de parâmetros no modelo ($tr(H)$) na regressão de mínimos quadrados ordinário. Então, o valor de h^*_{pj} pode ser comparado com o ponto $2/J$ ou $3/J$.

Da mesma forma como para os resíduos os valores *leverage* podem ser observados com respeito a cada variável explicativa permitindo variar aleatoriamente em um nível particular. Isto é importante para identificar grupos (ou observações) incomuns e determinar algum efeito de tratamento. A influência que um certo ponto exerce sobre um coeficiente é definida em termos de seu valor *leverage* como um ponto de alavanca. Usando-se os valores *leverage* padronizados podem ser calculadas medidas de influência como o DFFITS (Belsley *et al*, 1980), que pode ser expresso como:

$$D_{pj} = |u^*_{qj}| \sqrt{\frac{h^*_{pj}}{(1 - h^*_{pj})}},$$

que pode ser comparado com o valor $\frac{2}{\sqrt{J}}$.

Capítulo 6

Aplicação do MLH na Área de Educação

Alguns trabalhos aplicando MLH na área de Educação também foram feitos anteriomente como o realizado por Natis (2000) que aplicou MLH nos dados do SARESP 97, estudando a habilidade (obtida a partir da TRI) em Língua Portuguesa dos alunos da 4ª série de escolas públicas do Estado de São Paulo.

Este trabalho trata da aplicação do MLH aos dados do SAEPE (Sistema de Avaliação Educacional de Pernambuco) do ano de 2002, gentilmente cedido pela Secretaria de Educação do Estado de Pernambuco. O SAEPE foi implantado em Pernambuco no ano 2000 com objetivo de subsidiar uma estratégia de monitoria e de incentivos permanentes centrados na melhoria da qualidade e do desempenho do ensino básico no estado. O procedimento consistiu da aplicação de uma prova do domínio da língua (leitura e escrita) aos alunos da 2ª série do ensino fundamental, e de duas provas (português e matemática) aos alunos da 4ª e 8ª série do ensino fundamental e da 3ª série do ensino médio. Além do desempenho escolar o SAEPE avalia as condições de funcionamento da escola, o nível de eficiência das escolas na promoção de seus alunos, o perfil da direção das unidades escolares, as modalidades de gestão escolar, o perfil dos professores, os recursos pedagógicos utilizados em sala de aula e as características sócio-culturais dos alunos (Pernambuco, 2003).

O presente estudo se restringiu a análise do desempenho dos alunos da 4ª e 8ª série do ensino Fundamental e da 3ª série do ensino médio através das notas (percentual de acertos) que eles obtiveram nas provas de português e matemática.

Foram analisados três bancos de dados, um para cada série e, sabe-se que participaram muito mais alunos da 4ª do que da 8ª e uma quantidade ainda menor da 3ª série do ensino médio, como consta no Quadro 1 abaixo. No presente estudo os alunos que possuiam informações incoerentes e valores *missing* foram excluídos das análises, bem como as escolas em que menos de 30 alunos participaram da avaliação por dificultarem o processo de convergência dos parâmetros na análise de modelo linear hierárquico.

Quadro 1. Número de alunos que responderam as provas.

Série	Matemática	Português
4ª	72504	76250
8ª	42846	44822
3ª	22207	23142

Os questionários preenchidos pelos alunos possuiam muitas informações incoerentes prin-

cipalmente na questão da data de nascimento, isto foi observado com mais frequência dentre os alunos da 4ª série do ensino fundamental. Dessa forma, depois de retirar as informações incoerentes e faltantes dos alunos, o número de observações utilizado na modelagem foi bem menor em todas as séries e, para se ter uma visão melhor da distribuição das notas estudadas, algumas medidas estatística de interesse são apresentadas no Quadro 2 a seguir.

Quadro 2. Medidas estatísticas.

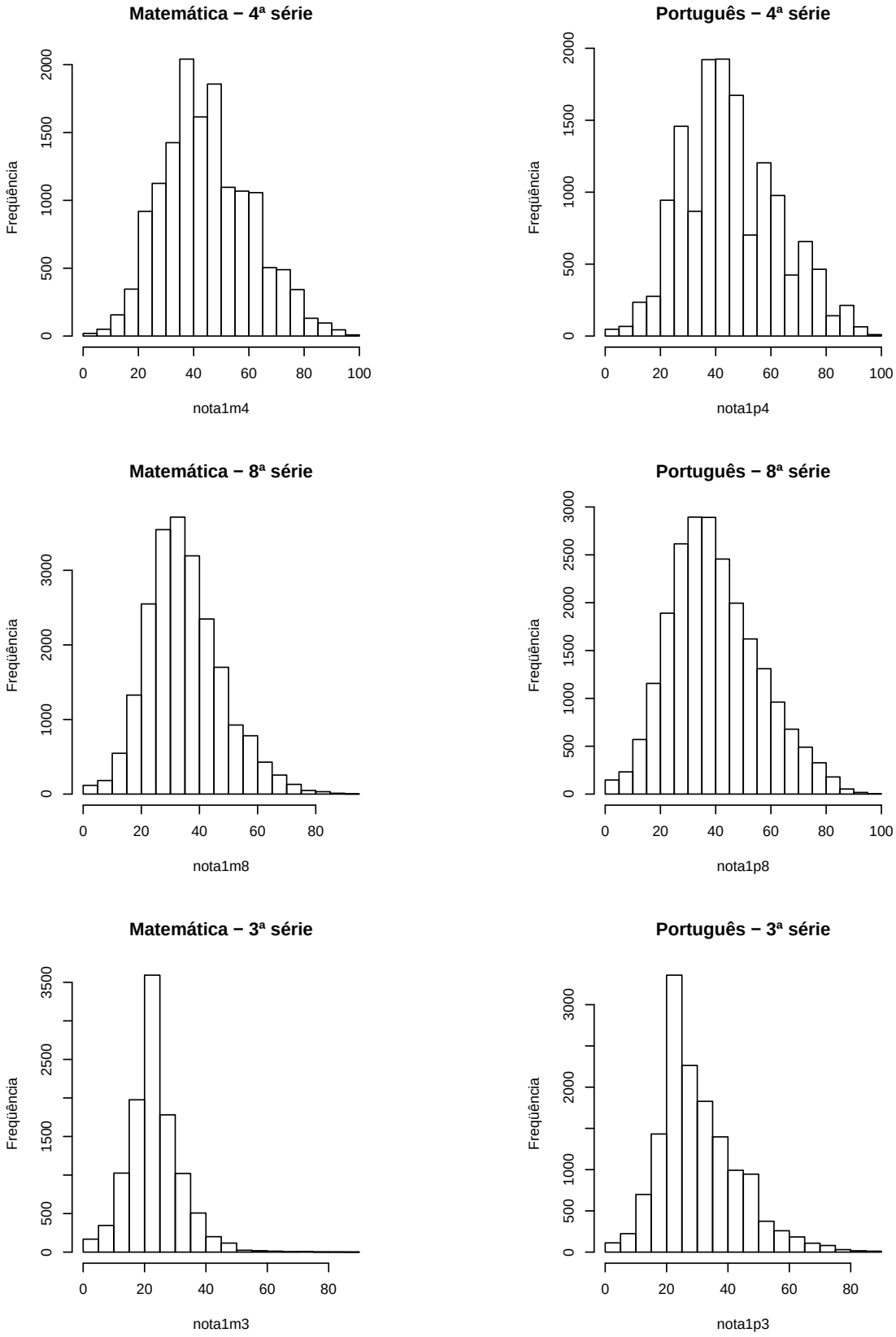
	4ª série		**8ª série**		**3ª série**	
Medidas	matemática	português	matemática	português	matemática	português
nº alunos	14954	15505	22157	23101	13764	14542
nº escolas	362	374	454	465	267	272
média	44,59	45,25	35,03	40,82	23,22	30,64
moda	41,18	41,18	30,77	37,50	22,70	27,30
mediana	42,42	44,12	33,33	40,00	22,73	29,54
DP	16,22	17,49	13,11	16,21	8,71	12,95
assimetria	0,33	0,33	0,47	0,40	0,60	0,82
curtose	-0,24	-0,31	0,47	-0,08	2,03	1,03

Observa-se no Quadro 2 que os alunos da 3ª série do ensino médio apresentam as mais baixas notas, tanto em português (média 30,64) quanto em matemática (média 23,22), quando são observadas as medidas de tendência central, média, moda e mediana.

Este fato é observado em todos os estados brasileiros não significando que os alunos do ensino médio aprendem menos do que os alunos do ensino fundamental, mas apenas que o grau de dificuldade das avaliações cresce com o aumento do nível da série cursada.

A seguir são apresentadas as representações gráficas, através dos histogramas, das notas nas provas em estudo. Analisando os gráficos e comparando os valores de assimetria e curtose observa-se que as notas obtidas pelos alunos da 4ª série em ambas as disciplinas tem um comportamento bem mais próximo da distribuição normal que as notas das demais séries.

Figura 1: Histogramas das notas de português e matemática da $4^{\underline{a}}$, $8^{\underline{a}}$ e $3^{\underline{a}}$ série.



Nos bancos de dados utilizados constavam informações dos alunos, da escola e diretores (questionários no anexo 2). Na modelagem foram testados diversos modelos, incluindo uma variável por vez ou combinando as diversas variáveis disponíveis, tanto as disponíveis no nível 1 (aluno) quanto as disponíveis no nível 2 (escola). Considerando os diversos modelos ajustados para explicar as notas estudadas foram então escolhidos aqueles modelos estatisticamente significativos, cujas variáveis selecionadas em pelo menos um dos modelos ajustados, são as seguintes.

• Variável a ser explicada:

Y_{ij}: nota obtida pelo aluno i da escola j;

• Variáveis do nível 1 (aluno):

acelera$_{ij}$(variável disponível para alunos da 4ª série do ensino fundamental e 3ª série do ensino médio): 1 - aluno na série normal, 0 - aluno com pendência (aluno que tem pendência na série anterior);

sexo$_{ij}$(sexo do aluno): 1 - masculino e 0 - feminino;

difidad$_{ij}$: diferença da idade que o aluno possuia na ocasião da prova e a idade que ele deveria ter para cursar a série (sendo na 4ª série: (idade - 9), na 8ª série: (idade - 13) e na 3ª série do ensino médio: (idade - 16)). Foram considerados os alunos com difidad ≥ -1, ou seja, se difidad $<$ -1 foi considerado erro na informação da idade visto que é bem provável um alunos estar dois anos adiantado em relação a sua idade.

entende$_{ij}$: variável dicotômica que indica se o aluno consegue entender o que o professor ensina, sendo 1 - entende tudo ou quase tudo e 0 - entende pouco ou não entende.

• Variáveis do nível 2 (escola):

tipo$_j$: 1 - escola municipal e 0 - escola estadual;

conserv$_j$: variável dicotômica que indica a situação de conservação da escola sendo 1 - boa conservação e 0 - caso contrário. A definição de escola em bom estado de conservação foi feita a partir da avaliação de 9 aspectos pesquisados (questionário da escola questões: 1 a 9) sendo considerado na escala de Lickert utilizado para a pontuação: bom, regular, ruim e não existe, os respectivos valores 3, 2, 1, 0. A partir da soma dos pontos nos diversos aspectos foi dado um ponto de corte no percentil 75 para indicar a escola em bom estado de conservação (1 - bom e 0 - ruim);

projeto$_j$: variável dicotômica que indica como foi desenvolvido o projeto pedagógico da escola no respectivo ano letivo: 1 - modelo dos professores, ou do diretor, ou da Secretaria de Educação e 0 - a escola não desenvolveu projeto;

idh$_j$: índice de desenvolvimento humano (IDH) do município onde está localizada a

escola (IPEA/FIBGE/FJP, 2000). O IDH é obtido pela média aritmética simples de três índices, referentes às dimensões Logenvidade (IDHM - Logenvidade), Educação (IDHM - Educação) e Renda (IDHM - Renda).

Um resumo de algumas medidas estatísticas das variáveis explicativas que fazem parte de pelo menos um dos modelos selecionados para as observações válidas de todas as escolas estão apresentadas no Quadro 3 a seguir.

Quadro 3. Medidas estatísticas das variáveis dos modelos selecionados para todas as escolas.

	4ª série		**8ª série**		**3ª série**	
Variáveis	matemática	português	matemática	português	matemática	português
idade do aluno (média)	12,35	12,33	17,19	17,24	21,54	21,66
idade do aluno (médiana)	11,84	11,84	16,18	16,22	20,10	20,10
% de alunos na série normal	98,10	98,10	-	-	85,10	85,20
% de alunos do sexo masculino	49,50	48,20	43,50	41,60	42,60	38,90
% de alunos de alunos que entendem a aula	78,10	77,70	61,80	63,30	59,10	62,00
% de alunos de escola municipal	76,10	76,80	34,70	-	-	-
% de escolas em boa conservação	-	-	56,80	56,70	-	-
% de escolas com projeto pedagógico	-	-	-	-	84,40	-

- Esta variável não foi incluída no modelo selecionado para esta disciplina e série.

O Quadro 4 apresenta um resumo de algumas medidas estatísticas das variáveis explicativas que fazem parte de pelo menos um dos modelos selecionados para as escolas em que mais de 30 alunos fizeram as provas.

Uma comparação dos valores constantes nos Quadros 3 e 4 revela que a sub-amostra de alunos considerada na modelagem apresenta características bem semelhantes as do total de alunos pesquisados, com uma diferença maior apenas na variável % de alunos de escola municipal em que os percentuais na sub-amostra serão menores do que os do total pesquisado.

Quadro 4. Medidas estatísticas das variáveis dos modelos selecionados.

	4ª série		**8ª série**		**3ª série**	
Variáveis	matemática	português	matemática	português	matemática	português
idade do aluno (média)	12,09	12,06	17,10	17,13	21,54	21,65
idade do aluno (médiana)	11,59	11,51	16,18	16,09	20,10	20,10
% de alunos na série normal	97,20	97,10	-	-	82,80	83,30
% de alunos do sexo masculino	48,50	48,30	43,60	41,80	42,20	39,00
% de alunos de alunos que entendem a aula	78,20	77,40	60,90	61,80	58,20	61,30
% de alunos de escola municipal	68,00	69,50	25,60	-	-	-
% de escolas em boa conservação	-	-	56,60	56,80	-	-
% de escolas com projeto pedagógico	-	-	-	-	84,30	-

- Esta variável não foi incluída no modelo selecionado para esta disciplina e série.

Observa-se no Quadro 4 que a idade média dos alunos nas duas disciplinas e nas séries estudadas é superior, em no mínimo três anos, a idade adequada para cursar a série. Em todas as séries o maior percentual de alunos é do sexo feminino. Mais da metade deles entende tudo ou quase tudo que o professor ensina em sala de aula e a maioria dos alunos (mais de 68%) da 4ª série estuda em escola municipal.

Os dados foram analisados nos programas HLM e R. O HLM é um programa que foi criado especificamente para análise de modelos hierárquicos, por isso apresenta os resultados de forma organizada separando os efeitos fixos dos efeitos aleatórios na saída dos resultados. Para se trabalhar no HLM é necessário montar um banco de dados para cada nível de hierarquia. Como neste trabalho foram considerados dois níveis de hierarquia então um banco de dados era referente ao nível 1, os alunos, e o outro banco de dados era referente ao nível 2, as escolas. Ambos os arquivos possuem uma variável "código" que os liga na análise dos dados que, no caso, identifica a escola.

O programa R não foi desenvolvido especificamente para análise de modelos hierárquicos como o HLM, desta forma não apresenta os resultados de maneira tão organizada como o HLM, por níveis hierárquicos. Para se trabalhar com modelos hierárquicos no R é necessário solicitar a biblioteca *nlme* que foi desenvolvida especificamente para este tipo de modelagem. Ao contrário do HLM, o R necessita de um único banco de dados, qualquer que seja a

quantidade de níveis que se deseja trabalhar. Neste caso as variáveis explicativas do nível 2, as escolas, são repetidas para todos os alunos da mesma escola, de forma que cada linha deste banco de dados representa um aluno. O R não fornece o nível descritivo (ou p-valor) para os efeitos aleatórios, mas tem a opção de apresentar intervalos de confiança para as estimativas dos parâmetros dos efeitos fixos e para os desvios padrão dos efeitos aleatórios.

Na análise descrita a seguir foram considerados, inicialmente, modelos apenas com o intercepto para verificar se o mesmo poderia ser considerado como aleatório verificando o p-valor. Em seguida, foram incluídas variáveis explicativas apenas no nível 1 para verificar a significância das mesmas, sendo retiradas (uma por vez) aquelas que não apresentaram significância. Por fim, as variáveis significantes no nível 1 foram mantidas no modelo e aos parâmetros que poderiam ser considerados como aleatórios foram acrescentadas (uma após outra) variáveis explicativas no nível 2, mantendo-se aquelas com nível de significância abaixo de 10%. De acordo com Hox (1995) o HLM provê o p-valor como um indicador para testar a significância do parâmetro; a estatística *t-Student* é usada para testar os efeitos fixos, enquanto a estatística χ^2 é usada para testar as componentes de variância.

6.1. Análise das Notas da 4ª Série do Ensino Fundamental

O modelo selecionado para explicar as notas dos alunos da 4ª série que fizeram a prova de matemática foi:

Nível 1:

$$Y_{ij} = \beta_{0j} + \beta_{1j}(acelera_{ij}) + \beta_{2j}(difidad_{ij}) + \beta_{3j}(sexo_{ij}) + \beta_{4j}(entende_{ij}) + e_{ij}$$

Nível 2:

$$\beta_{0j} = \gamma_{00} + \gamma_{01}(tipo_j) + u_{0j}$$
$$\beta_{1j} = \gamma_{10} + u_{1j}$$
$$\beta_{2j} = \gamma_{20}$$
$$\beta_{3j} = \gamma_{30}$$
$$\beta_{4j} = \gamma_{40}$$

Através do resultado do teste para as componentes de variância os parâmetros β_{2j}, β_{3j} e β_{4j} foram considerados como fixos.

O Quadro 5 mostra o resultado das estimativas do modelo hierárquico que melhor se ajustou as notas obtidas pelos alunos de matemática da 4ª série, usando os programas HLM e o R constatando-se a grande semelhança dos valores estimados pelos referidos programas.

Quadro 5. Parâmetros estimados do modelo final para as notas de matemática, 4ª série (1).

Efeitos fixos	**HLM**			**R**			
	Estimativas	**EP**	**p-valor**	**Estimativas**	**EP**	**IC95%**	**p-valor**
β_{0j} - Intercepto1							
γ_{00} - intercepto	36,900	1,203	0,000	36,908	1,155	34,644;39,172	0,000
γ_{01} - tipo	-3,355	0,631	0,000	-3,346	0,592	-4,511;-2,181	0,000
β_{1j} - acelera							
γ_{10} - intercepto	8,904	1,024	0,000	8,891	1,036	6,860;10,921	0,000
β_{2j} - difidad							
γ_{20} - intercepto	-0,954	0,076	0,000	-0,954	0,061	-1,073;-0,835	0,000
β_{3j} - sexo							
γ_{30} - intercepto	2,492	0,241	0,000	2,491	0,254	1,993;2,989	0,000
β_{4j} - entende							
γ_{40} - intercepto	4,219	0,307	0,000	4,219	0,306	3,619;4,818	0,000
Efeitos aleatórios	**Estimativas**	**DP**	**p-valor**		**DP**	**IC95%**	
u_{0j} - intercepto1	25,029	5,003	0,000		5,270	3,409;8,148	
u_{1j} - acelera	11,373	3,374	0,014		3,901	1,793;8,486	
e_{ij} - nível 1	222,818	14,927			14,926	14,752;15,102	

(1) EP: erro padrão; DP: desvio padrão e IC95%: é o intervalo com 95% de confiança.

Através dos dados apresentados no Quadro 5 verifica-se que os alunos da 4ª série em média tiraram 36,900 pontos na prova de matemática quando são mantidas fixas no modelo as variáveis explicativas, enquanto que a média global das notas desses alunos foi de 44,59 pontos com um desvio padrão de ± 16,22 pontos (já referidos no Quadro 2).

Os alunos em aceleração (ou seja, cursam disciplinas de duas séries em um ano) apresentam um desempenho 8,904(EP: 1,024) pontos menor que os alunos da série normal, fixadas as demais variáveis do modelo. Os que estudam em escolas municipais têm suas notas diminuídas em 3,355(EP: 0.631) pontos quando o efeito das demais variáveis explicativas está fixo, ou seja, o desempenho dos alunos nas escolas estaduais é melhor do que nas municipais. Em relação a defasagem idade-série observa-se que a cada ano de diferença da idade do aluno para a idade ideal há um decréscimo de 0,954(EP: 0.076) pontos na nota média por ano de atraso, considerando que outras variáveis explicativas do modelo são mantidas constantes. Isto indica que o aluno que está atrasado em relação a idade normal tem um desempenho médio quase 1 ponto a menos a cada ano de diferença. Os alunos do sexo masculino têm a média acrescida em 2,492(EP: 0.241) pontos em relação aos alunos do sexo feminino, mantidas fixas as demais variáveis explicativas e, aqueles que entendem tudo ou quase tudo que o professor ensina tem nota média 4,219(EP: 0.307) pontos superior em relação aos que entendem pouco ou nada entendem.

Para os alunos da 4ª série o modelo selecionado para modelar as notas de português foi:

Nível 1:

$$Y_{ij} = \beta_{0j} + \beta_{1j}(acelera_{ij}) + \beta_{2j}(difidad_{ij}) + \beta_{3j}(sexo_{ij}) + \beta_{4j}(entende_{ij}) + e_{ij}$$

Nível 2:

$$\beta_{0j} = \gamma_{00} + \gamma_{01}(tipo_j) + u_{0j}$$
$$\beta_{1j} = \gamma_{10} + u_{1j}$$
$$\beta_{2j} = \gamma_{20}$$
$$\beta_{3j} = \gamma_{30}$$
$$\beta_{4j} = \gamma_{40}$$

Através do resultado do teste para as componentes de variância os parâmetros β_{2j}, β_{3j} e β_{4j} foram considerados como fixos.

As variáveis selecionadas foram as mesmas que modelaram as notas de matemática. O resultado das estimativas do modelo hierárquico, usando os programas HLM e R, que melhor se ajustou aos dados está apresentado no Quadro 6.

Quadro 6. Parâmetros estimados do modelo final para as notas de português, 4ª série (1).

Efeitos fixos	**HLM**			**R**			
	Estimativas	**EP**	**p-valor**	**Estimativas**	**EP**	**IC95%**	**p-valor**
β_{0j} - Intercepto1							
γ_{00} - intercepto	37,511	1,288	0,000	37,527	1,224	35,127;39,927	0,000
γ_{01} - tipo	-1,706	0,617	0,006	-1,700	0,614	-2,907;-0,493	0,059
β_{1j} - acelera							
γ_{10} - intercepto	10,204	1,088	0,000	10,186	1,091	8,047;12,325	0,000
β_{2j} - difidad							
γ_{20} - intercepto	-0,807	0,091	0,000	-0,808	0,066	-0,938;-0,678	0,000
β_{3j} - sexo							
γ_{30} - intercepto	-3,508	0,275	0,000	-3,506	0,278	-4,051;-2,961	0,000
β_{4j} - entende							
γ_{40} - intercepto	4,323	0,332	0,000	4,322	0,331	3,673;4,972	0,000
Efeitos aleatórios	**Estimativas**	**DP**	**p-valor**		**DP**	**IC95%**	
u_{0j} - intercepto1	27,457	5,240	0,000		5,393	3,459;8,408	
u_{1j} - acelera	11,671	3,417	0,053		3,853	1,627;9,124	
e_{ij} - nível 1	265,065	16,281			16,280	16,089;16,473	

(1) EP: erro padrão; DP: desvio padrão e IC95%: é o intervalo com 95% de confiança.

A nota média dos alunos da 4ª série na prova de português, quando todas as variáveis explicativas são mantidas constantes é 37,511, enquanto que a média geral é de 45,25 com um desvio padrão de ± 17,49 pontos. Os alunos que estudam em escola municipal diminuem a estimativa média em 1,706(EP: 0,617) pontos em comparação aos que estudam em escola estadual, mantidas as demais variáveis explicativas fixas. Quanto aos alunos que têm defasagem na idade-série estes apresentam um decréscimo na estimativa média de 0,807(EP: 0,091) pontos por ano de atraso quando comparados àqueles que têm idade normal para a série, e tendo as demais variáveis explicativas constantes, aqueles que aprendem tudo ou quase tudo que o professor ensina têm um aumento na estimativa média de 4,323(EP: 0,332) pontos em relação aos que pouco ou nada aprendem.

Observando os resultados apresentados pelo R, as estimativas são próximas às apresentadas pelo HLM. Verifica-se que a variação das notas dentro da escola é de 16,280 variando entre 16,089 a 16,473 pontos com 95% de confiança, enquanto que a variação nos interceptos é de 5,393 variando entre 3,459 e 8,408 pontos, e nas inclinações é de 3,853 variando de 1,627 a 9,124 pontos.

6.2. Análise das Notas da 8ª Série do Ensino Fundamental

O modelo selecionado para as notas dos alunos da 8ª série de matemática foi:

Nível 1:

$$Y_{ij} = \beta_{0j} + \beta_{1j}(difidad_{ij}) + \beta_{2j}(sexo_{ij}) + \beta_{3j}(entende_{ij}) + e_{ij}$$

Nível 2:

$$\beta_{0j} = \gamma_{00} + \gamma_{01}(conserv_j) + u_{0j}$$
$$\beta_{1j} = \gamma_{10} + \gamma_{11}(conserv_j) + u_{1j}$$
$$\beta_{2j} = \gamma_{20} + \gamma_{21}(tipo_j) + u_{2j}$$
$$\beta_{3j} = \gamma_{30}$$

O parâmetro β_{3j} foi considerado como fixo pelo resultado do teste para as componentes de variância.

Na 8ª série não existe o aluno com pendência na série anterior. Com exceção da variável "acelera", as variáveis escolhidas no nível 1 são as mesmas selecionadas nos modelos da 4ª série e, no nível 2, além do tipo da escola foi selecionado também o estado de conservação da mesma.

O resultado das estimativas do modelo hierárquico que melhor se ajustou aos dados está apresentado no Quadro 7.

Quadro 7. Parâmetros estimados do modelo final para as notas de matemática, 8ª série (1).

Efeitos fixos	**HLM**			**R**			
	Estimativas	**EP**	**p-valor**	**Estimativas**	**EP**	**IC95%**	**p-valor**
β_{0j} - Intercepto1							
γ_{00} - intercepto	33,813	0,509	0,000	33,791	0,538	32,737;34,845	0,000
γ_{01} - conserv	2,149	0,684	0,002	2,147	0,691	0,789;3,504	0,002
β_{1j} - difidad							
γ_{10} - intercepto	-0,487	0,052	0,000	-0,489	0,051	-0,588;-0,389	0,000
γ_{11} - conserv	-0,182	0,068	0,008	-0,183	0,068	-0,316;-0,050	0,007
β_{2j} - sexo							
γ_{20} - intercepto	3,529	0,218	0,000	3,542	0,222	3,107;3,977	0,000
γ_{21} - tipo	0,822	0,409	0,044	0,773	0,430	-0,069;1,615	0,072
β_{3j} - entende							
γ_{30} - intercepto	4,887	0,169	0,000	4,888	0,171	4,554;5,222	0,000
Efeitos aleatórios	**Estimativas**	**DP**	**p-valor**		**DP**	**IC95%**	
u_{0j} - intercepto1	28,912	5,377	0,000		5,348	4,727;6,051	
u_{1j} - difidad	0,200	0,447	0,000		0,450	0,378;0,536	
u_{2j} - sexo	3,645	1,909	0,000		1,896	1,446;2,488	
e_{ij} - nível 1	143,114	11,963			11,961	11,845;12,077	

(1) EP: erro padrão; DP: desvio padrão e IC95%: é o intervalo com 95% de confiança.

A média global dos alunos que fizeram a prova de matemática da 8ª série foi de 35,03 pontos com um desvio padrão de ± 13,11 pontos, e a estimativa média apresentada pelo modelo foi de 33,813 pontos, quando todas as variáveis explicativas foram mantidas constantes. A escola em bom estado de conservação eleva a média das notas em 2,149(EP: 0,684) pontos em relação as escolas mal conservadas, fixadas as outras variáveis no modelo. Os alunos do sexo masculino têm seu desempenho elevado em 3,529(EP: 0,218) pontos em comparação aos alunos do sexo feminino e se eles estudam em escola municipal aumenta o desempenho em mais 0,822(EP: 0,409) pontos.

O modelo selecionado para os alunos da 8ª série na prova de português foi:

Nível 1:

$$Y_{ij} = \beta_{0j} + \beta_{1j}(difidad_{ij}) + \beta_{2j}(sexo_{ij}) + \beta_{3j}(entende_{ij}) + e_{ij}$$

Nível 2:

$$\beta_{0j} = \gamma_{00} + \gamma_{01}(conserv_j) + u_{0j}$$
$$\beta_{1j} = \gamma_{10} + u_{1j}$$
$$\beta_{2j} = \gamma_{20}$$
$$\beta_{3j} = \gamma_{30}$$

Os parâmetros β_{2j} e β_{3j} foram considerados como fixos, pelo resultado do teste para as componentes de variância.

Também para o desempenho em português as variáveis escolhidas no nível 1 são as mesmas selecionadas no modelo da 8ª série em matemática e no modelo da 4ª série, com exceção da variável "acelera"; no nível 2 foi significativa apenas a conservação da escola e não mais o tipo da escola. O resultado das estimativas do modelo hierárquico que melhor se ajustou aos dados está apresentado no Quadro 8.

Quadro 8. Parâmetros estimados do modelo final para as notas de português, 8ª série (1).

Efeitos fixos	**HLM**			**R**			
	Estimativas	**EP**	**p-valor**	**Estimativas**	**EP**	**IC95%**	**p-valor**
β_{0j} - Intercepto1							
γ_{00} - intercepto	46,510	0,492	0,000	46,508	0,516	45,496;47,520	0,000
γ_{01} - conserv	0,900	0,449	0,045	0,899	0,450	0,014;1,784	0,046
β_{1j} - difidad							
γ_{10} - intercepto	-0,997	0,041	0,000	-0,997	0,041	-1,078;-0,916	0,000
β_{2j} - sexo							
γ_{20} - intercepto	-3,954	0,206	0,000	-3,954	0,202	-4,349;-3,559	0,000
β_{3j} - entende							
γ_{30} - intercepto	5,587	0,205	0,000	5,587	0,206	5,183;5,991	0,000
Efeitos aleatórios	**Estimativas**	**DP**	**p-valor**		**DP**	**IC95%**	
u_{0j} - intercepto1	46,607	6,827	0,000		6,822	6,081;7,656	
u_{1j} - difidad	0,317	0,563	0,000		0,562	0,475;0,665	
e_{ij} - nível 1	214,844	14,658			14,658	14,520;14,797	

(1) EP: erro padrão; DP: desvio padrão e IC95%: é o intervalo com 95% de confiança.

Os alunos que realizaram a prova de português da 8ª série tiveram uma média de 40,82 pontos com um desvio padrão de ± 16,21 pontos. No modelo apresentado no Quadro 8 a nota média estimada foi 46,510 pontos com as variáveis explicativas assumindo valor zero. Há um acréscimo da ordem de 0,900(EP: 0,449) pontos nesta média quando o estado de conservação da escola é considerado bom e são mantidas constantes as demais variáveis explicativas do modelo. O aluno que tem defasagem na idade-série tem um decréscimo na média de 0,997(EP: 0,041) pontos (a cada ano de defasagem), e para aqueles que entendem tudo ou quase tudo que o professor ensina o aumento médio é de 5,587(EP: 0,205) pontos, considerando o efeito apenas da variável citada e as demais fixas. Avaliando os resíduos a variação das notas dos alunos dentro da escola é de 14,658 pontos, enquanto a variação nos interceptos e inclinações é de 6,827 e 0,563 respectivamente.

6.3. Análise das Notas da 3ª Série do Ensino Médio

O modelo selecionado para os alunos da 3ª série na prova de matemática foi:

Nível 1:

$$Y_{ij} = \beta_{0j} + \beta_{1j}(acelera_{ij}) + \beta_{2j}(difidad_{ij}) + \beta_{3j}(sexo_{ij}) + \beta_{4j}(entende_{ij}) + e_{ij}$$

Nível 2:

$$\beta_{0j} = \gamma_{00}$$
$$\beta_{1j} = \gamma_{10} + u_{1j}$$
$$\beta_{2j} = \gamma_{20} + \gamma_{21}(projeto_j) + u_{2j}$$
$$\beta_{3j} = \gamma_{30}$$
$$\beta_{4j} = \gamma_{40}$$

Pelo resultado do teste para as componentes de variância os parâmetros β_{0j}, β_{3j} e β_{4j} foram considerados como fixos.

As variáveis selecionadas no nível 1 são as mesmas identificadas nos modelos da 4ª série do ensino fundamental e, no nível 2 a variável selecionada corresponde a ocorrência de projeto pedagógico elaborado pelo diretor, enquanto para a 4ª série a variável do nível 2 foi o tipo da escola, e para a 8ª série foi o tipo e mais a conservação da escola.

O resultado das estimativas do modelo hierárquico que melhor se ajustou aos dados está apresentado no Quadro 9.

Quadro 9. Parâmetros estimados do modelo final para as notas de matemática, 3ª série (1).

Efeitos fixos	**HLM**			**R**			
	Estimativas	**EP**	**p-valor**	**Estimativas**	**EP**	**IC95%**	**p-valor**
β_{0j} - Intercepto1							
γ_{00} - intercepto	21,454	0,293	0,000	21,531	0,512	20,528;22,534	0,000
β_{1j} - acelera							
γ_{10} - intercepto	0,758	0,272	0,006	0,774	0,279	0,228;1,320	0,006
β_{2j} - difidad							
γ_{20} - intercepto	-0,102	0,027	0,000	-0,109	0,045	-0,197;-0,021	0,016
γ_{21} - projeto	-0,108	0,030	0,001	-0,102	0,050	-0,199;-0,005	0,039
β_{3j} - sexo							
γ_{30} - intercepto	2,209	0,180	0,000	2,210	0,165	1,886;2,534	0,000
β_{4j} - entende							
γ_{40} - intercepto	2,323	0,182	0,000	2,320	0,166	1,995;2,645	0,000
Efeitos aleatórios	**Estimativas**	**DP**	**p-valor**		**DP**	**IC95%**	
u_{1j} - acelera	5,375	2,318	0,000		2,330	2,011;2,699	
u_{2j} - difidad	0,014	0,119	0,000		0,130	0,091;0,185	
e_{ij} - nível 1	68,805	8,295			8,290	8,178;8,403	

(1) EP: erro padrão; DP: desvio padrão e IC95%: é o intervalo com 95% de confiança.

Os alunos que cursam a 3ª série, estando com alguma pendência na série anterior, têm um decréscimo (mantendo as demais variáveis explicativas constantes) de 0,758(EP: 0,272) pontos na estimativa média, que foi de 21,454 pontos, quando comparados com aqueles que cursam a série normal. O fato do aluno ser do sexo masculino eleva a estimativa média em 2,209(EP: 0,180) pontos e os que entendem tudo ou quase tudo que o professor ensina na sala de aula têm um aumento de 2,323(EP: 0,182) pontos em relação aos alunos que entendem pouco ou nada entendem.

As estimativas dos parâmetros e dos desvios padrão dos efeitos aleatórios obtidos pelo R, também apresentam valores próximos as estimativas obtidas pelo HLM.

O modelo selecionado para explicar as notas dos alunos da 3ª série de português foi:

Nível 1:

$$Y_{ij} = \beta_{0j} + \beta_{1j}(acelera_{ij}) + \beta_{2j}(difidad_{ij}) + \beta_{3j}(sexo_{ij}) + \beta_{4j}(entende_{ij}) + e_{ij}$$

Nível 2:

$$\beta_{0j} = \gamma_{00} + \gamma_{01}(idh_j)$$
$$\beta_{1j} = \gamma_{10} + u_{1j}$$
$$\beta_{2j} = \gamma_{20} + u_{2j}$$
$$\beta_{3j} = \gamma_{30}$$
$$\beta_{4j} = \gamma_{40}$$

Pelo resultado do teste para as componentes de variância os parâmetros β_{3j} e β_{4j} foram considerados como fixos.

O resultado das estimativas do modelo hierárquico que melhor se ajustou aos dados está apresentado no Quadro 10.

Quadro 10. Parâmetros estimados do modelo final para as notas de português, 3ª série (1).

Efeitos fixos	**HLM**			**R**			
	Estimativas	**EP**	**p-valor**	**Estimativas**	**EP**	**IC95%**	**p-valor**
β_{0j} - Intercepto1							
γ_{00} - intercepto	23,562	1,898	0,000	23,611	1,968	19,753;27,470	0,000
γ_{01} - idh	6,365	2,565	0,013	6,181	2,663	0,938;11,423	0,021
β_{1j} - acelera							
γ_{10} - intercepto	1,779	0,366	0,000	1,859	0,373	1,129;2,589	0,000
β_{2j} - difidad							
γ_{20} - intercepto	-0,266	0,024	0,000	-0,266	0,024	-0,313;-0,219	0,000
β_{3j} - sexo							
γ_{30} - intercepto	-1,750	0,215	0,000	-1,746	0,212	-2,161;-1,330	0,000
β_{4j} - entende							
γ_{40} - intercepto	5,280	0,224	0,000	5,279	0,213	4,862;5,696	0,000
Efeitos aleatórios	**Estimativas**	**DP**	**p-valor**		**DP**	**IC95%**	
u_{0j} - intercepto1	7,776	2,789	0,000		2,656	1,658;4,255	
u_{1j} - acelera	7,680	2,771	0,000		2,742	1,691;4,446	
u_{2j} - difidad	0,036	0,191	0,000		0,189	0,138;0,261	
e_{ij} - nível 1	146,739	12,114			12,111	11,969;12,256	

(1) EP: erro padrão; DP: desvio padrão e IC95%: é o intervalo com 95% de confiança.

A média global para a prova de português obtida pelos alunos da 3ª série foi 30,64 com um desvio padrão de ± 12,95 pontos, a estimativa média do modelo foi de 23,562 pontos considerando o efeito das variáveis explicativas nulo. Tanto a defasagem idade-série quanto o fato do aluno ser do sexo masculino diminuem a nota média estimada. Os alunos do sexo masculino apresentam menor desempenho em português que os alunos do sexo feminino. O Índice de Desenvolvimento Humano usado para explicar as notas de português mostrou-se estatisticamente significante. Considerando por exemplo o município de Caetés que possui o menor IDH dentre os municípios da amostra de Pernambuco (0,521) há um acréscimo no desempenho do aluno de 3,316 pontos, e para o município de Fernando de Noronha que possui o maior IDH de Pernambuco (0,862) há um acréscimo no desempenho de 5,487 pontos, quando as demais variáveis explicativas estão fixas no modelo.

Destacando as variáveis do nível 2 pode-se então concluir que a condição de vida dos estudantes medida através do IDH serve para explicar o desempenho de português enquanto a existência de projeto pedagógico do diretor explica o desempenho em matemática, para os alunos da 3ª série do ensino médio.

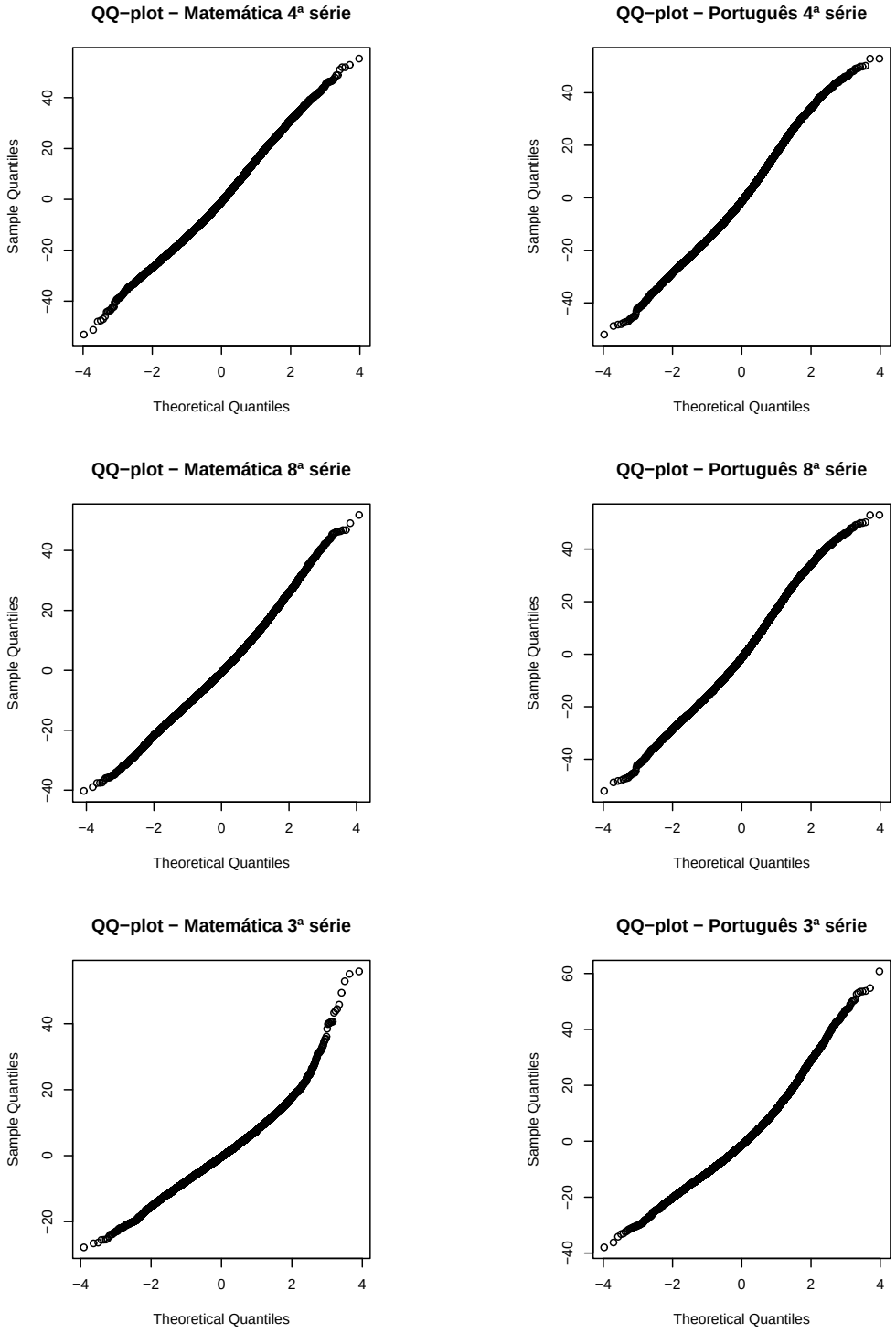
6.4. Análise Gráfica dos Resíduos

Levando em consideração que as estimativas residuais dependem geralmente dos efeitos fixos foram construídos QQ-*plot* dos resíduos dos modelos apresentados nos Quadros numerados de 5 a 10, para averiguar os presupostos de normalidade e de variância constante.

Verificando os gráficos (Figura 2) apresentados pelos 6 modelos observa-se que de um modo geral, os resíduos dos modelos da 4ª e 8ª série do ensino fundamental aparentam estar em conformidade com a suposição de normalidade, mas os resíduos dos modelos selecionados para a 3ª série do ensino médio não aparentam estar em conformidade com esta suposição.

Esta constatação sugere a necessidade de uma transformação na variável dependente do modelo (nota do aluno) a fim de modelar uma estrutura com as presuposições necessárias confirmadas.

Figura 2: QQ-*plot* dos resíduos dos modelos apresentados nos Quadros: 5 a 10.



6.5. Comparação com o Modelo de Mínimos Quadrados Ordinários

Modelos de regressão múltipla MMQO (modelos de mínimos quadrados ordinários) com as mesmas variáveis existentes nos respectivos modelos lineares hierárquicos foram construídos com a finalidade de comparar os coeficientes de determinação (R^2), e as estimativas dos parâmetros e os respectivos erros padrão para os modelos de regressão múltipla MMQO estão apresentados nos Quadros 11, 12 e 13. Em todos os modelos todas as variáveis foram significantes com nível de significância inferior a 1%.

Quadro 11. Modelo de MMQO para a 4ª.

	Matemática		**Português**	
Variáveis	**Estimativas**	**EP**	**Estimativas**	**EP**
β_0 - Intercepto	37,712	0,871	35,115	0,939
β_1 - Acelera	8,767	0,790	9,755	0,849
β_2 - Difidad	-1,092	0,061	-0,926	0,066
β_3 - Sexo	2,547	0,263	3,532	0,286
β_4 - Aprendiz	4,422	0,316	4,575	0,340
β_5 - Tipo	-3,791	0,287	-2,244	0,318

Quadro 12. Modelo de MMQO para a 8ª.

	Matemática		**Português**	
Variáveis	**Estimativas**	**EP**	**Estimativas**	**EP**
β_0 - Intercepto	33,211	0,344	41,899	0,349
β_1 - Difidad	-0,428	0,036	-0,940	0,029
β_2 - Dif*conserv	-0,236	0,048	-	-
β_3 - Sexo	-3,442	0,188	4,103	0,208
β_4 - Sexo*tipo	-0,975	0,287	-	-
β_5 - Aprendiz	5,176	0,174	5,866	0,212
β_6 - Conserv	2,841	0,426	1,149	0,209

Quadro 13. Modelo de MMQO para a 3ª.

Variáveis	Matemática		Português	
	Estimativas	**EP**	**Estimativas**	**EP**
β_0 - Intercepto	21,409	0,269	23,024	1,119
β_1 - Acelera	0,838	0,222	1,747	0,291
β_2 - Difidad	-0,109	0,027	-0,263	0,020
β_3 - Sexo	2,274	0,167	-1,707	0,216
β_4 - Aprendiz	2,432	0,167	5,412	0,216
β_5 - Idh	-	-	7,113	1,454
β_6 - Proj*difidad	-0,103	0,028	-	-

Observando os erros padrão dos modelos hierárquicos (Quadros 5 a 10) e os erros padrão dos modelos de regressão múltipla MMQO (Quadros 11 a 13) verifica-se que eles são maires nos modelos hierárquicos porque os mesmos explicam mais a variabilidade da variável resposta visto que considera a estrutura de variância de cada grupo, o que não ocorre nos modelos de regressão múltipla MMQO onde toda a variabilidade é explicada pelo resíduo.

Os resultados dos R^2 dos modelos de regressão múltipla MMQO apresentados acima e dos modelos hierárquicos apresentados nos Quadros numerados de 5 a 10 são mostrados no Quadro 14. Como os programas estatísticos utilizados não apresentam o resultado do R^2 para os modelos hierárquicos, eles foram calculados pelo procedimento de mínimos quadrados através da fórmula descrita abaixo:

$$R^2 = \frac{SQ_{reg}}{SQ_{tot}}.$$

Sabendo-se que

$$\sum_j \sum_i (Y_{ij} - \overline{Y}_{.j})^2 = \sum_j \sum_i (Y_{ij} - \widehat{Y}_{ij})^2 + 2 \sum_j \sum_i [(Y_{ij} - \widehat{Y}_{ij})(\widehat{Y}_{ij} - \overline{Y}_{.j})] + \sum_j \sum_i (\widehat{Y}_{ij} - \overline{Y}_{.j})^2$$

têm-se que

$2 \sum_j \sum_i [(Y_{ij} - \widehat{Y}_{ij})(\widehat{Y}_{ij} - \overline{Y}_{.j})] + \sum_j \sum_i (\widehat{Y}_{ij} - \overline{Y}_{.j})^2 = SQ_{reg}$ é a soma de quadrados da regressão, e

$\sum_j \sum_i (Y_{ij} - \widehat{Y}_{ij})^2 + 2 \sum_j \sum_i [(Y_{ij} - \widehat{Y}_{ij})(\widehat{Y}_{ij} - \overline{Y}_{.j})] + \sum_j \sum_i (\widehat{Y}_{ij} - \overline{Y}_{.j})^2 = SQ_{tot}$ é a soma de quadrados total.

$\widehat{Y}_{ij}$ é o valor ajustado da variável resposta para o indivíduo i do grupo j.

O Quadro 13 abaixo apresenta os coeficiêntes de determinação para os modelos de regressão múltipla MMQO e para os modelos lineares hierárquicos.

Quadro 14. Coeficientes de Determinação (R^2)

Série	MMQO	Mod. Hierárquico
4ª - Mat	0,0604	0,1626
4ª - Port	0,0529	0,1436
8ª - Mat	0,0884	0,1863
8ª - Port	0,0951	0,1972
3ª - Mat	0,0627	0,1272
3ª - Port	0,0646	0,1399

Verificando o quadro acima nota-se que em todas as situações os valores dos R^2 são cerca de duas vezes superiores nos modelos hierárquicos comparativamente aos modelos de mínimos quadrados ordinários, demonstrando que os modelos hierárquicos são bem melhores que os MMQO.

Capítulo 7

Conclusões

A presente dissertação teve como principal objetivo apresentar o modelo linear hierárquico como um método alternativo para avaliar o desempenho escolar e mostrar que quando se tem um conjunto de dados cuja estrutura seja hierárquica, este tipo de modelagem é mais apropriado.

Analisando os resultados dos bancos de dados fornecidos pela Secretaria de Educação e Cultura do Estado de Pernambuco, observou-se que os alunos da 3ª série do ensino médio obtiveram as menores notas médias nas provas de português e matemática, chegando a ter 47,92% a menos em matemática e 32,29% a menos em português, quando comparados com os alunos da 4ª série do ensino fundamental.

Ao analisar as variáveis do nível 1 dos modelos selecionados pode-se concluir que os alunos que têm pendência na série anterior apresentam sempre um desempenho inferior se comparados aos que cursam a série normal, isto pôde ser constatado tanto na 4ª série do ensino fundamental quanto na 3ª série do ensino médio, em ambas as provas. Os alunos que têm defasagem idade-série, ou seja, que possuem idade além da que deveriam ter para cursar a série, tendem a obter notas menores que aqueles com idade adequada para cursar a série, isto foi observado nas três séries estudadas. Quanto ao sexo dos alunos observa-se que tratando-se de matemática os alunos do sexo masculino apresentam desempenho superior aos do sexo feminino; já em português ocorre o inverso, os alunos do sexo feminino têm melhor desempenho que os do sexo masculino, isto constata-se nas três séries estudadas. Os alunos que entendem pouco ou não entendem a aula dada pelo professor apresentam, nas três séries estudadas, menor desempenho que aqueles que entendem tudo ou quase tudo que é dado pelo professor.

Considerando as variáveis selecionadas para o nível 2, conclui-se que para os alunos da 4ª série do ensino fundamental o tipo da escola (municipal ou estadual) tem influência tanto em matemática quanto em português, e os resultados mostram que os alunos que estudam em escolas municipais têm desempenho inferior aqueles da 4ª série que estudam nas escolas estaduais. O tipo da escola também é importante para os alunos da 8ª série do ensino fundamental que fizeram a prova de matemática, mas mostra o comportamento oposto do que é observado na 4ª série, ou seja, o desempenho dos alunos da escola estadual é inferior ao dos alunos da escola municipal. A conservação do prédio escolar também se mostra importante para os alunos da 8ª série nas duas disciplinas, aqueles que estudam em escolas em bom estado de conservação apresentam melhor desempenho que os alunos que estudam em escolas cujo estado de conservação do prédio não é bom. Quanto aos alunos da 3ª série

do ensino médio não houve uma variável no nível 2 em comum para as duas disciplinas. O fato do aluno ter defasagem idade-série e do diretor da escola ter informado que desenvolveu algum tipo de projeto pedagógico influencia no desempenho escolar dos alunos que fizeram a prova de matemática; já para os alunos que fizeram a prova de português, estudantes de escolas localizadas em municípios com maiores valores do Índice de Desenvolvimento Humano (IDH) apresentam melhores desempenho.

Em relação aos programas usados na análise dos dados não houve praticamente diferença nas estimativas dos parâmetros obtidos pelo HLM e pelo R. Os dois programas podem ser usados para análise de dados hierárquicos com a mesma segurança. Mas, tendo em vista que a versão gratuita para estudantes do HLM (`www.philscience.com/economics/ssicentral /hlm/hlmstu.htm`) não comporta um banco de dados com um grande número de observações sendo necessária a sua versão profissional que não está disponível gratuitamente, é mais vantajoso neste caso utilizar o R (`www.r-project.org`) que comporta bancos de dados de grande porte.

Modelos de regressão múltipla (MMQO) foram construídos com as mesmas variáveis selecionadas nos modelos lineares hierárquicos e foram obtidos os coeficientes de determinação (R^2) para as duas classes de modelos e os R^2 dos modelos hierárquicos foram, no mínimo, duas vezes maiores que os R^2 dos modelos de regressão múltipla mostrando um melhor desempenho do modelo linear hierárquico, sendo necessária a utilização do mesmo quando os dados estudados possuírem estrutura hierárquica.

Referências Bibliográficas

[1] Abbad, G. & Torres, C.V. Regressão múltipla stepwise e hierárquica em psicologia organizacional: Aplicações, problemas e soluções. *Estudos de Psicologia*, **7**, p.19-29, 2002.

[2] Aitkin, M., Anderson, D. & Hinde, J. Statistical modelling of data on teaching styles. *Journal of the Royal Statistical Society*, série A, **144**, p.148-161, 1981.

[3] Aitkin, M. & Longford, N. Statistical modelling in school effectiveness studies. *Journal of the Royal Statistical Society*, série A, **149**, p.1-43, 1986.

[4] Antunes, R.C. Avanços e perspectivas do melhoramento genético de suínos. `http://www.zoonews.com.br`, 2002.

[5] Atkinson, A.C. *Plots, Transformations and Regressions.* Oxford Statistical Science Series, Oxford, 1985.

[6] Barnett, V. & Lewis, T. *Outliers in Statistical Data*, 3ed. New York : Wiley, 1994.

[7] Beale, E.M.L. & Little, R.J.A. Missing values in multivariate analysis. *Journal of the Royal Statistical Society*, série B, **37**, p.129-145, 1975.

[8] Belsley, D.A., Kuh, E. & Welsch, R.E. *Regression Diagnostics*, John Wiley, New York, 1980.

[9] Beltrão, K., Leite, I. & Ferrão, M.E. O ambiente escolar no desempenho acadêmico do aluno: Criação de uma escola a partir do SAEB-99. *Estudos em avaliação educacional*, 2002.

[10] Bennett, N. *Teaching styles and pupil progress. Cambrige: Harvard University Press*, 1976.

[11] Bergamo, G.C. *Aplicação de modelos multiníveis na análise de dados de medidas repetidas no tempo.* Dissertação de mestrado defendida na Escola Superior de Agricultura Luiz de Queiroz. Universidade de São Paulo, 2002.

[12] Bickel, P.J. & Doksum, K.A. *Mathematical Statistics*, Holden-Day, 1976.

[13] Blight, B.J.N. Estimation from a censored sample for the exponential family. *Biometrika*, **57**, p.389-395, 1970.

[14] Brown, M.L. Identification of the sources of significance in two-way tables. *Appl. Statistical*, **23**, p.405-413, 1974.

[15] Bryk, A.S. & Raudenbush, S.W. *Hierarchical Linear Models: Applications and data analysis methods*. Sage Publications, 1992.

[16] Bryk, A.S., Raudenbush, S.W. & Congdon, R. Hierarchical linear and nonlinear modeling with the HLM/2L and HLM/3L programs. Chicago: *Scientific Software International*, 1996.

[17] Bryk, A.S., Raudenbush, S.W., Congdon, R. & Cheong, Y.F. Hierarchical linear and nonlinear modeling HLM 5 Chicago: *Scientific Software International*, 2001.

[18] Bryk, A.S. & Raudenbush, S.W. *Hierarchical Linear Models: Applications and data analysis methods*. 2^{ed} edition, Newbury Park, CA: Sage, 2002.

[19] Carter, W.H.Jr & Myers, R.H. Maximum liklihood estimation from linear combinations of discrete probability functions.*Journal of the American Statistical Association*, **68**, p.203-206, 1973.

[20] Cook, R.D. Detection of influential observations in linear regressions.*Technometrics*, **19**, p.15-18, 1977.

[21] Cook, R.D. Assessment of local influence. *Journal of the Royal Statistical Society*, série B, **48**, p.133-169, 1986.

[22] Dempster, A.P., Laird, N.M. & Rubin, D.B. Maximum likelihood from incomplete data via the EM algorithm.*Journal of the Royal Statistical Society*, série B, **39**, p.1-8, 1977.

[23] Dempster, A.P., Rubin, D.B. & Tsutakawa, R.K. Estimation in covariance components models. *Journal of the American Statistical Association,* **76**, p.341-353, 1981.

[24] Diez-Roux, A.V. Multilevel Analysis in Public Health Research.*Annual Reviews Public Health*, **21**, p.171-192, 2000.

[25] Ferrão, M.E., Beltrão, K. & Santos, D. Modelo de regressão multinível: aplicação ao estudo do impacto da política de não-repetência no desempenho escolar dos alunos da $4^{\underline{a}}$ série *Pesquisa e Planejamento Econômico*, **32**, 3, 2002.

[26] Giampaoli, V. *Inferência Estatística para Modelos Lineares com Restrições nos Parâmetros em Condições Regulares e Não Regulares*. Tese de Doutorado defendida no Instituto de Matemática e Estatística da Universidade de São Paulo, 1999.

[27] Goldstein, H. *Multilevel Statistical Models*, $2^{\underline{a}}$ed. London: Edward Arnold, 1995.

[28] Goldstein, H. *Multilevel Statistical Models*, $1^{\underline{a}}$ed. Internet London: Institute of Education, Multilevel Models Project, april 1999.

[29] Goldstein, H. *Multilevel Statistical Models*, 3ªed. Kendall's library of statistics, 2003.

[30] Hartley, H.O. Maximum liklihood estimation from incomplete data.*Biometrics*, **14**, p.174-194, 1958.

[31] Healy, M. & Westmacott, M. Missing values in experiments analysed on automatic computers.*Appl. Statistical*, **5**, p.203-206, 1956.

[32] Henderson, C.R., Kempthorne, O., Searle, S.R. & Vom Krosig, C.M. Estimation of environmental and genetic trends from records subjecting to culling. *Biometrics*, **15**, n.1, p.192-218, 1959.

[33] Hoaglin, D.C. & Welsch, R. The hat matrix in regression and ANOVA.*Am. Statistn*, **32**, p.17-22, 1978.

[34] Hox, J.J. *Applied Multilevel Analysis*, TT - Publikaties, 1995.

[35] Iemma, M. *Uso do melhor preditor linear não viesado em análises dialéticas e predição de híbridos*. Dissertação de mestrado defendida na Escola Superior de Agricultura Luiz de Queiroz, Universidade de São Paulo, 2003.

[36] IPEA/FIBGE/FJP. Programa das Nações Unidas para o Desenvolvimento, Atlas do Desenvolvimento Humano no Brasil (IDH-M), Brasília, 2000.

[37] Kreft, I. & Leeuw, D.J. *Introducing multilevel modeling*, Sage Publications, 1998.

[38] Langford, H.I. & Lewis, T. Outliers in multilevel data.*Journal of the Royal Statistical Society*, série A, **161**, part 2, p.121-160, 1998.

[39] Lima, C.R.O.P. *Predição linear não-tendenciosa ótima*. Dissertação de mestrado defendida no Instituto de Matemática, Estatística e Ciência de Computação, UNICAMP, 1987.

[40] Lindley, D.V. & Smith, A.F.M. Bayes estimates for the linear model. *Journal of the Royal Statistical Society*, série B, **34**, p.1-41, 1972.

[41] Longford, N.T. *Random coefficient models.* Oxford: Clarendon press, 1993.

[42] Magalhães, M.N. & Lima, A.C.P. *Noções de probabilidade e Estatística*, Edusp, 2002.

[43] Mason, W.M., Wong, G.M. & Entwistle, B. *Contextual analysis through the multilevel linear model. In S. Leinhardt, Sociological methodology.* San Francisco: Jossey- Bass, 1983.

[44] McCullagh, P. & Nelder, J.A. *Generalized Linear Models*, 2ªed. London:Chapman and Hall, 1989.

[45] Natis, L. *Modelos Lineares hierárquicos.* Dissertação de mestrado defendida no Instituto de Matemática e Estatística da Universidade de São Paulo, 2000.

[46] Orchard, T. & Woodbury, M.A. A missing information principle: Theory and applications. *Proc. 6th Berkeley Symposium on Math. Statistical and Probability*, **1**, p.697-715, 1972.

[47] Panter, D.M. & Allen, F.L. Using best linear unbiased predictions to enhence breeding for yield in soybean: choosing parents. *Crop Science*, **35**, n.2, p.397-405, mar-apr 1995.

[48] Paula, A.G. *Modelos de Regressão - Com apoio computacional*, IME - USP, jun, 2004.

[49] Pernambuco, Governo do Estado de. Sistema de Avaliação Educacional de Pernambuco: SAEPE: relatório 2002/Secretaria de Educação e Cultura, Recife, 2003.

[50] Pinheiro, J.C. & Bates, D.M. `http://www.r-project.org, ver pacote nlme, 2000`.

[51] Pregibon, D. Logistic regression diagnostics.*Annals of Statistics*, **9**, p.705-724, 1981.

[52] Roberts, K. & Burstein, L. *New directions for methodology for social and behavioral sciences*, 1980.

[53] Sorensen, D.A. & Kennedy, B.W. Estimation of response to selection using least-square and mixed model methodology.*Journal of Animals Science*, **58**, n.4, p.1097-1106, apr 1984.

[54] Sousa, S.O. *Estimação robusta no modelo de calibração.* Dissertação de mestrado defendida na Universidade Federal de Pernambuco, 2002.

[55] Sullivan, L.M., Dukes, K. & Losina, E. Tutorial in biostatistics an introduction to hierarchical linear modelling.*Statistics in Medicine*, **18**, p.855-888, 1999.

[56] Sundberg, R. Maximum likelihood theory for incomplete data from an exponential family. *Scand. Journal Statist.*, **1**, p.49-58, 1974.

[57] Torres, C.V. & Curtin, S. *Instrument for assessment of carreer development needs. Instrumento de avaliação.* San Diego: The city of San Diego Press, 1999.

[58] Tseloni, A. Comparing multilevel and single-level negative binomial regression models of personal crimes: evidence from the national crime victimization survey. *Department of Criminalogy and criminal justice*, University of Naryland, 1999.

[59] Turnbull, B.W. The empirical distribuition function with arbitrarily grouped, censored and truncated data.*Journal of the Royal Statistical Society*, série B, **38**, p.290-295, 1976.

[60] Van den Eeden, P. & Hüttner, H.J.M. *Multilevel research.* Current Sociology, 1982.

Anexo 1 – Programas Utilizados

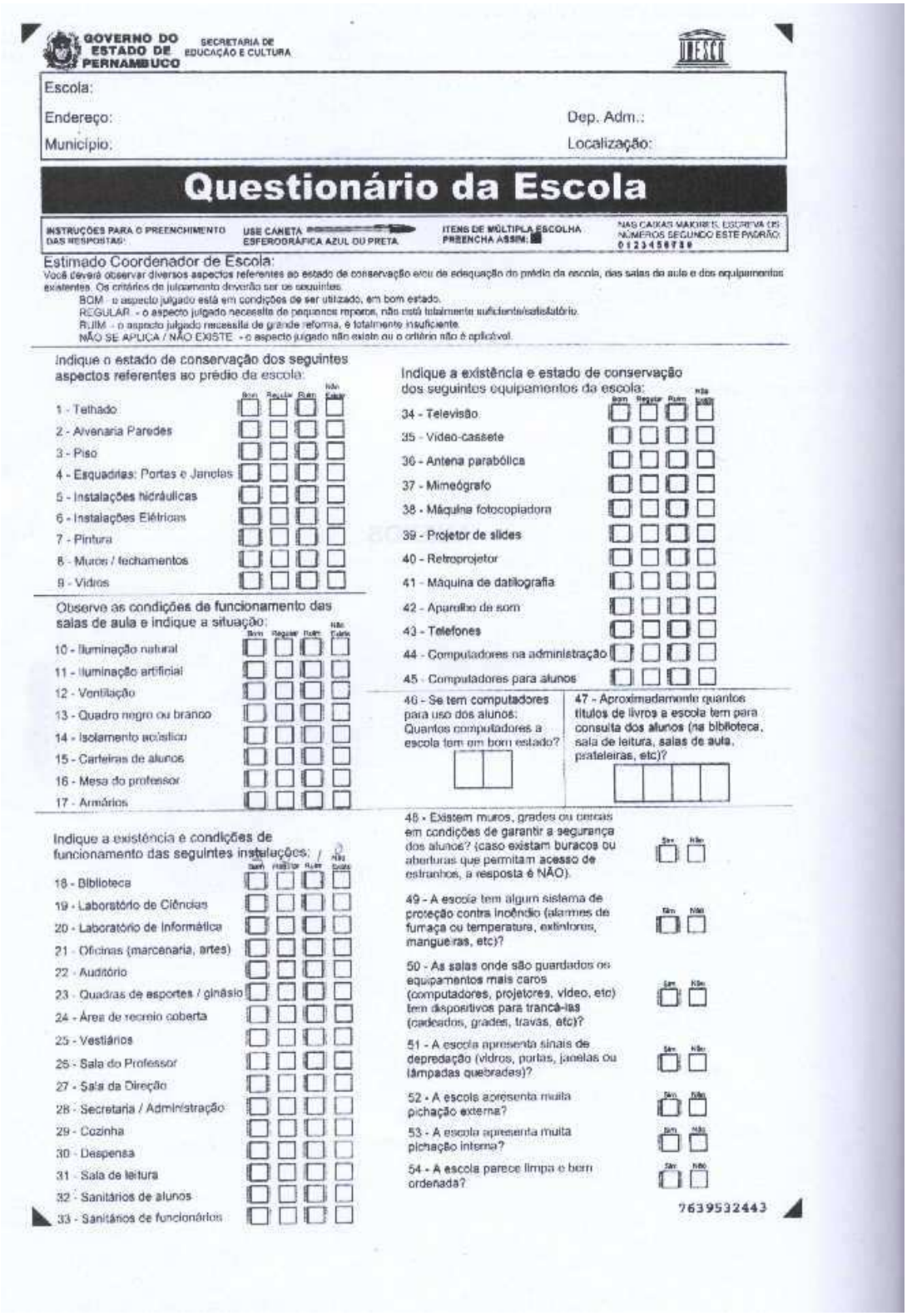
```
*****************************Modelos Finais*********************************


mat4<- read.table("c:/userroot/Alunos/sandraP/bancos30/R_4_mat_30.dat",header=T)
dadosm4=data.frame(mat4)

matemat4 = lme(notam4 ~ tipo +  acelera + difidad + q2sexo + aprendiz, random = ~1 +
acelera | codigo, data=dadosm4)
summary(matemat4)
intervals(matemat4)

*********
port4<- read.table("c:/userroot/Alunos/sandraP/bancos30/R_4_port_30.dat",header=T)
dadosp4=data.frame(port4)

portug4 = lme(notap4 ~ tipo +  acelera + difidad + q2sexo + aprendiz, random = ~1 +
acelera  | codigo, data=dadosp4)
summary(portug4)
intervals(portug4)

*********

mat8<- read.table("c:/userroot/Alunos/sandraP/bancos30/R_8_mat_30.dat",header=T)
dadosm8=data.frame(mat8)

matemat8 = lme(notam8 ~ tipo +  conserv7+ difidad + q2sexo + aprendiz +
difidad*conserv7 + q2sexo*tipo , random = ~1 + difidad + q2sexo | codigo,
data=dadosm8)

summary(matemat8)
intervals(matemat8)

*********

port8<- read.table("c:/userroot/Alunos/sandraP/bancos30/R_8_port_30.dat",header=T)
dadosp8=data.frame(port8)

portug8 = lme(notap8 ~ conserv7 + difidad + q2sexo + aprendiz, random = ~1 +
difidad | codigo,data=dadosp8)

summary(portug8)
intervals(portug8)

*************

mat3<- read.table("c:/mestrado/sandraP2/R_3_mat_30.dat",header=T)
dadosm3=data.frame(mat3)

matemat3 = lme(notam3 ~ acelera +  difidad + q2sexo + aprendiz + difidad*projeto,
```

```
 random = ~ -1 + acelera + difidad | codigo, data=dadosm3)

summary(matemat3)
intervals(matemat3)

**********

port3<- read.table("c:/userroot/Alunos/sandraP/bancos30/R_3_port_30.dat",header=T)
dadosp3=data.frame(port3)

portug3 = lme(notap3 ~ acelera +  difidad + q2sexo + aprendiz + idh, random = ~ 1
+ acelera + difidad | codigo, data=dadosp3)

summary(portug3)
intervals(portug3)

**********GRFICOS - HISTOGRAMA**********

 mat4<- read.table("c:/userroot/Alunos/sandraP/bancos30/R_4_mat_30.dat",header=T)
 dadosm4=data.frame(mat4)
 postscript("c:/userroot/Alunos/sandraP/bancos30/histogs.ps", horizontal=FALSE,
 paper="letter")
 par(mfrow=c(3,2),pty="s")
 #fix(dadosm4)
 nota1m4<-dadosm4[,5]
 hist(nota1m4,main="Matemtica - 4 srie")

 port4<- read.table("c:/userroot/Alunos/sandraP/bancos30/R_4_port_30.dat",header=T)
 dadosp4=data.frame(port4)
 #fix(dadosp4)
 nota1p4<-dadosp4[,5]
 hist(nota1p4,main="Portugus - 4 srie")

 mat8<- read.table("c:/userroot/Alunos/sandraP/bancos30/R_8_mat_30.dat",header=T)
 dadosm8=data.frame(mat8)
 #fix(dadosm8)
 nota1m8<-dadosm8[,4]
 hist(nota1m8,main="Matemtica - 8 srie")

 port8<- read.table("c:/userroot/Alunos/sandraP/bancos30/R_8_port_30.dat",header=T)
 dadosp8=data.frame(port8)
 #fix(dadosp8)
 nota1p8<-dadosp8[,3]
 hist(nota1p8,main="Portugus - 8 srie")

 mat3<- read.table("c:/userroot/Alunos/sandraP/bancos30/R_3_mat_30.dat",header=T)
 dadosm3=data.frame(mat3)
 #fix(dadosm3)
 nota1m3<-dadosm3[,5]
 hist(nota1m3,main="Matemtica - 3 srie")

 port3<- read.table("c:/userroot/Alunos/sandraP/bancos30/R_3_port_30.dat",header=T)
 dadosp3=data.frame(port3)
```

```
 #fix(dadosp3)
 nota1p3<-dadosp3[,5]
 hist(nota1p3,main="Portugus - 3 srie")

 dev.off()

*******************************Grficos QQ-Plots************************************

mat4<- read.table("c:/userroot/Alunos/sandraP/bancos30/R_4_mat_30.dat",header=T)
dadosm4=data.frame(mat4)
matemat4 = lme(notam4 ~ tipo +  acelera + difidad + q2sexo + aprendiz, random = ~1
+ acelera | codigo, data=dadosm4)
postscript("c:/userroot/Alunos/sandraP/qqplots.ps", horizontal=FALSE, paper="letter")
errom4=resid(matemat4)
par(mfrow=c(3,2),pty="s")
qqnorm(errom4,main="QQ-plot - Matemtica 4 srie")

port4<- read.table("c:/userroot/Alunos/sandraP/bancos30/R_4_port_30.dat",header=T)
dadosp4=data.frame(port4)
portug4 = lme(notap4 ~ tipo +  acelera + difidad + q2sexo + aprendiz, random = ~1 +
acelera  | codigo, data=dadosp4)
errop4=resid(portug4)
qqnorm(errop4,main="QQ-plot - Portugus 4 srie")

mat8<- read.table("c:/userroot/Alunos/sandraP/bancos30/R_8_mat_30.dat",header=T)
dadosm8=data.frame(mat8)
matemat8 = lme(notam8 ~ tipo +  conserv7+ difidad + q2sexo + aprendiz +
difidad*conserv7 + q2sexo*tipo , random = ~1 + difidad + q2sexo | codigo,
data=dadosm8)
errom8=resid(matemat8)
qqnorm(errom8,main="QQ-plot - Matemtica 8 srie")

port8<- read.table("c:/userroot/Alunos/sandraP/bancos30/R_8_port_30.dat",header=T)
dadosp8=data.frame(port8)
portug8 = lme(notap8 ~ conserv7 + difidad + q2sexo + aprendiz, random = ~1 + difidad
| codigo,data=dadosp8)
errop8=resid(portug8)
qqnorm(errop4,main="QQ-plot - Portugus 8 srie")

mat3<- read.table("c:/userroot/Alunos/sandraP/bancos30/R_3_mat_30.dat",header=T)
dadosm3=data.frame(mat3)
matemat3 = lme(notam3 ~ acelera +  difidad + q2sexo + aprendiz + difidad*projeto,
 random = ~ -1 + acelera + difidad | codigo, data=dadosm3)
errom3=resid(matemat3)
qqnorm(errom3,main="QQ-plot - Matemtica 3 srie")

port3<- read.table("c:/userroot/Alunos/sandraP/bancos30/R_3_port_30.dat",header=T)
dadosp3=data.frame(port3)
portug3 = lme(notap3 ~ acelera +  difidad + q2sexo + aprendiz + idh, random = ~ 1 +
acelera + difidad | codigo, data=dadosp3)
errop3=resid(portug3)
qqnorm(errop3,main="QQ-plot - Portugus 3 srie")
```

```
dev.off()

**********Mnimos Quadrados (para cada modelo)**********

yajust=fitted(portug4)
yobs=dadosp4[,5]
ymedia=mean(yobs)

dif1=(yobs-ymedia)^2
sqt=sum(dif1)
sqt

dif2=(yobs-yajust)^2
sqres=sum(dif2)
sqres

part1=((yobs-yajust)*(yajust-ymedia))
sompart1=2*sum(part1)
sompart1

part2=(yajust-ymedia)^2
sompart2=sum(part2)

sqreg=sompart1+sompart2
sqreg
rq=sqreg/sqt
rq
```

Anexo 2 – Questionários do SAEPE



GOVERNO DO ESTADO DE PERNAMBUCO — SECRETARIA DE EDUCAÇÃO E CULTURA

UNESCO

Escola:

Endereço: Dep. Adm.:

Município: Localização:

Questionário da Escola

INSTRUÇÕES PARA O PREENCHIMENTO DAS RESPOSTAS: USE CANETA ESFEROGRÁFICA AZUL OU PRETA. ITENS DE MÚLTIPLA ESCOLHA PREENCHA ASSIM: ■ NAS CAIXAS MAIORES, ESCREVA OS NÚMEROS SEGUNDO ESTE PADRÃO: 0 1 2 3 4 5 6 7 8 9

Estimado Coordenador de Escola:

Você deverá observar diversos aspectos referentes ao estado de conservação e/ou de adequação do prédio da escola, das salas de aula e dos equipamentos existentes. Os critérios de julgamento deverão ser os seguintes:

BOM - o aspecto julgado está em condições de ser utilizado, em bom estado.
REGULAR - o aspecto julgado necessita de pequenos reparos, não está totalmente suficiente/satisfatório.
RUIM - o aspecto julgado necessita de grande reforma, é totalmente insuficiente.
NÃO SE APLICA / NÃO EXISTE - o aspecto julgado não existe ou o critério não é aplicável.

Indique o estado de conservação dos seguintes aspectos referentes ao prédio da escola:

	Bom	Regular	Ruim	Não existe
1 - Telhado	☐	☐	☐	☐
2 - Alvenaria Paredes	☐	☐	☐	☐
3 - Piso	☐	☐	☐	☐
4 - Esquadrias: Portas e Janelas	☐	☐	☐	☐
5 - Instalações hidráulicas	☐	☐	☐	☐
6 - Instalações Elétricas	☐	☐	☐	☐
7 - Pintura	☐	☐	☐	☐
8 - Muros / fechamentos	☐	☐	☐	☐
9 - Vidros	☐	☐	☐	☐

Observe as condições de funcionamento das salas de aula e indique a situação:

	Bom	Regular	Ruim	Não existe
10 - Iluminação natural	☐	☐	☐	☐
11 - Iluminação artificial	☐	☐	☐	☐
12 - Ventilação	☐	☐	☐	☐
13 - Quadro negro ou branco	☐	☐	☐	☐
14 - Isolamento acústico	☐	☐	☐	☐
15 - Carteiras de alunos	☐	☐	☐	☐
16 - Mesa do professor	☐	☐	☐	☐
17 - Armários	☐	☐	☐	☐

Indique a existência e condições de funcionamento das seguintes instalações:

	Bom	Regular	Ruim	Não existe
18 - Biblioteca	☐	☐	☐	☐
19 - Laborstório de Ciências	☐	☐	☐	☐
20 - Laboratório de Informática	☐	☐	☐	☐
21 - Oficinas (marcenaria, artes)	☐	☐	☐	☐
22 - Auditório	☐	☐	☐	☐
23 - Quadras de esportes / ginásio	☐	☐	☐	☐
24 - Área de recreio coberta	☐	☐	☐	☐
25 - Vestiários	☐	☐	☐	☐
26 - Sala do Professor	☐	☐	☐	☐
27 - Sala da Direção	☐	☐	☐	☐
28 - Secretaria / Administração	☐	☐	☐	☐
29 - Cozinha	☐	☐	☐	☐
30 - Despensa	☐	☐	☐	☐
31 - Sala de leitura	☐	☐	☐	☐
32 - Sanitários de alunos	☐	☐	☐	☐
33 - Sanitários de funcionários	☐	☐	☐	☐

Indique a existência e estado de conservação dos seguintes equipamentos da escola:

	Bom	Regular	Ruim	Não existe
34 - Televisão	☐	☐	☐	☐
35 - Vídeo-cassete	☐	☐	☐	☐
36 - Antena parabólica	☐	☐	☐	☐
37 - Mimeógrafo	☐	☐	☐	☐
38 - Máquina fotocopiadora	☐	☐	☐	☐
39 - Projetor de slides	☐	☐	☐	☐
40 - Retroprojetor	☐	☐	☐	☐
41 - Máquina de datilografia	☐	☐	☐	☐
42 - Aparelho de som	☐	☐	☐	☐
43 - Telefones	☐	☐	☐	☐
44 - Computadores na administração	☐	☐	☐	☐
45 - Computadores para alunos	☐	☐	☐	☐

46 - Se tem computadores para uso dos alunos: Quantos computadores a escola tem em bom estado? ☐☐

47 - Aproximadamente quantos títulos de livros a escola tem para consulta dos alunos (na biblioteca, sala de leitura, salas de aula, prateleiras, etc)? ☐☐☐☐

48 - Existem muros, grades ou cercas em condições de garantir a segurança dos alunos? (caso existam buracos ou aberturas que permitam acesso de estranhos, a resposta é NÃO). Sim ☐ Não ☐

49 - A escola tem algum sistema de proteção contra incêndio (alarmes de fumaça ou temperatura, extintores, mangueiras, etc)? Sim ☐ Não ☐

50 - As salas onde são guardados os equipamentos mais caros (computadores, projetores, vídeo, etc) tem dispositivos para trancá-las (cadeados, grades, travas, etc)? Sim ☐ Não ☐

51 - A escola apresenta sinais de depredação (vidros, portas, janelas ou lâmpadas quebradas)? Sim ☐ Não ☐

52 - A escola apresenta muita pichação externa? Sim ☐ Não ☐

53 - A escola apresenta muita pichação interna? Sim ☐ Não ☐

54 - A escola parece limpa e bem ordenada? Sim ☐ Não ☐

7639532443

GOVERNO DO ESTADO DE PERNAMBUCO
SECRETARIA DE EDUCAÇÃO E CULTURA

SAEPE 2002
SISTEMA DE AVALIAÇÃO EDUCACIONAL DE PERNAMBUCO

UNESCO

GOVERNO DO ESTADO DE PERNAMBUCO
SECRETARIA DE EDUCAÇÃO E CULTURA
DIRETORIA DE POLÍTICA E PROGRAMAS EDUCACIONAIS

QUESTIONÁRIO DO DIRETOR

Prezado(a) Diretor(a):

A Secretaria de Educação do Estado de Pernambuco, em articulação com as Secretarias Municipais de Educação do Estado, representadas pela União Nacional dos Dirigentes Municipais de Educação (UNDIME), e com o apoio técnico da Organização das Nações Unidas para a Educação, a Ciência e a Cultura (UNESCO) está coletando dados para o 2º Ciclo do Sistema de Avaliação Educacional de Pernambuco (SAEPE). A finalidade é apresentar às escolas informações e elementos para melhorar os serviços educacionais do Estado.

É importante ressaltar que não é nosso objetivo avaliar individualmente diretores, professores ou alunos. As informações só serão divulgadas por Escola, Município ou Região. Por isso, damos garantia absoluta de sigilo sobre as informações fornecidas neste questionário.

Antes de responder às perguntas do questionário, algumas informações lhe serão úteis:

1. Não há respostas certas ou erradas. Você terá que responder com base naquilo que tem vivido e percebido em sua escola.
2. Leia atentamente cada questão. Em geral, você deverá escolher apenas uma das alternativas. Em alguns itens você poderá escolher mais de uma alternativa; Quando isso acontecer, será indicado no cabeçalho do item.
3. Se você trabalha em mais de uma escola, **as respostas devem ser dadas levando em consideração apenas a unidade escolar que lhe forneceu o questionário.**
4. As questões devem ser respondidas com base no trabalho desenvolvido neste ano 2002.
5. Você deverá responder diretamente no Questionário, marcando com **caneta preta ou azul** a alternativa escolhida.
6. Não deixe questão sem resposta.

Agradecemos de antemão sua participação no processo e esperamos, juntos, melhorar a qualidade da educação em nosso Estado.

exo
A) Feminino.
B) Masculino.

lade: _____ anos.

ual o seu nível de escolaridade completo?
A) Nenhum.
B) Ensino Fundamental - 4ª série (antigo 1º Grau Menor).
C) Ensino Fundamental - 8ª série (antigo 1º Grau Maior).
D) Ensino Médio - Magistério (antigo 2º Grau).
E) Ensino Médio - Outros (antigo 2º Grau).
F) Superior - Pedagogia.
G) Superior - Outra Licenciatura.
H) Superior - Outros.

ocê completou algum curso de pós-graduação?
A) Não.
B) Curso de extensão.
C) Curso de aperfeiçoamento.
D) Curso de especialização.
E) Mestrado.
F) Doutorado.

ocê tem escolaridade especifica na área de dministração Escolar?
A) Não tem.
B) Graduação com habilitação em Administração Escolar.
C) Especialização na área de Administração Escolar.
D) Mestrado na área de Administração Escolar.
E) Doutorado na área de Administração Escolar.

sua experiência e necessidades como diretor scola, como avaliaria a formação que recebeu juintes aspectos?

CTOS	Muito boa	Boa	Regular	Ruim
ra o planejamento s atividades da cola.	(A)	(B)	(C)	(D)
ra a elaboração do ojeto pedagógico da cola.	(A)	(B)	(C)	(D)
ra a mocratização da stão da escola.	(A)	(B)	(C)	(D)
ra gerir os aspectos dagógicos da cola	(A)	(B)	(C)	(D)
a gerir os recursos manos da escola.	(A)	(B)	(C)	(D)
a gerir os recursos teriais da escola.	(A)	(B)	(C)	(D)
a um bom acionamento com a munidade escolar.	(A)	(B)	(C)	(D)

13. **Você participou de alguma atividade de formação continuada (capacitação, treinamento ou atualização) nos 2 últimos anos?**
(A) Sim.
(B) Não. *(Passe para a questão 24)*

14. **Neste ano, quantas horas de cursos de capacitação, treinamento ou atualização você freqüentou?**
_____ horas.

As questões 15 a 23 apresentam alguns conteúdos de cursos. Responda se cada um deles constava, ou não, das atividades de formação continuada anteriormente indicada (*Marque* **Sim** *ou* **Não** *em cada linha)* e se foi útil (**Muito, Pouco** *ou* **Nada**) para seu trabalho como diretor desta escola.

Conteúdos	a) Constava?		b) Foi útil?		
	Não	Sim	Muito	Pouco	Nada
15. Aspectos Administrativos, Financeiros e Legais da Gestão Escolar.	(A)	(B)	(A)	(B)	(C)
16. Aspectos Pedagógicos da Gestão Escolar.	(A)	(B)	(A)	(B)	(C)
17. Fundamentos da Educação.	(A)	(B)	(A)	(B)	(C)
18. Estrutura e Funcionamento do Ensino.	(A)	(B)	(A)	(B)	(C)
19. Princípios da Avaliação Institucional.	(A)	(B)	(A)	(B)	(C)
20. Atualização Cultural. (arte, informação, cidadania, etc.).	(A)	(B)	(A)	(B)	(C)
21. Relações Interpessoais na Escola.	(A)	(B)	(A)	(B)	(C)
22. Estilos de Liderança.	(A)	(B)	(A)	(B)	(C)
23. Novas Tecnologias Aplicadas à Educação (televisão, videocassete, computador, etc.).	(A)	(B)	(A)	(B)	(C)

24. **Há quantos anos você trabalha em educação?**
_____ anos.

25. **Quantos anos de trabalho você tem em funções administrativas, tais como diretor, vice-diretor, secretário de escola, supervisor ou coordenador?** _____ anos.

26. **Desses, quantos anos foram só de direção de escola?**
_____ anos.

27. **Há quantos anos você é diretor(a) desta escola?**
_____ anos.

28. **Você ingressou na carreira do magistério por meio de concurso público?**
(A) Sim.
(B) Não.

Como você assumiu a direção desta escola?
(A) Concurso público.
(B) Eleição por colegiado escolar.
(C) Eleição pela comunidade escolar.
(D) Exame de seleção.
(E) Exame de seleção e eleição.
(F) Indicação de técnicos.
(G) Indicação de políticos.

Qual é a sua situação trabalhista nesta escola?
(A) Estatutário.
(B) CLT.
(C) Prestador de serviço (contrato temporário).
(D) Sem contrato.

Qual é sua carga horária de trabalho nesta escola?
_____ horas semanais.

Qual é o seu salário bruto como diretor(a) nesta escola?
R$_____.

Você exerce outra atividade profissional além da direção desta escola?
(A) Sim. (B) Não.

O CONSELHO DE ESCOLA é um órgão colegiado, constituído por representantes da escola e da comunidade, que tem como objetivo acompanhar as atividades escolares. Esta escola já constituiu o Conselho de Escola?
(A) Sim. (B) Não.

O Conselho de Escola pode ser composto por vários representantes. Responda se cada uma das categorias apresentadas nas questões 35 a 39 pertence, ou não, ao Conselho desta Escola, e quantos de cada categoria formam parte do Conselho de Escola.

*Marque **UMA** alternativa em cada linha.*

	Representantes	Não	Sim	Quantos?
35.	Professores?	(A)	(B)	
36.	Alunos?	(A)	(B)	
37.	Funcionários?	(A)	(B)	
38.	Pais?	(A)	(B)	
39.	Comunidade?	(A)	(B)	

40. **ESTE ANO, quantas vezes o Conselho de Escola se reuniu?** *(se não houve reuniões, indique 0 (zero))*
_____ vezes.

41. **Qual a natureza das questões tratadas nessas reuniões?**
(A) Predominantemente administrativas.
(B) Predominantemente pedagógicas.
(C) O Conselho de Escola não se reuniu este ano.

42. **O Conselho de Classe é um órgão formado por todos os professores que lecionam em cada turma/série. ESTE ANO, quantas vezes se reuniram os Conselhos de Classe desta escola?**
(A) Não existe Conselho de Classe.
(B) Nenhuma vez.
(C) Uma vez.
(D) Duas vezes.
(E) Três vezes.
(F) Quatro vezes ou mais.

43. **ESTE ANO, que assunto predominou nas reuniões do Conselho de Classe?**

*Marque apenas **UMA** alternativa.*

(A) Não existe Conselho de Classe.
(B) O Conselho de Classe não se reuniu este ano.
(C) Acompanhamento do projeto da escola.
(D) Problemas de aprendizagem e rendimento escolar.
(E) Critérios e procedimentos de avaliação dos alunos.
(F) Supervisão e controle das atividades da equipe escolar.
(G) Alunos com problemas disciplinares.
(H) Outro.

44. **Como foi desenvolvido o projeto pedagógico desta escola para ESTE ANO letivo?**
*Marque apenas **UMA** alternativa.*
(A) Não foi desenvolvido projeto pedagógico este ano.
(B) Através da aplicação de modelo encaminhado pela Secretaria da Educação.
(C) Pela equipe de professores e técnicos da escola.
(D) Por mim, com base em minha experiência anterior.
(E) Outra.
(F) Não sei.

45. **Quem participou da elaboração ou atualização do plano de trabalho desta escola para o ano de 2000?**
(A) A direção elaborou sozinha.
(B) A direção e a equipe técnica.
(C) A direção, a equipe técnica e os professores.
(D) A direção, a equipe técnica e os professores, com a participação de pais e alunos.

Questionário do aluno da 4ª Série do Ensino Fundamental

01. Mês e ano de nascimento:

Mês Ano

		/				

02. Qual o seu sexo?
(A) Masculino.
(B) Feminino.

03. Com quem você mora?
(A) Com a própria família.
(B) Só com o pai.
(C) Só com a mãe.
(D) Com parentes.
(E) Com amigos.
(F) Com o pai ou a mãe em nova união.

04. Qual o nível de instrução do seu pai?
(A) Nunca freqüentou a escola.
(B) Ensino Fundamental – 1ª a 4ª série.
(C) Ensino Fundamental – 5ª a 8ª série.
(D) Ensino Médio.
(E) Superior.
(F) Pós-graduação.
(G) Não sei.

05. Qual o nível de instrução de sua mãe?
(A) Nunca freqüentou a escola.
(B) Ensino Fundamental – 1ª a 4ª série.
(C) Ensino Fundamental – 5ª a 8ª série.
(D) Ensino Médio.
(E) Superior.
(F) Pós-graduação.
(G) Não sei.

06. Seu pai (ou responsável) está trabalhando atualmente?
(A) Sim.
(B) Não, ele está desempregado.
(C) Não, ele é aposentado.

07. Sua mãe está trabalhando atualmente? *(Trabalho remunerado)*
(A) Sim.
(B) Não, ela está desempregada.
(C) Não, ela é aposentada.
(D) Não, ela não trabalha fora.
(E) Não sei.

08. Você já deixou de freqüentar a escola em algum período?
(A) Sim, por 1 ano.
(B) Sim, por 2 anos.
(C) Sim, por 3 anos.
(D) Sim, por 4 anos.
(E) Sim, por mais de 4 anos.
(F) Não.

09. Caso você tenha abandonado a escola por algum período, qual foi o motivo?
(A) Doença/problemas de saúde.
(B) Trabalho/recursos financeiros.
(C) Falta de interesse, motivação.
(D) Falta de vagas.
(E) Reprovação(ões).
(F) Nunca abandonei a escola.

10. Você já repetiu de ano?
(A) Não.
(B) Sim, 1 vez.
(C) Sim, 2 vezes.
(D) Sim, 3 vezes.
(E) Sim, 4 vezes.
(F) Sim, 5 vezes.
(G) Sim, 6 vezes ou mais.

11. Que séries você repetiu?
(marque todas as séries que você repetiu)

(A) 1ª (B) 2ª (C) 3ª (D) 4ª

12. Você costuma conversar em casa sobre o que acontece na escola?
(A) Sim, quase todos os dias.
(B) Sim, só uma vez por semana.
(C) Sim, só quando recebo as notas.
(D) Sim, só no final do ano.
(E) Não converso.

13. Seu pai, sua mãe ou responsável participa(m) das reuniões da escola?
(A) Sim, sempre.
(B) Sim, às vezes.
(C) Não, nunca.
(D) Não. A escola não chama para reuniões.

14. Em sua casa se compra jornal?
(A) Sim, todos os dias.
(B) Sim, de vez em quando.
(C) Sim, em fins de semana.
(D) Não.

15. Quantas horas por dia você assiste a programas de televisão?
(A) Não assisto à televisão.
(B) Menos de 2 horas.
(C) De 2 a 4 horas.
(D) De 5 a 6 horas.
(E) Mais de 6 horas.

16. Você sabe utilizar computador?
(A) Não sei usar.
(B) Sei usar só para jogos.
(C) Sei usar para fazer trabalhos (edição de textos

17. Em sua escola, você utiliza computador?

(A) Não, na escola não tem computador.
(B) Na escola tem computador, mas não uso.
(C) Sim, na escola tem computador, mas o utilizo poucas vezes.
(D) Sim, na escola tem computador e o utilizo freqüentemente.

18. E, em sua casa, você utiliza computador?

(A) Não, em minha casa não tem computador.
(B) Em minha casa tem computador, mas eu não utilizo.
(C) Sim, em minha casa tem computador, mas o utilizo poucas vezes.
(D) Sim, em minha casa tem computador e o utilizo freqüentemente.

19. Qual das duas provas você achou mais difícil?

(A) Matemática.
(B) Português.

20. Quantas horas, por dia, você dedica, em média, às tarefas da escola (trabalhos encomendados pelos professores, leitura ou estudo)?

(A) Nenhuma, porque os professores da escola não dão tarefas.
(B) Nenhuma, porque não gosto de fazer tarefas.
(C) Menos de 1 hora por dia.
(D) Entre 1 e 2 horas por dia.
(E) Entre 2 e 3 horas por dia.
(F) Entre 3 e 4 horas por dia.
(G) Entre 4 e 5 horas por dia.
(H) Mais de 5 horas por dia.

21. Você gosta de ler?

(A) Gosto muito.
(B) Gosto.
(C) Gosto pouco.
(D) Não gosto.

22. Qual das disciplinas exige mais tarefas de casa?

(A) Matemática.
(B) Português.

23. Você consegue entender tudo que seu professo ensina?

(A) Sim, sempre entendo.
(B) Sim, consigo entender a maior parte.
(C) Consigo entender pouco.
(D) Nunca entendo.

Questionário do aluno da 8ª Série do Ensino Fundamental

01. Mês e ano de nascimento:

Mês Ano

		/				

02. Qual o seu sexo?
(A) Masculino.
(B) Feminino.

03. Com quem você mora?
(A) Com a própria família.
(B) Só com o pai.
(C) Só com a mãe.
(D) Com parentes.
(E) Com amigos.
(F) Com o pai ou a mãe em nova união.

04. Qual o nível de instrução do seu pai?
(A) Nunca freqüentou a escola.
(B) Ensino Fundamental – 1ª a 4ª série.
(C) Ensino Fundamental – 5ª a 8ª série.
(D) Ensino Médio.
(E) Superior.
(F) Pós-graduação.
(G) Não sei.

05. Qual o nível de instrução de sua mãe?
(A) Nunca freqüentou a escola.
(B) Ensino Fundamental – 1ª a 4ª série.
(C) Ensino Fundamental – 5ª a 8ª série.
(D) Ensino Médio.
(E) Superior.
(F) Pós-graduação.
(G) Não sei.

06. Seu pai (ou responsável) está trabalhando atualmente?
(A) Sim.
(B) Não, ele está desempregado.
(C) Não, ele é aposentado.

07. Sua mãe está trabalhando atualmente? *(Trabalho remunerado)*
(A) Sim.
(B) Não, ela está desempregada.
(C) Não, ela é aposentada.
(D) Não, ela não trabalha fora.
(E) Não sei.

08. Você já deixou de freqüentar a escola em algum período?
(A) Sim, por 1 ano.
(B) Sim, por 2 anos.
(C) Sim, por 3 anos.

09. Caso você tenha abandonado a escola por algum período, qual foi o motivo?
(A) Doença/problemas de saúde.
(B) Trabalho/recursos financeiros.
(C) Falta de interesse, motivação.
(D) Falta de vagas.
(E) Reprovação(ões).
(F) Nunca abandonei a escola.

10. Você já repetiu de ano?
(A) Não.
(B) Sim, 1 vez.
(C) Sim, 2 vezes.
(D) Sim, 3 vezes.
(E) Sim, 4 vezes.
(F) Sim, 5 vezes.
(G) Sim, 6 vezes.
(H) Sim, 7 vezes.
(I) Sim, 8 vezes.
(J) Sim, 9 vezes ou mais.

11. Que séries você repetiu?
(marque todas as séries que você repetiu)

(A) 1ª	(B) 2ª	(C) 3ª	(D) 4ª
(E) 5ª	(F) 6ª	(G) 7ª	(H) 8ª

12. Você costuma conversar em casa sobre o que acontece na escola?
(A) Sim, quase todos os dias.
(B) Sim, só uma vez por semana.
(C) Sim, só quando recebo as notas.
(D) Sim, só no final do ano.
(E) Não converso.

13. Seu pai, sua mãe ou responsável participa(m) das reuniões da escola?
(A) Sim, sempre.
(B) Sim, às vezes.
(C) Não, nunca.
(D) Não. A escola não chama para reuniões.

14. Em sua casa se compra jornal?
(A) Sim, todos os dias.
(B) Sim, de vez em quando.
(C) Sim, em fins de semana.
(D) Não.

15. Quantas horas por dia você assiste a programas de televisão?
(A) Não assisto à televisão.
(B) Menos de 2 horas.

16. Você sabe utilizar computador?

(A) Não sei usar.
(B) Sei usar só para jogos.
(C) Sei usar para fazer trabalhos (edição de textos e/ou planilha eletrônica).

17. Em sua escola, você utiliza computador?

(A) Não, na escola não tem computador.
(B) Na escola tem computador, mas não uso.
(C) Sim, na escola tem computador, mas o utilizo poucas vezes.
(D) Sim, na escola tem computador e o utilizo freqüentemente.

18. E, em sua casa, você utiliza computador?

(A) Não, em minha casa não tem computador.
(B) Em minha casa tem computador, mas eu não utilizo.
(C) Sim, em minha casa tem computador, mas o utilizo poucas vezes.
(D) Sim, em minha casa tem computador e o utilizo freqüentemente.

19. Qual das duas provas você achou mais difícil?

(A) Matemática.
(B) Português.

20. Quantas horas, por dia, você dedica, em média, às tarefas da escola (trabalhos encomendados pelos professores, leitura ou estudo)?

(A) Nenhuma, porque os professores da escola não dão tarefas.
(B) Nenhuma, porque não gosto de fazer tarefas.
(C) Menos de 1 hora por dia.
(D) Entre 1 e 2 horas por dia.
(E) Entre 2 e 3 horas por dia.
(F) Entre 3 e 4 horas por dia.
(G) Entre 4 e 5 horas por dia.
(H) Mais de 5 horas por dia.

21. Você gosta de ler?

(A) Gosto muito.
(B) Gosto.
(C) Gosto pouco.
(D) Não gosto.

22. Qual das disciplinas exige mais tarefas de casa?

(A) Matemática.
(B) Português.

23. Você consegue entender tudo que seu professo ensina?

(A) Sim, sempre entendo.
(B) Sim, consigo entender a maior parte.
(C) Consigo entender pouco.
(D) Nunca entendo.

Questionário do aluno da 3ª Série do Ensino Médio

01. Mês e ano de nascimento:

Mês ☐☐ / Ano ☐☐☐☐

02. Qual o seu sexo?
(A) Masculino.
(B) Feminino.

03. Com quem você mora?
(A) Com a própria família.
(B) Só com o pai.
(C) Só com a mãe.
(D) Com parentes.
(E) Com amigos.
(F) Com o pai ou a mãe em nova união.

04. Qual o nível de instrução do seu pai?
(A) Nunca freqüentou a escola.
(B) Ensino Fundamental – 1ª a 4ª série.
(C) Ensino Fundamental – 5ª a 8ª série.
(D) Ensino Médio.
(E) Superior.
(F) Pós-graduação.
(G) Não sei.

05. Qual o nível de instrução de sua mãe?
(A) Nunca freqüentou a escola.
(B) Ensino Fundamental – 1ª a 4ª série.
(C) Ensino Fundamental – 5ª a 8ª série.
(D) Ensino Médio.
(E) Superior.
(F) Pós-graduação.
(G) Não sei.

06. Seu pai (ou responsável) está trabalhando atualmente?
(A) Sim.
(B) Não, ele está desempregado.
(C) Não, ele é aposentado.

07. Sua mãe está trabalhando atualmente? *(Trabalho remunerado)*
(A) Sim.
(B) Não, ela está desempregada.
(C) Não, ela é aposentada.
(D) Não, ela não trabalha fora.
(E) Não sei.

08. Você já deixou de freqüentar a escola em algum período?
(A) Sim, por 1 ano.
(B) Sim, por 2 anos.
(C) Sim, por 3 anos.
(D) Sim, por 4 anos.
(E) Sim, por mais de 4 anos.
(F) Não.

09. Caso você tenha abandonado a escola por algum período, qual foi o motivo?
(A) Doença/problemas de saúde.
(B) Trabalho/recursos financeiros.
(C) Falta de interesse, motivação.
(D) Falta de vagas.
(E) Reprovação(ões).
(F) Nunca abandonei a escola.

10. Você já repetiu de ano?
(A) Não.
(B) Sim, 1 vez.
(C) Sim, 2 vezes.
(D) Sim, 3 vezes.
(E) Sim, 4 vezes.
(F) Sim, 5 vezes.
(G) Sim, 6 vezes.
(H) Sim, 7 vezes.
(I) Sim, 8 vezes.
(J) Sim, 9 vezes.
(K) Sim, 10 vezes.
(L) Sim, 11 vezes ou mais.

11. Que séries você repetiu?
(marque todas as séries que você repetiu)

ENSINO FUNDAMENTAL	(A) 1ª (E) 5ª	(B) 2ª (F) 6ª	(C) 3ª (G) 7ª	(D) 4ª (H) 8ª
ENSINO MÉDIO	(I) 1ª	(J) 2ª	(K) 3ª	

12. Você costuma conversar em casa sobre o que acontece na escola?
(A) Sim, quase todos os dias.
(B) Sim, só uma vez por semana.
(C) Sim, só quando recebo as notas.
(D) Sim, só no final do ano.
(E) Não converso.

13. Seu pai, sua mãe ou responsável participa(m) das reuniões da escola?
(A) Sim, sempre.
(B) Sim, às vezes.
(C) Não, nunca.
(D) Não. A escola não chama para reuniões.

14. Em sua casa se compra jornal?
(A) Sim, todos os dias.
(B) Sim, de vez em quando.
(C) Sim, em fins de semana.
(D) Não.

15. Quantas horas por dia você assiste a programas de televisão?
(A) Não assisto à televisão.
(B) Menos de 2 horas.
(C) De 2 a 4 horas.
(D) De 5 a 6 horas.
(E) Mais de 6 horas.

16. Você sabe utilizar computador?
(A) Não sei usar.
(B) Sei usar só para jogos.
(C) Sei usar para fazer trabalhos (edição de textos e/ou planilha eletrônica).

17. Em sua escola, você utiliza computador?
(A) Não, na escola não tem computador.
(B) Na escola tem computador, mas não uso.
(C) Sim, na escola tem computador, mas o utilizo poucas vezes.
(D) Sim, na escola tem computador e o utilizo freqüentemente.

18. E, em sua casa, você utiliza computador?
(A) Não, em minha casa não tem computador.
(B) Em minha casa tem computador, mas eu não utilizo.
(C) Sim, em minha casa tem computador, mas o utilizo poucas vezes.
(D) Sim, em minha casa tem computador e o utilizo freqüentemente.

19. Qual das duas provas você achou mais difícil?
(A) Matemática.
(B) Português.

20. Quantas horas, por dia, você dedica, em média, às tarefas da escola (trabalhos encomendados pelos professores, leitura ou estudo)?
(A) Nenhuma, porque os professores da escola não dão tarefas.
(B) Nenhuma, porque não gosto de fazer tarefas.
(C) Menos de 1 hora por dia.
(D) Entre 1 e 2 horas por dia.
(E) Entre 2 e 3 horas por dia.
(F) Entre 3 e 4 horas por dia.
(G) Entre 4 e 5 horas por dia.
(H) Mais de 5 horas por dia.

21. Você gosta de ler?
(A) Gosto muito.
(B) Gosto.
(C) Gosto pouco.
(D) Não gosto.

22. Qual das disciplinas exige mais tarefas de casa?
(A) Matemática.
(B) Português.

23. Você consegue entender tudo que seu professo ensina?
(A) Sim, sempre entendo.
(B) Sim, consigo entender a maior parte.
(C) Consigo entender pouco.
(D) Nunca entendo.