UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL

INSTITUTO DE MATEMÁTICA

DEPARTAMENTO DE ESTATÍSTICA

MODELOS MULTINÍVEIS:

CARACTERIZAÇÃO E APLICAÇÃO

KARINA PRETTO

ORIENTADOR: JOÃO RIBOLDI

Monografia apresentada para obtenção

do grau de Bacharel em Estatística

Porto Alegre, fevereiro de 2003.

AGRADECIMENTOS

Agradeço a todos que, de alguma forma, contribuíram para a realização deste trabalho.

Em especial agradeço ao professor João Riboldi, pelo incentivo, atenção e convivência, que foram essenciais durante o decorrer do curso, principalmente na execução deste trabalho. Agradeço pela oportunidade de aprender e trabalhar, sob sua orientação, no LEMAE.

À professora Stela Castro, agradeço pela atenção, amizade e orientação durante o período de iniciação científica.

Aos demais professores, agradeço pelos conhecimentos transmitidos neste período. E acima de tudo, obrigada pela oportunidade de conhecer, descobrir e admirar a ciência Estatística.

Agradeço ao Professor Bruce B. Duncan pela permissão para a utilização do banco de dados e assim enriquecer o conteúdo deste trabalho.

Aos meus amigos e colegas, Letícia, Mariana, Michele, Fernando, Talita e Patrícia, que sempre se fizeram presentes em todos os momentos alegrando estes quatro anos de convivência. Um obrigado especial aos meus queridos amigos, Luana, Leonardo e Daniel, que serão sempre meus irmãos de coração, e por quem tenho profundo respeito e admiração.

À minha prima e amiga Angélica, que sempre me apoiou e acreditou em mim e na minha escolha pela Estatística.

Aos meus queridos pais, Pedro e Dejane, e a minha irmã Natália, pelo carinho, respeito e apoio incondicional que sempre recebi, sendo essencial para que eu pudesse enfrentar todos as eventuais dificuldades.

ÍNDICE

1. INTRODUÇÃO

A seleção de um modelo que capte de maneira mais realista possível a variabilidade de um conjunto de dados é um dos principais pontos de uma análise estatística. Em alguns casos temos a presença de uma estrutura de hierarquia nos dados que deve merecer maior atenção por parte do pesquisador a fim de reproduzir de forma mais realista o comportamento das observações.

Modelos multiníveis, designação dada muitas vezes a modelos lineares hierárquicos, têm sua importância e utilidade destacadas por pesquisadores das mais diversas áreas do conhecimento, pois tratam justamente da modelagem de estudos em que os dados apresentam-se de forma hierárquica. Estes modelos caracterizam-se por propiciar a explicação da variabilidade da variável resposta, através de variáveis preditoras incluídas em diferentes níveis hierárquicos, e, além disso, possibilitam comparação direta entre os níveis.

Um problema multinível se preocupa com as relações entre as variáveis que são medidas em um número diferente de níveis hierárquicos. Eles são particularmente utilizados em pesquisas complexas que envolvem amostragem de unidades agrupadas ou delineamentos em múltiplos estágios e, além disso, a população em estudo possui estrutura hierárquica. O objetivo da análise multinível, então, é determinar o efeito direto das variáveis explicativas do indivíduo e do grupo, e determinar se as variáveis explicativas no mesmo nível do grupo servem como indicadores de relações do nível-indivíduo. Se o grupo das variáveis de um nível apresenta relações com os níveis inferiores, pode-se encarar como uma interação estatística entre as variáveis explicativas de diferentes níveis.

No passado, tais dados eram geralmente analisados usando análise de regressão múltipla clássica com uma variável dependente no menor nível (indivíduo) e uma coleção de variáveis explicativas de todos os níveis disponíveis (Hox, 1995).

A análise de dados que apresentam uma estrutura hierárquica é descrita sob vários nomes na literatura, tais como modelos hierárquicos, modelos de coeficientes aleatórios, modelos de curvas latentes, modelos de curvas de crescimento ou modelos

multiníveis. Estes modelos são mais flexíveis que os tradicionais modelos lineares generalizados, pois permitem a análise de conjunto de dados, cujas observações não são independentes ou apresentam-se agrupadas.

1.1. CONCEITOS DE MACRO E MICRO NÍVEIS

Dados com estrutura hierárquica, caracterizam-se por exemplo, quando sujeitos estão arranjados segundo tipos ou grupos. Desta forma, a hierarquia consiste em observações de níveis inferiores estarem agrupadas em unidades de níveis mais elevados. Por exemplo, alunos (nível 1) arranjados em salas de aula (nível 2) e estas, por sua vez pertencentes a escolas (nível 3).

As medidas dos níveis mais inferiores são também chamadas de medidas de *micro nível*. Já medidas referentes a unidades de níveis mais elevados são denominadas como medidas de *macro nível* e se referem, geralmente, aos grupos ou contextos. (Kreft & Leeuw, 1998).

Relações que envolvem macro e micro níveis são de grande interesse e são conhecidas como relações de níveis cruzados. Interessa saber se variáveis de níveis mais elevados interagem com aquelas pertencentes ao nível inferior.

1.2. TRATAMENTO DE DADOS AGRUPADOS

Em situações onde há a existência de dados agrupados, técnicas que supõem a independência entre as observações não são as mais adequadas, uma vez que se espera que sujeitos que pertençam ao mesmo grupo possuam uma semelhança maior, ou seja, estejam mais correlacionados do que sujeitos pertencentes a grupos diferentes. Isto porque, os dados correlacionados originalmente aparecem quando as observações em uma amostra não são selecionadas aleatoriamente uma da outra, ou seja, unidades primárias de amostragem são selecionadas e posteriormente dentro de cada uma delas seleciona-se novamente unidades secundárias.

A falta de independência das observações pode ser conseqüência do delineamento amostral, ou então do delineamento do estudo, como por exemplo, em estudos que envolvem múltiplas observações na mesma unidade experimental. Outro fator que influencia é a distribuição espacial da população em estudo, unidades mais próximas no espaço tendem a estar mais correlacionadas.

A utilização das estimativas de modelos lineares clássicos, nas condições de falta de independência das observações deve ser feita cautelosamente. Com o avanço computacional, soluções foram encontradas para este tipo de situação. A possibilidade de lidar com estes problemas têm estimulado o desenvolvimento de estudos, especialmente estudos longitudinais, que coletam dados correlacionados, às vezes em delineamentos muito complexos. A avaliação destes dados bem como a implementação de soluções mais poderosa para estes problemas têm desenvolvido novas ferramentas estatísticas (Laplante, 1999).

Em estudos que apresentam estrutura de hierarquia ou agrupamento nas observações temos fontes de variabilidade referentes a cada nível observado. Temos diferenças entre os sujeitos pertencentes ao mesmo grupo, em relação a sua média e também diferenças entre grupos de sujeitos. Dessa forma, quando temos amostras agrupadas, a variância da variável fica decomposta em dois componentes, conhecidos como componentes de variância, um referente ao processo aleatório (erro de amostragem) entre os grupos e outro originário do fato de indivíduos selecionados em um mesmo grupo tendem a estarem relacionados. Este raciocínio é similar ao utilizado na análise de variância tradicional.

1.3. POSSÍVEIS ABORDAGENS MULTINÍVEIS

Muitas análises estatísticas tradicionais podem ser abordadas segundo uma visão multinível, tais como a Meta-análise, Regressão Logística, Modelos Multivariados, Modelos de Equações Estruturais, dentre outras.

1.3.1. META-ANÁLISE

A Meta-análise é uma metodologia estatística que integra os resultados de vários estudos independentes, considerando que os mesmos possam ser combinados, fornecendo a estimativa do efeito do tratamento e permitindo calcular a heterogeneidade entre os resultados dos estudos individuais. Segundo Hox (1995), a idéia básica é aplicar métodos estatísticos formais para os resultados de um conjunto específico de estudos, aumentando assim a precisão das conclusões da literatura, fazendo, de forma objetiva e acurada, comparação entre as diferentes intervenções, e desta forma, auxiliando a resolver controvérsias e possibilitando aos pesquisadores uma determinada estratégia de ação. Se os resultados dos estudos não se diferenciam muito, aplicam-se procedimentos estatísticos que combinam todos os resultados em uma única resposta média. Caso os estudos variem muito, o principal objetivo da meta-análise passa a ser o de responder por que os resultados variam. Assim, a análise considera os diferentes resultados como conseqüências das diferenças entre as características dos estudos (tais como delineamentos utilizados ou sujeitos avaliados).

Na meta-análise, uma questão preliminar é avaliar se os resultados se diferenciam mais uns dos outros do que o correspondente pela variação amostral que é esperada (dada pelos tamanhos das amostras dos estudos). Se os resultados não se diferenciam mais que o esperado eles são ditos *homogêneos*, indicando que eles vêm de uma mesma população. Dessa forma, a próxima etapa seria estimar um valor comum para o parâmetro de interesse da população. No caso de haver diferenças acima das esperadas, os estudos são denominados *heterogêneos* sugerindo que vem de populações diferentes. Nesta situação, o objetivo principal não é encontrar um valor comum para os estudos, mas sim analisar os excessos de variação como uma função das características conhecidas dos estudos (características amostrais dos estudos) ou características metodológicas (referente à qualidade metodológica do estudo).

Há diversos métodos para analisar e combinar resultados de estudos separados. A semelhança com o problema multinível está na combinação de vários estudos considerando pequenos modelos de diferentes grupos ou contextos. A meta análise pode ser vista também como um modelo linear hierárquico com dois níveis (Bryk e Raudesbush, 1988,1992). Em cada estudo, um modelo dentro do estudo é estimado, e

em um segundo nível, um modelo entre estudos é adicionado para explicar a variação dos parâmetros dentro do estudo como função das diferenças entre os estudos.

1.3.2. MODELOS DE EQUAÇÕES ESTRUTURAIS

O enfoque de Equações Estruturais surge quando há grande dificuldade em se medir uma resposta diretamente, assim, pode-se pensar em medir um número de indicadores que forneçam a informação necessária para a realização do estudo.

Uma qualidade da modelagem de equações estruturais é a de possuir suposições pouco restritivas nos seus modelos, e por isso apresenta-se como uma técnica cada vez mais utilizada. Os modelos de equações estruturais não analisam as observações individuais em sua forma bruta, mas sim fazem uso das covariâncias, ou correlações obtidas entre todas as variáveis avaliadas. O modelo completo de equações estruturais contém várias variáveis, aleatórias ou latentes e, em menor número, variáveis não aleatórias ou fixas. Ainda possuem parâmetros estruturais, que fazem a ligação entre as variáveis e fornecem uma relação casual entre elas.

Principalmente em pesquisas da área das ciências sociais encontramos a necessidade de utilizar modelos de equações estruturais, e geralmente encontramos os objetos em estudo, por exemplos pessoas, dispostos de forma hierárquica, daí a necessidade de considerar uma estrutura multinível na análise. Considere por exemplo um modelo com dois níveis, sendo que em cada um deles seja descrito por um conjunto de variáveis. Pode ocorrer ainda que alguma variável associada ao nível 1 possa aparecer também no nível 2 do modelo, assim um conjunto de equações simultâneas ou estruturais estaria presente no estudo. Um procedimento para a estimação dos parâmetros desse modelo pode ser composto por dois estágios. O primeiro a estimação das matrizes de covariâncias em cada um dos níveis separadamente e o segundo envolve uma análise fatorial destas matrizes separadas usando qualquer procedimento padrão. Caso os dados não sejam balanceados este procedimento não se mostra muito eficiente, no entanto tem uma vantagem em que pode ser usado para equações estruturais fechadas (Goldstein,1995). Este procedimento estende-se também a modelos com mais de dois níveis de hierarquias ou ainda quando se pretende ajustar um modelo de

trajetórias incondicionais, com ou sem variáveis latentes, isso, é claro, desde que as matrizes de covariâncias em cada um dos níveis seja suficientes para estes modelos.

1.3.3. MODELOS PARA RESPOSTA DISCRETA

Existem muitos tipos de modelagem estatística que trabalham com dados de respostas discretas, tais como eventos de contagem. Esta contagem pode ser o número de vezes que um evento ocorre dentro de um número fixo de realizações. Por exemplo, temos a proporção de mortes em uma população classificada por idade. Pode-se ainda ter um vetor de contagem representando o número de diferentes tipos de eventos que ocorrem dentro de um total de eventos.

Métodos de regressão linear não são adequados para o ajuste de variáveis discretas, uma vez que podem gerar resultados fora da amplitude aceitável. Uma alternativa é o uso da técnica de regressão logística, indicada principalmente para casos de resposta dicotômica, ou a regressão de Poisson, indicada a casos em que a resposta avaliada é proveniente de dados de contagem.

Ambas regressões podem ser escritas como um modelo linear generalizado, e na presença de estrutura hierárquica no problema a ser analisado pode-se fazer uma extensão do modelo multinível chamando-o de Modelo Multinível Generalizado ou Modelo Hierárquico Generalizado.

1.3.3.1. REGRESSAO LOGÍSTICA MULTINÍVEL

A estrutura básica de dados para uma regressão logística, com dois níveis, é um conjunto de dados de N grupos, representando as unidades do segundo nível, contendo uma amostra aleatória de n_j unidades, ou indivíduos, representantes do primeiro nível de hierarquia. O índice **j** diz respeito ao j-ésimo grupo e assume os valores de 1 a N. Já cada individuo é identificado pelo subscrito **i** e varia de i=1 ate o total de indivíduos amostrados no grupo j, denotado por n_j. A variável resposta (Y_{ij}) é dicotômica, e seus valores representam falha(0) e sucesso(1). Ainda, a probabilidade de sucesso (P_j) é

constante em cada grupo. Em um modelo de coeficientes aleatórios os grupos são considerados como sendo extraídos de uma população de grupos e as probabilidades de sucesso referentes aos grupos (P_j) são tidas como variáveis aleatórias definidas nesta população. Maiores detalhes podem ser encontrados em Snijders & Bosker (1999).

1.3.3.2. REGRESSAO MULTINÍVEL DE POISSON

A análise de dados de contagem tem como importante distribuição de probabilidade a distribuição de Poisson, que é uma aproximação da distribuição binomial para a situação que o número de experimentos é grande e a probabilidade de sucesso é baixa. A distribuição de Poisson possui uma relação direta entre sua média e sua variância, onde ambas são iguais.

A distribuição de Poisson pode ser aproximada para uma distribuição contínua, desde que um grande número de contagens sejam registradas. Se for possível aproximação, um modelo linear hierárquico pode ser estimado sem nenhum problema, mas no caso do tamanho da amostra ser pequeno, se faz necessária a utilização de um modelo hierárquico generalizado.

1.3.4. MODELOS MULTINÍVEIS MULTIVARIADOS

Quando se deseja modelar simultaneamente mais que uma variável resposta como função de um conjunto de variáveis explicativas, é indicado trabalhar com modelos multivariados. No caso do conjunto de dados apresentar-se hierarquicamente, é aconselhável adotar um modelo que leve em consideração este aspecto do problema, ou seja, aconselha-se adotar um modelo multinível. Trabalha-se com variáveis no nível 1 de indivíduos e com variáveis do nível 2, referentes aos grupos em que os sujeitos estão organizados.

Algumas das razões para se trabalhar com o conjunto de variáveis dependentes simultaneamente na análise é o fato de que as conclusões podem ser tiradas sobre a correlação entre as variáveis dependentes e estender para as correlações dependentes ao

nível de indivíduos ou de grupos, devido à partição da variâncias entre as variáveis dependentes nos níveis de análise. Os testes para os efeitos para uma única variável dependente são mais poderosos na análise multivariada, alem disso se é objetivo testar efeitos conjuntos de variáveis explicativas em várias variáveis dependentes, é mais recomendado testes multivariados. No entanto, embora a análise multivariada seja mais indicada, aconselha-se inicialmente trabalhar com uma análise univariada para cada um das variáveis dependentes e somente depois unir todas elas em uma única análise. Exemplos e maiores detalhes podem ser encontrados em Snijders & Bosker (1999).

1.3.5. ESTUDOS MULTICÊNTRICOS

Algumas pesquisas científicas podem gerar resultados diferentes, o que pode gerar contradição entre os pesquisadores. Unificar resultados de estudos é muito interessante, assim, uma conclusão mais ampla poderá ser obtida e possíveis resultados ambíguos eliminados. O tamanho da amostra neste tipo de estudos também se apresenta maior, uma vez que ocorre simultaneamente em diferentes lugares.

A área médica é a que apresenta maior concentração de estudos multicêntricos, que surgiram da necessidade de responder a questões que se mantinham controvertidas nos estudos básicos, na revisão sistemática e na meta-análise. Por definição, são trabalhos com uma amostra grande, em geral mais de 1000 pacientes por centro. Neste tipo de estudo diversos centros, localizados em diferentes áreas, realizam o mesmo experimento, com o intuito de coletar evidências significativas sobre o estudo em andamento. Assim, incorpora-se no estudo uma maior generalização nos resultados, uma vez que a diferença existente entre os pacientes avaliados vai além da variabilidade intrínseca de cada indivíduo. Outros fatores como classe social, costumes regionais e equipe envolvida podem estar se destacando na variabilidade da resposta final.

A modelagem multinível se torna uma ferramenta muito útil neste caso, pois é possível considerar um nível de hierarquia como sendo o centro onde foi desenvolvido a pesquisa e no nível inferior a este primeiro o paciente avaliado e em seguida suas medidas (no caso de medidas repetidas).

1.4. OBJETIVOS

O objetivo geral deste trabalho é apresentar de forma consistente a estruturação básica de modelos multiníveis e enfocar, de forma específica, uma de suas inúmeras aplicações.

O objetivo específico é, utilizando dados de um estudo epidemiológico, que buscam relacionar dados de níveis de glicose em gestantes, participantes de um estudo brasileiro multicêntrico, ilustrar uma aplicação da metodologia multinível, em dados que envolvem estudos multicêntricos e medidas repetidas.

1.5. ESTRUTURA DA MONOGRAFIA

No capítulo inicial apresenta-se, de forma simplificada, os modelos de análise de covariância e de coeficientes aleatórios.

No capitulo 3, apresenta-se de forma estruturada, a metodologia de modelos multiníveis, enfocando-se conceitos fundamentais, suposições, métodos de estimação, hipóteses testadas e estudos de diagnóstico.

No capitulo 4 são apresentados alguns softwares utilizados para o ajuste de modelos multiníveis, dentre eles, o SAS e o MlwiN.

No capitulo 5 ilustra-se uma aplicação da metodologia multinível em um estudo epidemiológico que busca relacionar o nível de glicose de gestantes participantes de um estudo multicêntrico, com medidas repetidas. Utilizando a metodologia multinível para analisar dados de medidas repetidas, pretende-se relacionar estes dados com algumas covariáveis tais como altura, idade, índice de massa corporal, dentre outras.

No capítulo final são apresentadas conclusões deste trabalho..

2. MODELOS DE ANÁLISE DE COVARIÂNCIA E DE COEFICIENTES ALEATÓRIOS

Os modelos de regressão tradicionais apresentam diferentes maneiras de decomposição da variação presente nos dados. Quando variáveis referentes ao indivíduo e ao grupo são consideradas na construção do modelo, esta análise passa a ser chamada de análise contextual, pois o interesse da pesquisa concentra-se tanto no enfoque do indivíduo como no enfoque do grupo a que o sujeito pertence.

Alguns conceitos de Modelos de Análise de Covariância e de Coeficientes Aleatórios são utilizados em análise multinível e são exemplos de análises contextuais.

2.1. ANÁLISE DE COVARIÂNCIA (ANCOVA)

A análise de covariância (ANCOVA) é uma forma tradicional de análise de dados agrupados, e focaliza mais seus resultados nos efeitos de grupo, estes tratados como fatores, do que nos efeitos individuais, cuja função destes é atuar como covariáveis.

O modelo de ANCOVA foi desenvolvido inicialmente para o auxílio na análise de delineamentos experimentais , onde é muito comum a necessidade de inclusão de uma variável auxiliar, ou covariável, em um modelo a fim de ajustar de forma mais adequada o comportamento da resposta de interesse. Outros exemplos de utilização de covariáveis podem ser encontrados em estudos educacionais onde se pode estar interessado em ajustar um modelo para a proficiência de um aluno segundo a escola que ele estuda. Uma covariável importante a ser considerada neste caso é o número médio de anos de experiência dos professores das escolas. Em estudos agronômicos que avaliam o rendimento de uma planta frente à aplicação de diferentes doses de adubos usualmente utiliza-se o número de plantas por parcela como covariável. Na epidemiologia, o uso de covariáveis é muito freqüente, como por exemplo, em estudos sobre o estado nutricional de crianças entre diferentes classes sociais uma covariável de interesse seria o tempo de amamentação da criança. No Estudo Brasileiro de Diabetes

Gestacional (EBDG) inúmeras covariávies, tais como idade, peso, índice de massa corporal foram medidas a fim de auxiliar a modelagem do comportamento do nível de glicose em gestantes.

Os Modelos de análise de Covariância (ANCOVA) utilizam variáveis auxiliares na interpretação dos dados referentes a uma variável de interesse. Estes modelos combinam conceitos de análise de variância e análise de regressão. Este procedimento é muito útil no fornecimento de estimativas, intervalos de confiança e testes de hipóteses baseados no modelo linear normal. No entanto, a ANCOVA não é restrita somente ao modelo linear normal, o aumento de precisão pela introdução de covariáveis, é, por exemplo, possível também em classes mais amplas de modelos, inclusive em modelos lineares generalizados (Nelder & Wedderbun, 1972).

Nos modelos de ANCOVA , assume-se que a relação de X e Y seja linear e que o coeficiente de inclinação seja o mesmo para todas os grupos (tratamentos). Ainda que os grupos não tenham efeitos sobre a variável explicativa X adicionada ao modelo. O modelo linear para ANCOVA, apresentado em (1) considera um experimento com t níveis de um único fator, instalado no delineamento completamente casualizado e é descrito como:

$$Y_{ij} = \mu + \tau_i + \beta\ (X_{ij} - \overline{X..}) + \varepsilon_{ij} \qquad (1)$$

onde Y_{ij}: observação da variável resposta para a j-ésima unidade experimental do i-ésimo tratamento;

μ é a média geral do modelo;

τ_i mede o efeito do i-ésimo tratamento;

β é o coeficiente de inclinação (se for nulo tem-se um modelo de Análise de Variância tradicional);

$(X_{ij} - \overline{X..})$ é a covariável centrada em relação à sua própria média.

ε_{ij} é o erro do modelo

No caso do coeficiente de inclinação da covariável adotada ser considerado aleatório, temos um modelo de coeficientes aleatórios combinado com um modelo de análise de covariância.

2.2. MODELOS DE COEFICIENTES ALEATÓRIOS

A linguagem de efeitos fixos ou aleatórios é oriunda da pesquisa experimental, quando um pesquisador está avaliando um conjunto de tratamentos em sua pesquisa. Se todos os tratamentos de interesse do pesquisador estiverem presentes no experimento dizemos que este modelo é de efeitos fixos, enquanto que o modelo é dito de efeitos aleatórios se uma amostra dos possíveis tratamentos foi selecionada para participar do estudo. A inferência cabível em cada um dos modelos também é diferente, pois no caso de efeitos fixos, só será possível inferir para os tratamentos pesquisados, enquanto que quando se trabalha com efeitos aleatórios é possível generalizar os resultados para toda a população de possíveis tratamentos.

É possível avaliar uma variável também como sendo fixa ou aleatória, que é diferente do fato de um modelo possuir coeficientes fixos ou aleatórios. Uma variável é aleatória se possui uma distribuição de probabilidade enquanto que é dita fixa, se seus valores são conhecidos antecipadamente.

Um modelo de regressão pode também conter coeficientes de regressão fixos ou aleatórios ou ainda apresentar simultaneamente os dois tipos de coeficientes. A escolha de um ou outro tipo de coeficiente vai depender do comportamento da variável resposta dentro e entre as unidades de nível mais elevado.

O modelo de regressão tradicional representado em (2), expressa funcionalmente como a média ou resposta média de uma variável dependente(Y) varia com uma variável independente, ou explicativa (X).

$$Y = \beta_0 + \beta_1 X + \varepsilon \qquad (2)$$

Uma representação gráfica deste modelo apresentada pela Figura 1, apresenta uma relação linear direta entre as variáveis X e Y.

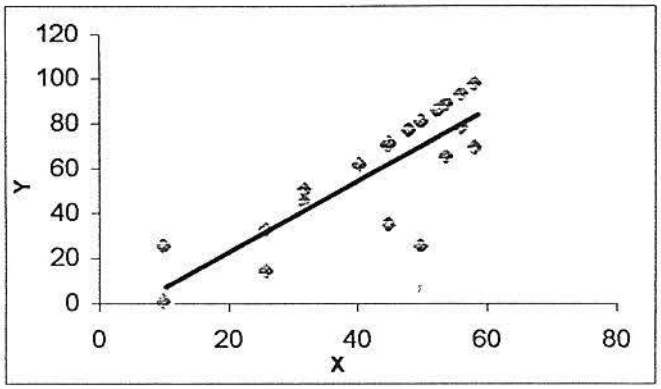

Figura1. Modelo de regressão tradicional

Outros modelos, tais como os modelos de coeficientes aleatórios, permitem maior flexibilização na estimação da equação de regressão, como por exemplo, estimar um modelo que permita a variação do intercepto entre os grupos que compõem a amostra. Os modelos de coeficientes aleatórios permitem ajustar modelos de interceptos ou de coeficientes de inclinação aleatórios ou ainda ambos aleatórios.

Quando a resposta média de uma variável se altera de um grupo a outro, ou seja, sofre maior ou menor variação dependendo dos valores assumidos pela variável explicativa entre os grupos considerados no estudo, tem-se um exemplo de modelo com inclinação aleatória, cuja equação deste tipo de modelo é apresentada em (3).

$$Y = \beta_0 + (\beta_1 + \zeta)X + \varepsilon \qquad (3)$$

Conforme apresentado na Figura 2, cada uma das equações estimadas possuem o mesmo intercepto, no entanto diferentes inclinações, indicando que alguns grupos apresentam relações mais fortes que outros entre as variáveis X e Y.

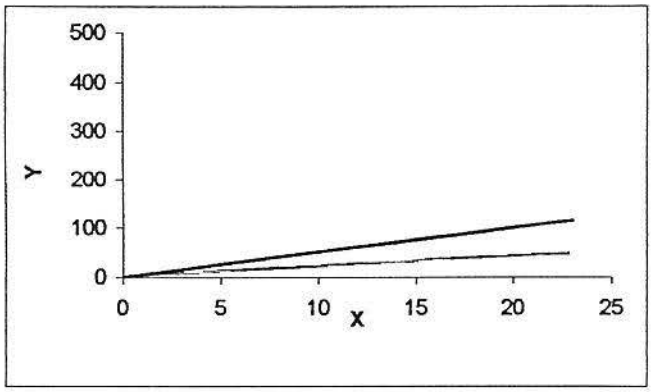

Figura 2. Modelo com Inclinação Aleatória

Podem ainda existir relações entre variáveis dependente e explicativa em que um modelo de intercepto aleatório (4) é de interesse, uma vez que neste tipo de modelo, a variação da variável explicativa sobre a variável independente é igual para todos os grupos e o que difere de um grupo para outro é a resposta média global.

$$Y = (\beta_0 + \varsigma) + \beta_1 X + \varepsilon \qquad (4)$$

Neste caso obtém-se um gráfico de linhas paralelas, apresentado pela Figura 3, e nesta situação temos uma semelhança muito grande com o modelo de ANCOVA, onde interceptos desiguais, mas relações semelhantes entre Y e X são assumidas, com a diferença que uma distribuição subjacente é admitida para os interceptos.

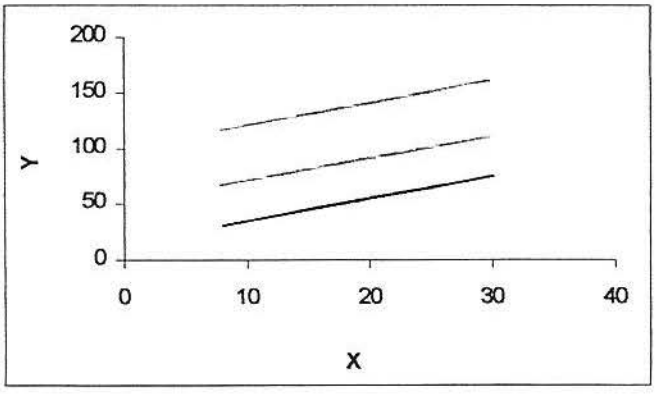

Figura 3. Modelo com Intercepto Aleatório.

Uma situação mais realista, na maioria dos casos é apresentada em (5), onde se admite que ocorra variação tanto no intercepto quanto no coeficiente de inclinação.

$$Y = (\beta_0 + \varsigma) + (\beta_1 + \zeta)X + \varepsilon \qquad (5)$$

Neste caso temos diferentes influências da variação da variável X na variação média da variável Y, conforme observa-se na Figura 4.

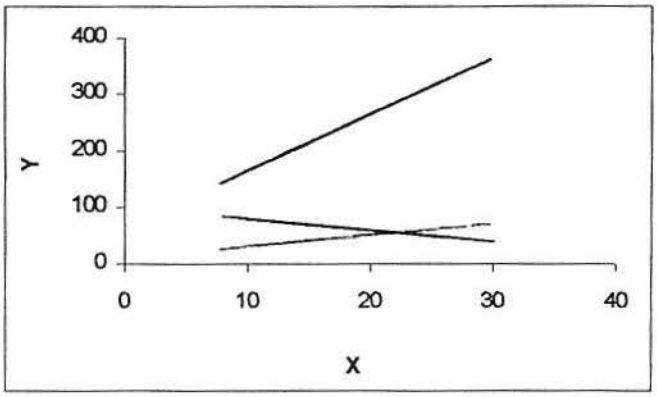

Figura 4. Modelo com Intercepto e Inclinação Aleatórios.

Em modelos de coeficientes aleatórios, considera-se que os valores atuais da variável X foram aleatoriamente amostrados de uma população de valores e que há razões para crer que os efeitos de X em Y e que as estimativas $\hat{\beta}_0$ e $\hat{\beta}_1$ poderiam variar de amostra a amostra de valores de X. Dessa forma, é possível estimar diferentes modelos com coeficientes de inclinações, ou intercepto, ou ambos, diferentes para cada grupo presente na análise, permitindo assim estimar um modelo cujo intercepto assuma diferentes valores no modelo (4); ou ainda, estimar um modelo com diferentes inclinações como no modelo (3); além disso estimar um modelo que ambos coeficientes, de inclinação e intercepto variem como no modelo (5).

O coeficiente resultante de um modelo de coeficientes aleatórios consiste de duas partes: uma média, ou parte fixa, e uma variância, ou parte aleatória. A porção aleatória refere-se ao desvio de uma solução global e é também chamada de variância macro, ou nível macro de variância, pois os coeficientes se diferem um dos outros em um contexto de níveis mais elevados ou níveis macro.

Os coeficientes em um modelo de coeficientes aleatórios são estimados como efeitos principais com uma variância em torno deles, que representa o desvio do contexto de todos os demais. Podem ser definidos como componentes fixas mais um distúrbio, com média zero, independentes das perturbações do nível mais inferior.

O modelo apresentado em (5) pode ser descrito como:

$$Y_{ij} = \beta_{0i}^{*} + \beta_{1i}^{*} X_{ij} + \varepsilon_{ij} \qquad (6)$$

onde Y_{ij} é a resposta do i-ésimo indivíduo do j-ésimo grupo;

$\beta_{0i}^{*} e \beta_{1i}^{*}$ são respectivamente o intercepto e o coeficiente de inclinação, ambos aleatórios;

ε_{ij} representa o erro do modelo;

Assume-se neste modelo que:

$$\begin{pmatrix} \beta_{0i}^{*} \\ \beta_{1i}^{*} \end{pmatrix} \overset{iid}{\sim} Normal\left[\begin{pmatrix} \beta_{0i} \\ \beta_{1i} \end{pmatrix}, \psi \right], \quad \text{onde} \quad \psi = \begin{pmatrix} \sigma_a^2 & \sigma_{ab} \\ \sigma_{ab} & \sigma_b^2 \end{pmatrix}$$

e

$$\varepsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2) \qquad (7)$$

A estrutura básica de modelos de coeficientes aleatórios, sustenta-se numa estrutura similar a adotada em modelos de análise de covariância. Nesse caso modelos de regressão linear são utilizados para incluir variáveis contínuas (covariáveis) como variáveis independentes. Este modelo adota ainda um conceito de coeficientes de inclinação como variáveis resposta. Se, primeiramente, o pesquisador pretende concluir para cada um dos grupos específicos (tratamento), utilizados no estudo ou experimento é apropriado o uso de análise de covariância. Já a utilização de um modelo de coeficientes aleatórios deve ser feita quando as conclusões do estudo pretendem ser generalizadas para toda a população de tratamentos e não somente para os testados ou ainda se o pesquisador deseja testar os efeitos das variáveis do nível macro.

Em relação a pequenas amostras, o modelo de coeficientes aleatórios tem importante vantagem sobre a análise de covariância, uma vez que a suposição sobre a aleatoriedade dos coeficientes é razoável. O modelo de coeficientes aleatórios inclui uma suposição extra de independência e distribuição idêntica para os efeitos de grupo e na maioria das análises supõe-se que os termos aleatórios do modelo sigam uma distribuição normal.

3. MODELOS LINEARES HIERÁRQUICOS

3.1. DEFINIÇÃO

A presença de hierarquia na informação estudada é muito comum nas diferentes área de estudo. Temos um estudo hierárquico quando possuímos variáveis que descrevem indivíduos, os quais estão agrupados em unidades de níveis mais elevados. Por exemplo em estudos educacionais, os sujeitos avaliados, os alunos, estão agrupados em salas de aula; em estudos no campo industrial, quando são avaliados os funcionários agrupados em setores; em estudos epidemiológicos, com pacientes agregados em centros de pesquisa ou então medidas repetidas em um mesmo indivíduo. Ainda identifica-se a presença de hierarquia em pesquisas econômicas, e de marketing, considerando os consumidores pertencentes a diferentes regiões geográficas, dentre muitas outras aplicações.

O diagrama representado na Figura 5 exemplifica a caracterização de uma estrutura hierárquica em um estudo epidemiológico, em que foi avaliado um desfecho de interesse em pacientes de diferentes hospitais. Acredita-se que pacientes avaliados em um mesmo hospital sejam mais semelhantes do que pacientes atendidos em locais diferentes.

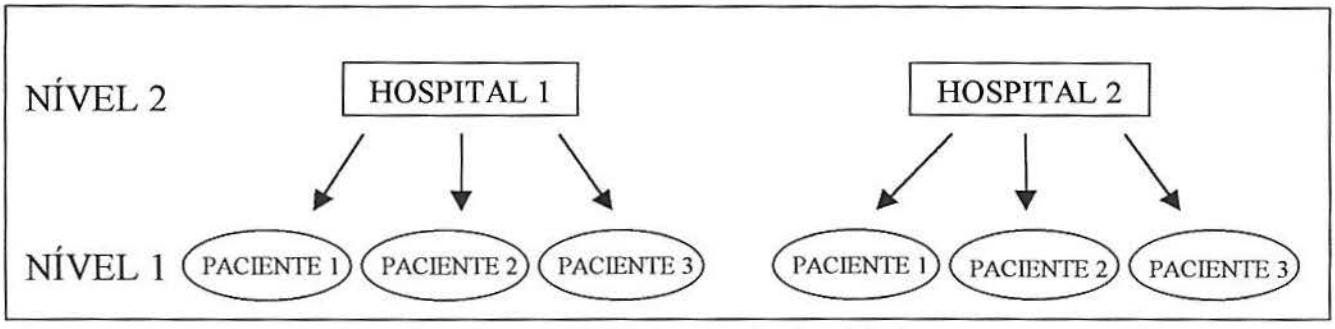

Figura 5. Representação de uma estrutura hierárquica.

Uma classe de modelos que se preocupa em avaliar a relação das observações na presença de hierarquia é a dos modelos lineares hierárquicos. Um modelo linear hierárquico leva em consideração o fato de que indivíduos pertencentes a um mesmo

grupo possuem maior semelhança, e, portanto apresentam autocorrelação nas suas respostas, em relação a indivíduos de grupos diferentes. Assim, diferente dos modelos de análise de regressão, que possuem quatro suposições básicas: linearidade, aditividade, normalidade, homocedasticidade e independência, os modelos lineares hierárquicos procuram manter as duas primeiras suposições e adaptar as demais a fim de expressar de maneira mais realista a estrutura de correlação presente nos dados.

Os modelos lineares hierárquicos diferenciam-se em três aspectos dos modelos que consideram somente um nível, ou modelos clássicos ou tradicionais de regressão:

- Os coeficientes (intercepto e inclinação) podem variar entre as unidades de nível mais elevado;
- Incluem parâmetros adicionais correspondendo à variância do intercepto e coeficientes de inclinação entre unidades de nível hierárquico superior;
- No caso da presença de coeficientes aleatórios no modelo, variáveis explicativas referentes ao segundo nível de hierarquia são incluídas no modelo auxiliando na explicação da variação entre as unidades deste nível.

Pode-se optar, então, pela estimação de um modelo de regressão diferente para cada grupo, cada qual com seus próprios intercepto e coeficiente de regressão. Partindo do pressuposto que estes grupos foram amostrados, podemos considerar tanto o intercepto quanto o coeficiente de regressão como aleatórios e assim estimar um *modelo de coeficientes aleatórios*. Quando se incorpora a estes modelos variáveis explicativas (covariáveis), o modelo hierárquico passa a ser analisado através da análise de covariância (ANCOVA). Se estas variáveis forem referentes a qualquer um dos níveis de hierarquia, está-se tratando de um *modelo multinível*.

3.2. MODELOS MULTINÍVEIS

3.2.1. DEFINIÇÃO

Os modelos de regressão tradicionais possuem a suposição de que os indivíduos em estudo são independentes entre si e em relação ao desfecho. Já quando tratamos com modelos multiníveis ou hierárquicos, dificilmente há independência entre os sujeitos, uma vez que estes estão agrupados e apresentam semelhanças. Além disso, indivíduos que compartilham de um meio semelhante podem responder de forma similar em relação aos desfechos, ou seja, há a presença de correlação entre as observações, conhecidas também como correlações intra-grupo.

Um modelo de regressão multinível completo assume que há um conjunto de dados hierárquicos com uma única variável dependente que é medida no nível mais baixo e várias variáveis explicativas medidas em todos os níveis. (Hox, 1995).

Se considerarmos um modelo com dois níveis de hierarquia, por exemplo, centros médicos e pacientes, podemos ajustar equações de regressão separadas em cada unidade do nível superior (por exemplo, centros médicos) para predizer a variável dependente Y através das variáveis explicativas X. Este modelo pode ser representado por:

$$Y_{ij} = \beta_{0j} + \beta_{1j} X_{ij} + e_{ij} \qquad (8)$$

onde β_{0j} representa o intercepto da equação para cada unidade do nível 2 (ou seja, para cada centro médico);

β_{1j} é o coeficiente de inclinação para cada um dos indivíduos pertencentes ao nível 1 (por exemplo, para cada paciente).

O índice i refere-se às unidades pertencentes ao nível mais baixo enquanto que o índice j está associado as unidades do segundo nível considerado.

Temos então um modelo em que o comportamento pode se diferenciar de centro para centro e até mesmo dentro de um mesmo centro médico. Assumir que tanto o intercepto e o coeficiente de inclinação da equação possam variar de centro para centro, como descrito anteriormente é assumir que estes coeficientes sejam aleatórios.

Reescrevendo os coeficientes do modelo de regressão tradicional segundo a composição de um modelo de coeficientes aleatórios temos:

$$Y_{ij} = \beta_{0i} + \beta_{1i} X_{ij} + e_{ij} \qquad (9)$$

$$\beta_{0j} = \gamma_{00} + u_{0j} \qquad \beta_{1j} = \gamma_{10} + u_{1j}$$

O termo γ_{00} representa o efeito principal de média, ou seja, o termo γ_{00} é o intercepto médio entre as unidades do nível 2 e u_{0j} mede os desvios de cada contexto de valor médio global. Já γ_{10} é a inclinação média entre as unidades do nível 2. Temos γ_{10} como um estimativa maior entre todos os níveis enquanto que u_{1j} representa os desvios das curvas em relação ao coeficiente de inclinação global da equação.

Valores elevados para o intercepto predizem valores altos para a resposta, ou seja, produzem um grande efeito na variável resposta, sugerindo que o grupo seja seletivo, enquanto que valores baixos sugerem grupos igualitários. Se diferentes valores para o intercepto forem encontrados temos diferentes relações entre os elementos do nível 2.

O termo de erro referente ao nível 2 (u_{ij}) é considerado independente do termo de erro do modelo (e_{ij}) apresentado em (8) sendo que ambos possuem esperança igual zero e variâncias iguais a σ^2_{kk} (com k = i, j) e σ^2, respectivamente. Pela suposição de independência entre esses termos de erros a covariância (σ_{12}) entre eles é assumida

como nula. Não se assume que os coeficientes γ variem, logo eles se aplicam a todas as unidades do nível a que pertencem e assim são tidos como efeitos fixos.

Tendo sido estimados os coeficientes do modelo, parte-se para a etapa de predição da variação dos coeficientes de regressão introduzindo variáveis explicativas (Z_j) no nível mais elevado. Para exemplificar, considere somente uma equação estimada dada por:

$$\beta_{0j} = \gamma_{00} + \gamma_{01} Z_j + u_{0j}$$

$$\text{e} \qquad \beta_{1j} = \gamma_{10} + \gamma_{11} Z_j + u_{1j} \qquad (10)$$

Os termos aleatórios u_{0j} e u_{1j} representam os resíduos do segundo nível, seus valores esperados nulos e suas dispersões dadas pela matriz:

$$E\begin{bmatrix} u_{0j} \\ u_{1j} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \qquad \text{e} \qquad Var\begin{bmatrix} u_{0j} \\ u_{1j} \end{bmatrix} = \begin{bmatrix} \sigma_{00}^2 & \sigma_{01} \\ \sigma_{10} & \sigma_{11}^2 \end{bmatrix} \qquad (11)$$

Coeficientes angulares (γ_{01} e γ_{10}) maiores que zero indicam que o grupo a que esta observação pertence é tido como mais seletivo. Caso ocorra o contrário os grupos são tidos como igualitários. O termo u_{0j} representa o resíduo do nível 2.

A relação entre X e Y depende do valor da variável Z uma vez que esta serve de indicador da relação ou proximidade entre as variáveis dependente e independente. A variável Z também é conhecida como variável moderadora ou termo moderador.

Caso o conjunto de variáveis Z adicionadas a equação do intercepto β_{0j} seja diferente daquele utilizado em β_{1j} deve-se prestar mais atenção na interpretação dos resultados, pois variáveis diferentes foram utilizadas na estimação de cada um destes parâmetros.

Um modelo mais geral é obtido quando substitui-se em (9) as equações descritas em (10). Este modelo possui uma parte fixa e outra aleatória, e é apresentado a seguir:

$$Y_{ij} = \underbrace{\gamma_{00} + \gamma_{10}X_{ij} + \gamma_{01}Z_j + \gamma_{11}Z_jX_{ij}}_{fixa} + \underbrace{u_{1j}X_{ij} + u_{0j} + e_{ij}}_{aleatória} \qquad (12)$$

Conforme pode ser observado na equação (12), o efeito moderador de Z na proximidade entre a variável dependente Y e X é expresso como uma interação de níveis cruzados, cuja interpretação pode ser complexa. Em geral, a interpretação de coeficientes no modelo com interação torna-se mais simples se as variáveis que estão envolvidas na interação são expressas como desvios de sua própria média. Uma vez que o termo de erro u_{1j} está ligado a X_{ij}, o resultado total do erro será diferente para cada valor de X_{ij}, caracterizando assim uma situação de heterocedasticidade. (Hox, 1995).

Em um modelo multinível tradicional, assume-se que a variável resposta seja normalmente distribuída e a notação mais condensada usualmente utilizada para o modelo apresentado em (9) pode ser representada por:

$$Y_{ij} \sim N(X\beta, \Omega) \qquad (13)$$

onde $X\beta$ é a parte fixa do modelo e Ω a matriz de covariâncias dos efeitos aleatórios. No caso da variável resposta não seguir uma distribuição Normal, pode-se ajustar um Modelo Multinível Generalizado. Neste trabalho serão considerados somente modelos cuja variável resposta seja normalmente distribuída.

Através de um modelo multinível pode-se estimar a correlação intra-classe. Neste caso, o coeficiente de correlação intra-classe representa uma estimativa da proporção da variância explicada na população. Segundo Kreft & Leeuw (1998), a correlação intra-classe é uma medida do grau de dependência dos indivíduos. Espera-se que indivíduos que pertençam ao mesmo grupo estejam mais correlacionados, ou seja, possuam características mais semelhantes, que indivíduos de grupos distintos. Exemplos

de alto grau de dependência é encontrado quando os indivíduos são gêmeos ou pertencem a mesma família. Ainda, medidas feitas na mesma pessoa em ocasiões diferentes são muito correlacionadas. Este coeficiente de correlação é calculado segundo a expressão (14) e pode ser interpretado também como a proporção da variância da variável resposta explicada entre as unidades do segundo nível, ou seja, a proporção explicada pela variação entre-grupos.

$$\rho = \frac{\sigma_{u0}^2}{\sigma_{u0}^2 + \sigma^2} \qquad (14)$$

onde σ_{u0}^2 representa a variância do grupo, ou nível 2 e σ^2 representa a variância do erro. A variância do erro representa o efeito de todas as variáveis omitidas no modelo e as medidas de erro, que se supõem serem não correlacionados. O coeficiente de correlação intra-classe pode variar de zero, quando se assume que as unidades estão homogeneamente distribuídas, até um, quando há fortes indícios de que a variação da variável resposta se deve exclusivamente a diferença na variação de um nível de grupo para outro .

A existência de uma correlação intra-unidades (ou intra-classe) resulta da presença de mais que um termo de resíduo no modelo, significando que a estimação tradicional por mínimos quadrados ordinários não é apropriada, uma vez que há a violação da suposição de independência das observações. Neste caso, as estimativas dos coeficientes de regressão obtidas a partir dos métodos tradicionais de regressão levarão à subestimação dos erros padrões dos coeficientes e por conseqüência decisões equivocadas nos testes de significância dos mesmos, aumentando a probabilidade de se cometer o erro tipo I.

Em estudos educacionais, onde se avalia o desempenho de alunos dentro de escolas, o coeficiente de correlação intra-classe é denominado "Efeito-Escola". Este permite aferir sobre a magnitude do efeito-escola, ou seja, se apresentar um valor muito elevado sugere que a variabilidade no desempenho dos alunos deve-se à diferença entre

as escolas, enquanto que na presença de um valor próximo a zero têm-se evidências de que a variabilidade no desempenho é devido à variação de cada aluno em particular.

3.2.2. MÉTODOS DE ESTIMAÇÃO DE UM MODELO MULTINÍVEL

Ao iniciar a análise dos dados é recomendada uma análise em cada nível separadamente. Análise esta, para averiguar a qualidade dos dados, presença de valores estranhos e até identificar possível necessidade de transformações nos dados. Em seguida, recomenda-se uma análise bivariada entre as variáveis contínuas para identificar possíveis relações não lineares entre as observações que pudessem influenciar a modelagem.

No modelo linear hierárquico as equações em cada nível devem obedecer às suposições da estimação por mínimos quadrados, pois a falha dessas suposições em um nível pode acarretar falha nas estimativas nos níveis superiores.

A regressão por Mínimos Quadrados Ordinários (OLS) assume que os erros sejam independentes e com variâncias iguais. Caso ocorra violação desta suposição, as estimativas não serão afetadas, mas seus erros padrões sim. A modelagem através de técnicas multiníveis não apresenta estudos muito profundos quanto às conseqüências da violação das suposições. Outras suposições feitas são que os coeficientes aleatórios ou os resíduos, de um nível, são independentes entre os grupos. Além disso, os coeficientes de um nível superior são independentes dos resíduos do nível inferior a ele. Quanto aos resíduos, temos que no primeiro nível eles possuem distribuição normal com média zero e variância constante, e os resíduos do segundo nível do modelo apresentam-se normalmente distribuídos com média zero e matriz de covariância constante.

Os parâmetros que podem ser estimados em uma análise hierárquica ou são parâmetros de efeitos fixos, ou coeficientes aleatórios do nível 1 ou ainda componentes de variância e covariância (Byrk & Raundebush, 1992).

O número de parâmetros a ser estimado aumenta à medida que mais variáveis explicativas são incorporadas nos níveis do estudo, e é por isso que na modelagem multinível, freqüentemente ocorrem problemas de convergência, principalmente quando um grande número de parâmetros de covariâncias estão sendo estimados e estes possuem valores muito pequenos, próximos de zero. Uma solução para este caso é a simplificação do modelo, retirando tais componentes.

O princípio mais comum usado pelos pesquisadores para a estimação dos parâmetros é o de máxima verossimilhança (ML) que, às vezes, também é conhecido como máxima verossimilhança com informação completa (FIML). As estimativas obtidas através deste método são sempre não negativas para os componentes de variância, no entanto, não considera a perda de graus de liberdade resultante da estimação dos efeitos fixos do modelo, produzindo estimativas viesadas.

Um processo variante do Método de Máxima Verossimilhança é a Máxima Verossimilhança Restrita (REML). Neste procedimento, cada observação é dividida em duas partes independentes, uma referente aos efeitos fixos e outra aos aleatórios, de modo que a função densidade de probabilidade das observações seja dada pela soma das funções densidades de cada parte. Neste método somente os componentes de variância são incluídos na função de verossimilhança. Pode-se considerar REML como sendo um método mais realístico, ao menos na teoria, no entanto, na prática a diferença entre os dois métodos não se apresenta muito grande. (Kreft, Leeuw & Kim, 1990).

Se os parâmetros da parte aleatória do modelo apresentado em (12) são conhecidos, então os coeficientes de regressão poderiam ser estimados pelo método dos mínimos quadrados generalizados(Snijdes & Bosker, 1999). Diferentemente dos estimadores de mínimos quadrados utilizados nos modelos de regressão tradicionais, o método de estimação de mínimos quadrados generalizados (GLS) leva em conta a informação de que a variabilidade das observações é diferente em cada grupo, ou seja, permite a modelagem na presença de heterocedasticidade entre os grupos. Já se todos os parâmetros γ da equação (12) fossem conhecidos, a matriz de covariâncias dos efeitos aleatórios poderia ser utilizada para estimar os parâmetros da parte aleatória. Estes dois

processos de estimação parcial poderiam ser alternados: usar valores provisórios para os parâmetros da parte aleatória para estimar os coeficientes de regressão e posteriormente estimar os parâmetros da parte aleatória do modelo novamente, gerando assim um processo iterativo que só finaliza quando convergir. Este processo é utilizado pelo algoritmo de estimação de Mínimos Quadrados Generalizados Iterativos (IGLS) que é um dos algoritmos que também calcula as estimativas de Máxima Verossimilhança (ML). Outros algoritmos que também calculam as estimativas de ML são o Escore de Fisher e o Algoritmo EM ("Expectation-Maximization".)

3.2.3. HIPÓTESES TESTADAS

Considere novamente o modelo apresentado em (12). Como na regressão linear simples, uma estatística t-Student pode ser utilizada para testar a significância dos interceptos e inclinações das equações do segundo nível.

$$Y_{ij} = \underbrace{\gamma_{00} + \gamma_{10}X_{ij} + \gamma_{01}Z_j + \gamma_{11}Z_jX_{ij}}_{fixa} + \underbrace{u_{1j}X_{ij} + u_{0j} + e_{ij}}_{aleatória} \qquad (12)$$

Para os efeitos fixos do modelo (12), testa-se se o valor do parâmetro γ é igual ou diferente de zero. Ou seja, está-se interessado em testar se há efeito médio significativo do coeficiente, ou ainda, se há efeito significativo das variáveis preditoras no segundo nível de análise. A hipótese de nulidade é expressa da seguinte forma:

$$H_0: \gamma = 0$$

A estatística de teste é dada por:

$$t = \frac{\hat{\gamma}}{\sqrt{\hat{Var}(\hat{\gamma})}} \qquad (15)$$

que segue assintoticamente uma distribuição normal, muito embora melhor aproximada pela distribuição t de Student.

No caso de testes com múltiplos parâmetros, testa-se a hipótese de que o vetor de efeitos fixos seja igual a zero contra a alternativa que são diferentes de zero, utilizando-se a estatística F.

Quando o parâmetro de interesse for considerado um coeficiente aleatório do primeiro nível testa-se a hipótese de que o parâmetro é nulo, contra a alternativa de que o parâmetro possui efeito significativo, ou não nulo. A hipótese e a estatística de teste são apresentadas a seguir:

$$H_0: \beta_{qj}=0 \qquad e \qquad z=\frac{\beta^*_{qj}}{\sqrt{V^*_{qqj}}} \tag{16}$$

onde β^*_{qj} é o estimador empírico de Bayes;

V^*_{qqj} é o q-ésimo elemento da diagonal das dispersões à posteriori dos coeficientes β^*_{qj};

z possui distribuição aproximadamente normal quando a hipótese de nulidade é verdadeira.

O estimador empírico de Bayes β^*_{qj} é uma combinação de dois estimadores de β_{qj}, o primeiro baseado no modelo do nível 1 (Equação 17) e o outro no nível 2 (Equação 18). No primeiro caso, um estimador não viesado seria $\overline{Y}_{.j}$ cuja variância seria dada por V_j. Já no segundo, teríamos $\hat{\gamma}_{00}=\frac{\sum \Delta_j^{-1}\overline{Y}_{.j}}{\sum \Delta_j^{-1}}$, onde Δ_j^{-1} é a matriz de dispersão dos dados.

$$\overline{Y}_{.j} = \beta_{qj} + \varepsilon_{ij}, \quad \varepsilon_{.j} \sim N(0, V_j) \tag{17}$$

$$\beta_{qj} = \gamma_{q0} + u_{qj}, \quad u_{0j} \sim N(0, \sigma^2_{q0}) \tag{18}$$

O estimador de Bayes é a combinação ponderada destes dois estimadores, dado por:

$$\beta^*_{qj} = \lambda_j \overline{Y}_{.j} + (1 - \lambda_j) \hat{\sigma}^2_{q0} \tag{19}$$

Um peso λ_j grande, indica grande confiança em $\overline{Y}_{.j}$ como estimador de β_{qj}, já quando se aplica um valor pequeno para λ_j atribui-se maior confiança a $\hat{\gamma}_{00} = \frac{\sum \Delta_j^{-1} \overline{Y}_{.j}}{\sum \Delta_j^{-1}}$ como estimador de β_{qj}.

O valor de λ_j é calculado através da equação (20), e mede a razão entre o escore verdadeiro (ou a variância do parâmetro) e o escore observado (ou variância da media amostral $\overline{Y}_{.j}$).

$$\lambda_j = \frac{Var(\beta_{qj})}{Var(\overline{Y}_{.j})} = \frac{\sigma^2_{q0}}{\sigma^2_{q0} + V_j} \tag{20}$$

Maiores detalhes podem ser encontrados em Bryk & Raudenbush (1992).

A terceira hipótese possível de ser testada é a dos componentes de covariância do modelo, pois se está interessado em avaliar se a variação aleatória existe, ou seja, se há variação entre as unidades de níveis mais elevados. A hipótese de nulidade é dada pela expressão a seguir, onde σ^2_{qq} é a variância de β_{qq}.

$$H_0: \sigma^2_{qq} = 0$$

Quando há a rejeição da hipótese nula, o investigador possui evidências para crer que há variação aleatória no coeficiente β_q.

A estatística de teste segue uma distribuição aproximadamente normal e é calculada conforme a seguinte expressão:

$$z = \frac{\hat{\sigma}^2_{qq}}{\sqrt{Var(\sigma^2_{qq})}}$$

sendo que

$$\mathrm{Var}(\sigma^2_{qq}) = \sum_j \frac{\left(\hat{\beta}_{qj} - \hat{\gamma}_{qs}\, W_{sj}\right)^2}{\hat{V}_{qqj}} \qquad (21)$$

$\hat{V}_{qqj}$ é o q-ésimo elemento da diagonal da matriz $\hat{V}_{qqj} = \hat{\sigma}^2 (X_j' X_j)^{-1}$

No caso de múltiplos parâmetros a hipótese testada é sobre a matriz de variâncias e covariâncias $Var\begin{bmatrix} u_{0j} \\ u_{1j} \end{bmatrix}$, apresentada em (11).

Uma forma de avaliar a estimação dos parâmetros é através da estatística de *deviance*, que quanto menor for seu valor, melhor foi o ajuste obtido. O teste da *deviance* é também conhecido como teste da razão da verossimilhança e utilizado em testes de muitos parâmetros e para testes da parte aleatória do modelo.

Os parâmetros são estimados segundo o método de Máxima Verossimilhança e a partir do valor de menos duas vezes o logaritmo da verossimilhança calculada obtém-se o valor da *deviance* do modelo ajustado.

Pode-se comparar os valores da *deviance* de diferentes ajustes feitos em um mesmo banco de dados, porem não é possível comparar *deviances* de conjuntos diferentes de dados. Esta estatística pode ser usada para comparar o ajuste de dois

modelos, se um estiver hierarquicamente encaixado no outro, e representa uma medida de ajuste para componentes de covariância do modelo. É possível ainda calcular o valor da *deviance* e a partir das estimativas obtidas através do método de Máxima Verossimilhança Restrita, no entanto só é possível comparar seu valor entre modelos que tenham os mesmos efeitos fixos mas se diferenciam quanto a sua parte aleatória (Snijders&Bosker,1999). A diferença na estatística de *deviance* para dois modelos encaixados terá distribuição Qui-Quadrado assintótica com graus de liberdade igual a diferença no número de seus respectivos parâmetros. A *deviance* é igual a menos duas vezes o valor da verossimilhança e é a mesma estatística de ajuste usada em modelos de equações estruturais sob o nome de "goodness of fit". Nos modelos lineares hierárquicos, especialmente em casos de estimação de modelos complexos, a estatística e *deviance* é usada no lugar da estatística R^2, mas no entanto sua interpretação não é direta como no caso do coeficiente de determinação.

O procedimento de estimação padrão dos modelos hierárquicos é a estimação por REML, assim, a *deviance* é primeiramente usada para decidir se:

(a) Um coeficiente no nível 1 deve ser considerado como aleatório ou fixo e se existe melhoria no ajuste do modelo em cada uma dessas situações;

(b) Houve a melhoria do ajuste quando há a inclusão de preditores nas equações referentes aos segundo nível de hierarquia.

Uma vez que o poder trabalha com a probabilidade de encontrar um efeito significativo quando tal efeito está realmente presente nos dados, é extremamente importante observar a estimativa dos erros padrões dos parâmetros estimados, pois a partir delas é que se obterá o poder do modelo de regressão. Erros padrões elevados , ou seja, erro padrões maiores que um meio da estimativa, sugerem um baixo poder, enquanto que erros padrões pequenos indicam um poder maior. Dessa forma, segundo apresentado por Kreft e Leeuw (1998), o poder está baseado na variância do erro dos efeitos fixos. A estimação por máxima verossimilhança, empregada na estimação de modelos multiníveis, fornece, em média, um maior poder e uma menor probabilidade de se cometer o erro Tipo I.

Para se obter um poder suficiente, é preciso, em geral, um grande número de observações, a menos que os efeitos sejam fortes e facilmente detectados. Quando se tem um pequeno número de grupos, os componentes aleatórios são subestimados (quando se utiliza o Mínimos Quadrados Generalizados Iterativos - IGLS) ou ainda apresentam grandes erros padrões (no caso de trabalhar com Mínimos Quadrados Generalizados Iterativos Restritivos - IGLS). Um poder suficiente pode ser encontrado quando os grupos não são muito pequenos, e o número de grupos é maior que 20.(Kreft & Leeuw, 1998). Isto depende de caso a caso, conforme a magnitude do coeficiente de correlação intra-classe e, portanto do tamanho do efeito estimado.

3.2.4. INTERPRETANDO AS INTERAÇÕES

É muito importante interpretar corretamente o significado das interações em um modelo, o qual apresenta ou não estrutura de hierarquia. Segundo Hox (1995), tendo sido detectada uma interação significativa, aconselha-se incluir ambos efeitos envolvidos na interação no modelo, mesmo que individualmente eles não sejam significativos.

As interações entre níveis são definidas como interações de variáveis medidas em diferentes níveis de hierarquia na estrutura dos dados. Por exemplo, em um estudo educacional avalia-se o sexo do aluno (variável do nível 1) e também o efeito do professor (variável do nível 2). Se esta se apresentar significativa, deve-se incluir os efeitos de sexo e de professor no modelo, mesmo que estes não tenham sido significativos individualmente. Como conclusão da análise, temos que existe um forte efeito de professor atuando em alguns tipos de estudantes, seja de forma positiva ou negativa.

3.2.5. DIAGNÓSTICO DO MODELO: ANÁLISE DE RESÍDUOS

Devida à complexa estrutura que compõe um modelo multinível, torna-se difícil a visualização de como os casos influenciam as estimativas dos parâmetros ou como uma unidade afeta os parâmetros da outra unidade. Logo, uma análise dos resíduos pode conduzir o investigador a identificar presença de *outliers*, quando uma unidade discorda totalmente das demais unidades dentro de um mesmo grupo (Hilden-Minton,1995).

Na regressão múltipla tradicional, os resíduos podem ser estimados simplesmente subtraindo os valores preditos pelos observados, para cada um dos indivíduos. Já em modelos multiníveis que possuem resíduos em vários níveis este procedimento de estimação dos resíduos é um pouco mais complexo.

Supondo que Y_j é o valor observado para o i-ésimo sujeito da j-ésima unidade e que seu valor foi predito por uma regressão linear, o resíduo bruto para o sujeito é $r_{ij} = y_{ij-} \hat{y}_{ij}$. O resíduo bruto para a j-ésima unidade é dado pela média dos resíduos dos indivíduos de uma mesma unidade (denotado por r_{+j}). O resíduo para as unidades do segundo nível é dado por r_{+j} multiplicado pelo fator apresentado a seguir:

$$\hat{u}_{0j} = \frac{\sigma^2_{u0}}{\sigma^2_{u0} + \sigma^2_{e0}/n_j} r_{+j} \qquad (22)$$

onde n_j é o tamanho da j-ésima unidade.

O fator que multiplica r_{+j} em (22) é menor que 1, logo o resíduo estimado dessa maneira é menor que o resíduo bruto e este fator pode também ser chamado de fator de redução, ou encolhimento (Rasbash, Browne, et. al., 2000).

Assim o estimador do resíduo para o primeiro nível pode ser descrito por:

$$\hat{e}_{0ij} = r_{ij} - \hat{u}_{0j} \qquad (23)$$

No caso da presença de unidades ou grupos atípicos é preciso tomar a decisão de excluí-los ou não da análise. Partir para uma análise de diagnóstico do modelo e assim avaliar se suposições do modelo foram atendidas em cada nível é extremamente importante nesse caso.

Pode-se prosseguir com a análise de diagnóstico do modelo simplesmente checando o mesmo ou então realizando uma análise de sensibilidade. Quanto à checagem do modelo, identifica-se possíveis violações do modelo e suposições básicas, tais como normalidade, heterocedasticidade. Muitos dos diagnósticos estão interligados de tal modo que os testes podem estar confundidos, ou ainda correlacionados e com pouco poder. Testes formais reduzem um problema complexo em uma única estatística de teste. Um procedimento gráfico pode ser proposto a fim de auxiliar o investigador nesta fase. A análise de sensibilidade concentra-se em quanto influencia a caracterização dos dados ou nas especificações das suposições feitas sobre as inferências baseadas no modelo. Na modelagem multinível, deveria ser avaliada cada fonte de variação do modelo separadamente, definindo um resíduo distinto para cada fonte de erro do modelo. Procura-se por resíduos que não apresentem distúrbios nas inferências.

Em modelos hierárquicos pode-se pensar o modelo como sendo formado por modelos menores e assim, os resíduos do modelo global podem estar confundindo os efeitos dos modelos menores. Recomenda-se avaliar os resíduos dos modelos menores e posteriormente os resíduos globais.

Os princípios básicos para um diagnóstico de um modelo multinível podem ser descritos como:

Decomposição: Entender o modelo como uma rede interligada de sub-modelos e atentar para a satisfação da suposição em cada um deles.

Ignorância a Priori: Alguns componentes de um modelo multinível podem atuar como um componente Bayesiano a priori para outros componentes. Por exemplo, em um modelo com dois níveis, o modelo entre as unidades atua como uma priori empírica para os modelos dentro de cada uma das unidades. Assim, o princípio da ignorância a priori pode conduzir ao uso de resíduos e diagnósticos equivocados. O uso de uma estrutura a priori adequada evita problemas que poderão afetar na análise de resíduos.

Recomposição: Depois de examinar os componentes separadamente deve-se recombinar os componentes para melhor estimar os resíduos e compará-los.

3.2.6. ETAPAS DA ANÁLISE MULTINÍVEL

É durante a especificação do modelo, tida como umas das etapas mais difíceis da análise, que se tem necessidade de levar em consideração tanto os aspectos estatísticos como as características do experimento em avaliação, tais como a teoria envolvida no estudo, detalhes da formulação do problema, além do conhecimento do pesquisador sobre esse assunto (Snijder & Bosker, 1999) .

Outro ponto importante a ser destacado é distinguir os efeitos que são incluídos por interesse inicial do pesquisador e aqueles necessários para se obter um adequado ajuste do modelo, e ainda definir quais serão analisados como fixos ou como aleatórios. Na busca de uma adequação no ajuste é preciso tomar cuidado para não incluir termos em excesso, pois isso pode dificultar na posterior interpretação dos coeficientes do modelo.

A seguir são sugeridas algumas etapas a serem seguidas na adequação de um modelo. Este é um procedimento indicado por Hox (1995). Inicia-se com um modelo, mais simples possível, com um único intercepto, e vários tipos de parâmetros vão sendo incluídos passo a passo. Em cada uma das etapas são realizadas avaliações do ajuste do modelo.

Etapa 1: *Análise de um modelo sem variáveis explicativas*

Primeiramente parte-se para a análise de um modelo que não contempla variáveis explicativas. Assim, é possível obter uma estimativa do coeficiente de correlação intraclasse (14). Além disso, este modelo fornece o valor da *deviance*, que é uma medida do grau da falta de ajuste do modelo que servirá como um valor padrão para a comparação com os valores obtidos posteriormente com a inclusão de mais variáveis no modelo.

Este modelo apresenta-se da seguinte forma:

$$Y_{ij} = \gamma_{00} + u_{0j} + e_{ij} \qquad (24)$$

Etapa 2: *Análise de um modelo com variáveis fixas no nível mais baixo*

Nesta fase da regressão inclui-se mais variáveis explicativas do nível 1 no modelo. Os coeficientes destas variáveis são considerados fixos e permitem avaliar a contribuição de cada variável explicativa individualmente.

As variáveis explicativas acrescentadas no modelo (denotadas por X na equação 25) auxiliam na explicação da variação da resposta final, e devem ser relevantes tanto sob o ponto de vista estatístico quanto do pesquisador e realmente possam trazer uma melhoria no ajuste.

$$Y_{ij} = \gamma_{00} + \gamma_{p0} X_{pij} + u_{0j} + e_{ij} \qquad (25)$$

Neste ponto da análise, já é possível identificar quais das variáveis apresentam maior contribuição na explicação da variação do modelo. Novamente utiliza-se a estatística de *deviance*, para avaliar a melhoria no ajuste com a inclusão de cada uma das variáveis. Compara-se a *deviance* com a obtida na Etapa 1, a diferença entre esses dois valores seguem aproximadamente uma distribuição Qui-quadrado com graus de

liberdade iguais a diferença no numero de parâmetros considerados em ambos os modelos.

O valor para a estatística de *deviance* é sempre um valor positivo, e à medida que entram variáveis explanatórias (ou covariáveis) no componente sistemático, decresce até se tornar zero para o modelo saturado. (Demétrio, 2001).

Etapa 3: *Avaliação dos coeficientes de regressão das variáveis explicativas*

Tendo sido escolhidas e adicionadas ao modelo as variáveis de nível mais baixo, define-se quais dos coeficientes de regressão do modelo adotados considerados inicialmente como fixos serão considerados como aleatórios e parte-se, então, para a verificação de significância dos mesmos.

O coeficiente de regressão de uma determinada variável será considerado fixo se o comportamento desta variável não se diferenciar entre os grupos, caso contrário, admite-se que o coeficiente assuma um comportamento aleatório. O modelo (24) passa a ser escrito como:

$$Y_{ij} = \gamma_{00} + \gamma_{p0} X_{pij} + u_{pj} X_{pij} + u_{0j} + e_{ij} \qquad (26)$$

É preciso testar cada um dos coeficientes de regressão separadamente. Variáveis omitidas em etapas anteriores podem ser analisadas novamente nesta. Isso porque é possível que variáveis explicativas não possuam significância na regressão mas possuam nos componentes de variância.

Tendo sido determinados quais coeficientes de regressão são significativos, pode-se testar os desvios de ajuste deste novo modelo confrontando com o modelo apresentado na segunda etapa utilizando novamente o valor da estatística de *deviance*, calculada na etapa 2.

Etapa 4: *Acrescentando as variáveis do nível mais elevado*

Com a adição de variáveis explicativas ligadas ao nível mais elevado, podemos avaliar se estas variáveis explicam a variação entre grupos junto a variável dependente. O modelo suposto nesta etapa tem a seguinte configuração:

$$Y_{ij} = \gamma_{00} + \gamma_{p0} X_{pij} + \gamma_{0q} Z_{qj} + u_{pj} X_{pij} + u_{0j} + e_{ij} \qquad (27)$$

Novamente deve-se buscar a inclusão de variáveis cujo efeito sobre a resposta seja plausível segundo a opinião do pesquisador. Realizam-se testes de significância para o modelo adotado e pode-se optar pela inclusão de novos termos ou ainda interações entre níveis, cuidando para não acrescentar termos em demasia.

Etapa 5: *Acrescentando as interações*

Por fim a quinta e última etapa adiciona as interações entre variáveis explicativas do nível de um grupo com variáveis explicativas do nível individual.

Estas interações devem ser escolhidas dando prioridade para aquelas entre variáveis explicativas mais importantes para cada um dos níveis avaliados, e devem ser definidas novamente em conjunto com o pesquisador.

O modelo final, então se apresenta da seguinte forma:

$$Y_{ij} = \gamma_{00} + \gamma_{p0} X_{pij} + \gamma_{0q} Z_{qj} + \gamma_{pq} Z_{qj} X_{pij} + u_{pj} X_{pij} + u_{0j} + e_{ij} \qquad (28)$$

Quando se trata de ajustes de modelos, não são raras às vezes em que dois pesquisadores estimem modelos com diferentes efeitos especificados, ambos demonstrando ajuste adequado, isso porque há uma indeterminação característica de ajustes de modelos de bases empíricas, ocasionada pela interpretação diferente dada ao mesmo problema. Nestes casos é aconselhável adiar a escolha final do modelo após a

realização de pesquisas futuras possibilitando então uma comparação mais razoável entre os dois modelos quanto ao seu potencial.

Em nenhum estágio da modelagem deve-se esquecer das suposições necessárias pela técnica adotada, tanto no caso quando a especificação do modelo foi feita de forma indutiva, através do auxilio do pesquisador, ou quando o ajuste de modelos sem hipóteses a priori é proposto.

3.2.7. MODELOS COM TRÊS NÍVEIS

A estrutura hierárquica que envolve os dados pode envolver quantos níveis forem precisos para caracterizar melhor o conjunto de dados que se está trabalhando. Se considerarmos um estudo epidemiológico realizado em vários centros de pesquisas, coordenados por diferentes equipes de médicos que atendem os pacientes participantes da pesquisa temos três níveis de hierarquia envolvidos, um para o local do estudo, outro referente ao grupo de pesquisadores e no nível mais baixo o paciente tratado.

Outro exemplo de um problema multinível com 3 níveis é na pesquisa educacional, em que se pretende avaliar o aprendizado dos alunos, agrupados em salas de aula que pertençam tanto a escolas públicas ou privadas. Pode-se estender aplicações destes modelos em pesquisas de opinião que envolve a entrevista de indivíduos agrupados em domicílios dentro de determinadas regiões geográficas, ou ainda trabalhadores dentro de firmas que por sua vez estão dentro de indústrias.

Os procedimentos de análise para modelos com três níveis são semelhantes aos apresentados no caso de modelos com dois níveis e se estendem a problemas que envolvam estrutura com níveis superiores. No entanto, cabe salientar que à medida que mais níveis são incluídos na análise mais complicadores serão adicionados ao modelo, e é preciso ter bom senso na escolha da melhor modelagem, buscando sempre a simplicidade e a validade do modelo.

O modelo mais simples é o modelo incondicional completo, onde nenhuma variável preditora é incluída nos níveis. Considerando somente o nível inferior temos a seguinte especificação do modelo:

$$Y_{ijk} = \pi_{0jk} + e_{ijk} \tag{29}$$

onde

Y_{ijk} é a resposta do indivíduo i pertencente a unidade j (nível 2) e a unidade k (nível 3)

π_{0jk} é a resposta média da unidade j dentro do grupo k.

e_{ijk} é o efeito aleatório do indivíduo (nível 1) e representa o desvio do indivíduo em relação a sua média. Supõe-se que este efeito é normalmente distribuído com média 0 e variância σ^2.

Para a adição da informação do segundo nível hierárquico, considera-se o termo π_{0jk} como variável de resposta que varia aleatoriamente em torno da média do segundo nível.

$$\pi_{0jk} = \beta_{00k} + r_{0jk} \tag{30}$$

onde

β_{00k} representa a resposta média para as unidades do segundo nível;

r_{0jk} indica o termo aleatório para as unidades do nível 2 que simboliza os desvios das unidades deste nível em torno de sua média. Para estes efeitos assume-se também que possuam distribuição normal com média 0 e variância $\sigma^2{}_{\pi}$. Assume-se também que a variabilidade dentro de cada unidade do nível 2 a variabilidade seja a mesma.

Por fim caracterizando um modelo de três níveis, acrescenta-se a expressão que representa a variabilidade entre as unidades do nível mais elevado. Novamente considera-se que as unidades deste nível variem aleatoriamente em torno de sua média:

$$\beta_{00k} = \gamma_{000} + u_{00k} \tag{31}$$

onde

γ_{000} é a média geral deste nível;

u_{00k} representa a dispersão das unidades do terceiro nível em torno da sua média e este termo de erro é normalmente distribuído com média 0 e variância $\sigma^2{}_\beta$;

Temos então que a variabilidade total da variável resposta Y_{ijk} é particionada em três componentes, um para cada nível. Estes componentes permitem estimar a proporção da variação explicada em cada um dos níveis conforme as expressões da Tabela 1. Sua interpretação é similar ao coeficiente de correlação intra-classe nos modelos de dois níveis.

Tabela 1. Expressões para a estimação da proporção de variância explicada em cada um dos níveis.

Nível	Significado	Expressão
Nível 1	Proporção da variação dentro das unidades do nível 1.	$\frac{\sigma^2}{\sigma^2 + \sigma_\pi^2 + \sigma_\beta^2}$
Nível 2	Proporção da variação entre unidades do nível 2 dentro das unidades do nível 3.	$\frac{\sigma_\pi^2}{\sigma^2 + \sigma_\pi^2 + \sigma_\beta^2}$
Nível 3	Proporção da variância entre unidades do terceiro nível.	$\frac{\sigma_\beta^2}{\sigma^2 + \sigma_\pi^2 + \sigma_\beta^2}$

A partir deste modelo base pode-se adicionar variáveis explicativas em cada um dos níveis de hierarquia. Os coeficientes de regressão associados a estas novas variáveis incluídas no modelo podem variar aleatoriamente nos níveis mais elevados, daí a possibilidade de formular um modelo estrutural em cada nível.

A forma e a notação do modelo passam a ficar mais complexas à medida que ocorrem novas inclusões dos termos. Considere o modelo para o nível 1 apresentado em (29)

$$Y_{ijk} = \pi_{0jk} + e_{ijk}$$

Acrescentando variáveis preditoras deste nível, ou seja, variáveis explicativas do sujeito sua expressão passa a ser dada por:

$$Y_{ijk} = \pi_{0jk} + \pi_{1jk}\alpha_{1ijk} + \pi_{2jk}\alpha_{2ijk} + \pi_{3jk}\alpha_{3ijk} + ... + e_{ijk} \qquad (32)$$

onde

Y_{ijk} é a resposta do indivíduo i pertencente a unidade j (nível 2) e a unidade k (nível 3).

π_{0jk} é o intercepto da unidade j dentro do grupo k.

α_{pijk} simboliza as P variáveis preditoras do nível 1(p=1...P).

π_{pjk} coeficientes que indicam a relação de força de associação entre as características do individuo e as unidades de níveis superiores.

e_{ijk} é o efeito aleatório do indivíduo (nível 1)

Quando se acrescentam variáveis preditoras do segundo nível de hierarquia, temos que cada coeficiente π_{pjk} pode ser visto como fixo ou aleatório. A expressão geral para o modelo que considera a variação entre elementos do primeiro nível dentro de cada uma das unidades do nível 2 é dada da seguinte forma:

$$\pi_{pjk} = \beta_{p0k} + \sum_{q=1}^{Q_p} \beta_{pqk} X_{qjk} + r_{pjk} \qquad (33)$$

onde

β_{p0k} é o intercepto para a unidade k no modelo de efeito da unidade π_{pjk} do primeiro nível.

X_{qjk} é a característica usada como preditor para os elementos do segundo nível ($q = 1...Q_p$).

β_{pqk} é o coeficiente que representa a força e a direção do efeito da associação entre as características X_{qjk} e π_{pjk}.

r_{pjk} é o efeito aleatório deste nível.

Temos neste nível P+1 equações, uma para cada coeficiente do nível 1. Os efeitos aleatórios destas equações são assumidos como sendo correlacionados e que o conjunto de efeitos aleatórios r_{pjk} possuem distribuição normal multivariada.

Por fim, pode-se ainda adicionar ao modelo variáveis preditoras do terceiro nível de hierarquia. Cada coeficiente β_{pqk} pode ser predito por algumas características do nível 3 através da seguinte expressão.

$$\beta_{pqk} = \gamma_{pq0} + \sum_{s=1}^{S_{pq}} \gamma_{pqs} W_{sk} + u_{pqk} \qquad (34)$$

onde

γ_{pq0} é o intercepto do nível 3 para o modelo de β_{pqk}.

W_{sk} é a variável explicativa para o terceiro nível ($s = 1...S_{pq}$).

γ_{pqs} é o coeficiente e representa a associação entre a variável explicativa W_{sk} e β_{pqk}.

u_{pqk} representa o efeito aleatório deste nível.

Abre-se um grande número de opções de formulações de modelos, entre aquele que não considera nenhuma variável explicativa até o modelo mais completo, que possui variáveis explicativas em todos os níveis de hierarquia. Numa análise é possível escolher se os preditores serão introduzidos em cada um dos níveis; especificar se os coeficientes de inclinação e intercepto são fixos ou aleatórios em cada nível; e por fim, especificar modelos alternativos para os componentes de variância e covariância.

Os testes de hipótese para os coeficientes deste nível são semelhantes ao aplicado no caso de um modelo multinível com dois níveis. Os testes univariados para

os coeficientes fixos ajudam na decisão da necessidade de inclusão dos mesmos no modelo e a escolha do modelo pode ser feita novamente pelo valor da deviance do modelo adotado comparada ao modelo mais simples.

4. SOFTWARES

Os aplicativos tradicionais de análise estatística abordam a modelagem através de modelos lineares generalizados, que consideram somente um nível de hierarquia e uma única variável aleatória. No entanto, quando tratamos de casos onde se faz necessário utilizar uma modelagem multinível, estes aplicativos de uma maneira geral não se mostram muito adequados.

Alguns dos pacotes existentes permitem trabalhar com dados multiníveis utilizando a metodologia de modelos lineares mistos, é claro que nem todos os casos podem ser abordados desta maneira e por isso se faz necessária a utilização de rotinas adequadas de análise.

A partir dos anos 80, com o avanço da computação, foram surgindo vários aplicativos que fornecem uma opção de se trabalhar com modelos multiníveis, dentre eles podemos citar, BMDP - 5V, GENMOD, HLM, MLN3 e VARCL. Posteriormente, surgiram outros, tais como o SAS, o MLwiN e o GENSAT.

Existem ainda dois outros pacotes, o EGRET e SABRE, que possuem procedimentos que ajustam os modelos logit com 2 níveis. Ainda há a opção do BUGS, um aplicativo bayesiano que utilizando o Gibbs Sampling é capaz de ajustar os mesmos modelos multiníveis que o MLwiN.

Espera-se, que em breve, módulos de modelagem multinível estejam disponíveis cada vez em mais aplicativos computacionais, e que possuam uma interface que facilite o entendimento da complexa estrutura que envolve esta classe de modelos.

A seguir apresenta-se um breve comentário de alguns dos aplicativos computacionais acima citados.

4.1. BMDP-5V

Jennrich e Schluchter (1986) desenvolveram o aplicativo BMDP5-V, também conhecido como BMDP, que é voltado a dados de medidas repetidas, incluindo situações de desbalanceamento e presença de dados "missing". Fornece várias opções de técnicas e algoritmos, além de permitir modelar a dispersão dos resíduos do segundo nível do modelo. Entretanto, seus comandos podem se tornar um pouco complicados. Possui alguma semelhança com o procedimento PROC MIXED do SAS.

Os modelos que podem ser ajustados pertencem à classe de modelos lineares multivariados generalizados, onde um conjunto de valores de coeficientes de regressão descrevem a estrutura dos valores observados e um conjunto de parâmetros de covariância fornecem uma parametrização mais geral da covariância dentro de sujeitos. Estas estruturas de covariâncias podem ser definidas e ajustadas pelo pesquisador.

O ajuste do modelo para medidas repetidas precisa ser feito considerando um modelo multivariado, em que as medidas de diferentes ocasiões são consideradas variáveis dependentes separadas. Quando se considera o contexto multinível, os indivíduos passam a fazer parte do segundo nível e as medidas do primeiro, onde as observações aparecem agrupadas.

O uso deste programa em dados que apresentam missing é um pouco restritivo, pois há casos em que os valores dos preditores dentro dos grupos são idênticos e a matriz de variáveis preditoras deve ser idêntica para todos os indivíduos.

As rotinas do BMDP-5V permitem ao usuário escolher três algoritmos diferentes para calcular as estimativas de máxima verossimilhança do modelo. Os algoritmos disponíveis são o de Newton-Raphson, Escores de Fisher e o algoritmo EM. Todos eles fornecem o mesmo resultado, o que os diferencia é o número de iterações para a estimação.

Para a análise dos dados, as instruções são semelhantes às utilizadas no SPSS ou no SAS, dentre as principais estão DESIGN, MODEL e STRUCTURE. A primeira especifica a estrutura dos dados, identificando variáveis que classificam o sujeito, e

varáveis de medidas. A instrução MODEL especifica o modelo enquanto que STRUCTURE trata da definição da estrutura de covariância dentro de cada sujeito. Dentre as estruturas de covariâncias disponíveis estão a auto-regressiva, componente simétrico ou ainda estruturas auto regressivas gerais ou "banded". Ainda há a opção do próprio pesquisador definir outras estruturas utilizando a linguagem de programação FORTRAN77.

Uma vez o comando PRINT acionado, os resultados das instruções do programa, especificação do modelo e os parâmetros são exibidos. Além disso, o critério de informação de Akaike também é fornecido a fim de avaliar a qualidade do ajuste segundo diversas estruturas de covariâncias; os erros padrões assintóticos para as estimativas de máxima verossimilhança também são listados e o teste de Wald para a significância dos termos de regressão do modelo também é apresentado. É possível, ainda, listar os valores preditos, resíduos e resíduos padronizados e avaliar a distância de Malahanobis com a finalidade de detectar possível presença de *outlier* (Kreft et al.,1994).

4.2. GENMOD

O GENMOD foi desenvolvido por Wong e Mason (Mason, Wong & Entwisle, 1984) e surgiu em pesquisas de estudos demográficos. Possibilita a realização de análises comparativas e contextuais, esta última exclusiva deste programa.

A linguagem de programação base é a FORTRAN77. O algoritmo adotado é o algoritmo EM. O modelo básico ajustado pelo GENMOD é um modelo com dois níveis, sendo que uma versão especial deste programa permite que diferentes contextos tenham variâncias diferentes para o primeiro nível, nos demais a variância é assumida como sendo a mesma.

A estruturação do banco de dados é similar a utilizada em pacotes como o SAS e o SPSS, contendo um cabeçalho que especifica o arquivo de respostas e opções de imputação de dados. Os resultados contêm as informações usuais básicas sobre

covergência, iterações calculadas, estimativas de Máxima Verossimilhança Restrita (REML), estimativas de parâmetros além de outras informações necessárias.

4.3. HLM

Segundo Kreft et. al. (1994), o pacote HLM é muito utilizado em estatísticas educacionais, e tornou-se o software oficial de análises multiníveis educacionais nos Estados Unidos, principalmente por ser de fácil manuseio, além de trabalhar com modelos lineares hierárquicos com dois níveis. O HLM fornece a opção de efetuar análise contextual e de curvas de crescimento. Duas versões deste programa estão disponíveis, ambas escritas em FORTRAN77.

Duas rotinas estão disponíveis, a primeira é do algoritmo EM e a outra é a "V-conhecida" que assume que os componentes de variância e covariâncias são valores conhecidos.

Como opção de resultados, o HLM fornece a opção de vários testes estatísticos para coeficientes de regressão (teste t), para componentes de variância (teste Qui-Quadrado) e teste para homogeneidade de variâncias.

O formato adotado pelo HLM é interativo tornando fácil o seu uso. Além disso, permite muitas possibilidades de exploração dos dados.

O manual é claro e o programa possui poucas opções para o usuário, com interface simples fornece estatísticas descritivas e testes de hipóteses muito utilizados.(Kreft & Leeuw,1998).

4.4. VARCL

Escrito por Longford (1990), o VARCL é um programa para análise de componentes de variância de dados hierárquicos e é designado como um programa para análise de coeficientes aleatórios, e não como um programa de análise multinível. O

usuário tem liberdade de decidir antecipadamente se o coeficiente tem efeito fixo ou aleatório, no entanto, não é possível criar interações entre os níveis de uma maneira direta. Há duas versões para o VARCL, a primeira analisa o intercepto e o coeficiente de inclinação como aleatórios, para modelos com mais de três níveis; e a segunda trata com os coeficientes aleatórios em modelos com mais de nove níveis de agrupamento, no entanto considera somente uma estrutura simples de componentes de variância para os efeitos aleatórios.

A linguagem de programação base é a FORTRAN77 e requer um ambiente de cálculo iterativo. Pode-se executar este programa a partir de sistemas operacionais como o VAX/VMS, MS-DOS, MSC-OS e em alguns ambientes UNIX.

O modelo ajustado pelo VARCL é diferente daqueles ajustados por outros softwares como GENMOD, HLM e ML3 pois não constrói as interações cruzadas de níveis por *default*. Contudo, há a possibilidade de adicionar uma matriz contendo as interações entre os níveis das variáveis. O algoritmo adotado é o escore de Fisher

A informação gerada pode ser salva em um arquivo contendo um resumo das especificações iniciais. O programa é de fácil uso e interativo, no entanto as variáveis que representam a interação entre os níveis devem ser criadas antes de iniciar a análise.

Uma vantagem do VARCL é velocidade de convergência, além da adaptação de quase-verossimilhança para conjuntos de resposta não normais.

4.5. SPSS

O Statistical Package for Social Science, SPSS, é um pacote estatístico muito difundido em vários meios de pesquisa, especialmente nas áreas de Ciências Sociais, onde ocorrem a maioria das aplicações de modelos multiníveis.

O módulo VARCOP do SPSS possibilita estimar modelos de intercepto aleatório e incluir efeitos cruzados aleatórios. Em termos de análise de variância, este módulo

permite selecionar os fatores e especificá-los como fixos ou aleatórios (Sinjders & Bosjer, 1999).

4.6. STATA

O programa STATA contém alguns módulos que permitem a estimação de certos modelos multiníveis. Um deles, o módulo "**loneway**" fornece estimativas de um modelo sem efeitos. Já a série de módulos XTs são designadas para análises de dados longitudinais que podem ser ajustados em modelos de intercepto aleatório de dois níveis.

Existem comandos específicos para estimar modelos com coeficientes aleatórios com médias a posteriori calculadas, e outros comandos que fornecem estimativas de modelos multiníveis de regressão de Poisson e Probit. Estas estimativas são baseadas em métodos de equações generalizadas.

Uma facilidade deste software é a utilização de um estimador conhecido como estimador de Huber, que quando usado combinado com a palavra "cluster" calcula os erros padrões que são assintoticamente corrigidos sob uma amostragem em dois estágios.

4.7. ML3

No início dos anos 90, desenvolveu-se o ML3, como parte de um projeto de modelos multiníveis do Instituto de Educação da Universidade de Londres, que tinha o propósito de difundir a teoria multinível e aplicá-la em conjunto de dados reais. Um aspecto positivo deste programa é o fato de fornecer um conjunto completo de janelas que oferecem facilidades no manuseio e operações de transformações nos dados.

Uma diferença do ML3 com os demais softwares é que os dados não são reduzidos a estatísticas suficientes e posteriormente armazenados, toda a matriz de

dados é armazenada, não apresentando grandes problemas quando ao tamanho do banco.

Basicamente estima-se modelos com dois ou três níveis através do ML3. Há possibilidades também de possuir estrutura de erro mais complexo incorporando mais termos de erro no nível 1. Modelos log-lineares e logísticos também podem ser avaliados usando extensões padrões de Modelos Lineares Generalizados .

Um algoritmo implementado é o Mínimo Quadrados Generalizados Iterativos (IGLS) e fornece estimativas dos parâmetros dos modelos que são equivalentes às estimativas encontradas pelo procedimento de máxima verossimilhança quando a suposição de normalidade está satisfeita. Há também a possibilidade de escolher outro algoritmo, o RIGLS, que fornece estimativas restritivas de máxima verossimilhança.

A leitura dos dados pode ser feita através de um arquivo que possua os dados em linha ou ainda num arquivo que contenha dados de níveis superiores (macro) e inferiores(micro) juntos. Os dados devem ser sorteados por contexto e se faz necessária à utilização de variáveis identificadoras em cada um dos níveis de hierarquia. Se houver a presença de missing, estes recebem um código, também numérico.

O programa ML3 permite ao usuário maior liberdade de ajuste durante a execução da análise, e possui um grande potencial de uso que ainda não foi explorado.

4.8. MODELAGEM MULTINÍVEL ATRAVÉS DO MlwiN

A partir do ML3 foi desenvolvido o MLwiN, que realiza diferentes análises de modelos multiníveis com uma interface gráfica e facilidades gráficas de diagnóstico e manipulação dos dados.

Este pacote, disponível em versão para ambiente Windows, permite ajustar um modelo composto de vários níveis além de casos ponderados, erros de medidas e estimativas robustas dos erros padrões. Possui também linguagens MACRO de alto nível propiciando que uma grande variedade de facilidades seja implementada.

O método de estimação do MLWin é o IGLS (mínimos quadrados generalizados iterativos) e o GIRGLS (máxima verossimilhança restrita) e junto a isso ainda possui diagnósticos gráficos para modelos Bayesianos, que utilizam o Método de Monte Carlo com cadeia de Markov (MCMC), bem como foi implementado o bootstrap iterativo para a estimação dos modelos lineares generalizados multiníveis.

A estrutura do MlwiN é essencialmente a de uma planilha com colunas representando as variáveis e linhas representando os casos.A janela principal do MlwiN apresenta uma barra de menu e outra de ferramentas, onde se encontram comandos referentes aos procedimentos de estimação dos modelos. A opção **File** da barra de ferramentas abre a possibilidade de abrir a planilha.

Para visualizar as planilhas de dados e escolher aquela de interesse, selecione a opção **Open Worksheet** do menu **File** e clique em **OK**. Já para visualizar o conjunto de dados é preciso selecionar alguma das opções fornecidas pelo menu **Data Manipulation**. Nesta ferramenta estão disponíveis operações realizadas sobre a base de dados, como visualizar suas características.

Escolhendo a opção **Names** é possível visualizar característica do conjunto de dados avaliado, tais como número de casos, número de dados "missing", valores mínimo e máximo para cada variável. Se for de interesse alterar o nome de uma variável, basta selecionar a palavra, escrever seu novo nome e pressionando a tecla **enter** e a substituição será efetuada.

Caso haja a presença de variáveis categóricas e é de interesse definir suas categorias, basta dar um duplo clique sobre a variável e então escolher a opção **categories**. Após detalhadas as categorias, clique em **Apply** e em seguida em **Quite** para confirmar a mudança. No menu **File** você pode também salvar a planilha com as novas alterações através da opção **Save Worksheet as...** .

O comando central para especificar os modelos é dado na janela **Equations**. Nela, pode-se definir as equações e obter as estimativas dos parâmetros do modelo. Clicando no botão **Names** pode-se incluir o nome das variáveis nas equações. A notação ~ N(XB, W) informa que o modelo assume uma distribuição normal, e para trocar a

distribuição para binomial basta clicar sobre o **N**. Há ainda a opção de colocar as estimativas e erros padrões no lugar dos betas e sigmas.

Clicando sobre os parâmetros do modelo é possível definir se são fixos ou aleatórios e a que nível de hierarquia eles pertencem. Quando a especificação de um parâmetro é mudada, imediatamente as equações se alteram e durante o processo de iteração pode-se visualizar as estimativas correntes. Outra janela mostra as estimativas em uma tabela e um gráfico das sucessivas estimativas no processo de iteração, no comando **Trajectories**. Caso as iterações estejam ainda sendo computadas, as estimativas vão sendo atualizadas continuamente. Este comando é muito usado em analises de convergência.

Na janela de resíduos, **Residuals,** é possível especificar o calculo dos resíduos e então usa-los na janela de gráficos, **Graphs**, para construir vários tipos de gráficos.

Os dois novos métodos de estimação do MlwiN disponíveis são o de MCMC e bootstraping. O método de bootstrap calcula amostras bootstrap adicionando aleatoriamente resíduos de uma distribuição de erros adequada. Este é usado para corrigir estimativas ou viés e construir erros padrões e intervalos de confiança caso haja desconfiança dos erros padrões assintóticos no caso de amostras pequenas.

Os métodos de simulação de Monte Carlo com Cadeias de Markov são diferentes do método apresentado anteriormente. São métodos Bayesianos e por isso é preciso fornecer estimativas a priori para as distribuições de probabilidade para os parâmetros de interesse. Por *default* o MlwiN utiliza uma distribuição Normal para os coeficientes de regressão e uma distribuição Uniforme entre 0 e um número muito grande para as estimativas iniciais das variâncias e covariâncias.

As estatísticas Bayesianas combinam a distribuição a priori em uma distribuição a posteriori dos parâmetros. Se a distribuição a posteriori for simples, pose-se calcular a media ou moda e estabelecer intervalos de confiança, no entanto, geralmente as distribuições são muito complexas. MCMC são métodos utilizados para estimar valores dos parâmetros das distribuições a posteriori. A partir de uma quantia suficiente de

amostras independentes de uma distribuição posteriori pode-se calcular a média, moda e intervalos de confiança nesta amostra.

O MlwiN tem dois procedimentos: Gibbs sampling e Metropolis-Hastings. Os métodos de MCMC tipicamente têm dois estágios: um de *burn-in* e outro de monitoramento ou amostragem. No estágio *burn-in*, os cálculos de MCMC são feitos para um número grande de iterações, para permitir uma convergência em uma correta distribuição a posteriori. A seguir, inúmeras iterações são coletadas para serem usadas para estimar os valores dos parâmetros e de seus erros padrões.

Dois problemas podem ocorrer nessa fase, ou não foi possível obter um número suficiente de iterações no primeiro estágio a fim de obter uma convergência a uma correta distribuição a posteriori, ou ainda podem existir problemas quando a coletar amostras independentes de uma distribuição posteriori para garantir boas estimativas. Muitas vezes os métodos de MCMC produzem resultados altamente correlacionados, assim, a suposição de coletar amostras independentes podem não estar satisfeita, para solucionar isto, é preciso aumentar o número de amostras coletadas.

O MlwiN fornece algumas inspeções muito úteis para diagnóstico dos resultados de MCMC, dentre as quais destaca-se gráficos de parâmetros para as sucessivas iterações e auto-correlação. Na janela de trajetórias é possível visualizar os valores dos parâmetros e clicando no gráfico é possível acessar ao diagnóstico do MCMC e então obter uma ajuda para decidir se o número de iterações mostrou-se suficiente para o estágio *burn-in* no monitoramento do processo.

É apresentado também um gráfico da Função de Auto-Correlação (ACF) onde se pode avaliar o grau de dependência dos resultados da simulação MCMC. O número apresentado em **NHAT** significa o número de iterações requeridas para a estimação dos intervalos de confiança.

Uma opção do MLwiN é a facilidade para comparar diferenças entre métodos de estimação. Através de uma análise de sensibilidade é possível comparar os resultados de vários métodos, comparando seus resultados. Se esses apresentarem grandes diferenças não se deve confiar nos seus resultados, ou por que os dados não possuem informação

suficiente, devendo então coletar mais observações, ou por que os parâmetros que estão sendo estimados devem ser modificados, sugerindo uma alteração no modelo ajustado.

A versão deste programa para o Windows está ainda sendo melhorada, mas já apresenta uma visualização clara e com muitas opções, outras estarão sendo implementadas em breve. Seu manual de ajuda é muito acessível, incluindo informações sobre a construção de macros.

4.8.1. IMPORTANDO DADOS do SPSS PARA O MlwiN

Considerando uma situação em que os dados tivessem terminação.**sav** (utilizada pelo SPSS), é preciso alterar a extensão do banco de dados para.**txt.**

A seguir serão listadas etapas a serem percorridas até os dados serem importados para a janela do MlwiN:

1. Converter o arquivo.**sav** para o tipo FIXED ASCII, cuja extensão é .**dat**.
2. Salvar o arquivo no Desktop do computador.
3. Para converter o arquivo para a terminação **.txt**, é preciso abrir a janela do Prompt do MS DOS.
4. Se o prompt do DOS estiver na pasta **C:\Windows**, escreva o seguinte comando **cd desktop.**
5. A inscrição deve ter sido alterada para **C:\WINDOWS\DESKTOP**.
6. Para renomear o arquivo digite os seguintes comandos **REN TESTE.DAT TESTE.TXT**(teste é o nome do arquivo que se deseja alterar o nome).
7. Para encerrar a seção basta digitar **EXIT**.
8. Se o arquivo contiver casas decimais é preciso converter as vírgulas para pontos.
9. Antes de iniciar o processo de importação dos dados é preciso anotar o número de caracteres máximo presente em cada uma das variáveis, incluindo sinais (- ou +) e vírgulas. Por exemplo, se uma determinada variável assume o valor –6.25, o número de caracteres contidos nela é 5.

10. Para importar os dados para o MlwiN, abra o programa e clique em **ASCII tex file input** conforme a figura abaixo:

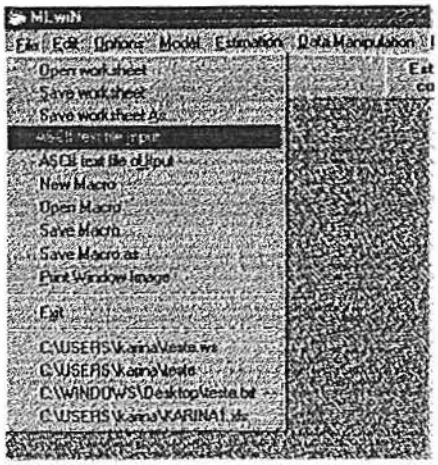

Figura 6. Importando dados com formato ASCII.

11. A caixa de diálogo que será ativada diz respeito a formatação do banco de dados a ser importado.
12. Deve-se, no primeiro campo, especificar quantas colunas formam o banco de dados.(no caso de 3 variáveis serão 3 colunas : **C1-C3**).
13. Supondo um banco de dados com três variáveis uma contendo variáveis com 3 caracteres, a outra com 6 e a última com 5 caracteres, a instrução seria dada por : **(3,-1,6,-1,5)**. O número –1 é utilizado para dar um espaço de 1 caracter entre uma coluna e outra do banco de dados.

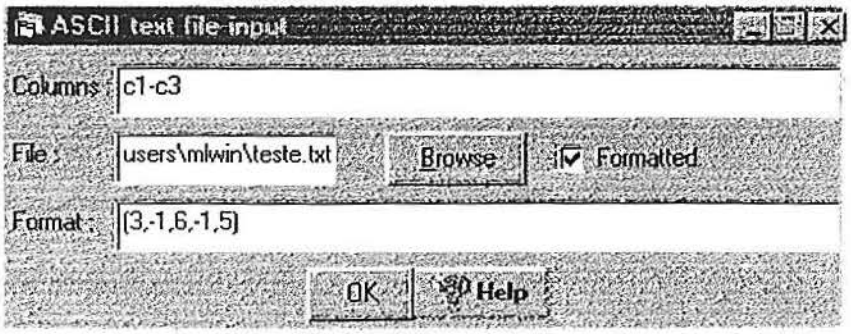

Figura 7. Definindo as colunas do banco de dados

14. No botão Browse defina a localização do arquivo para a importação.
15. Clique em **OK** para confirmar a importação.
16. Para salvar o banco de dados em formato **.ws** característico do MlwiN clique em **Save Worksheet** no comando **File** da barra de ferramentas.

17. Após salvar já é possível abrir e visualizar normalmente o banco de dados recém importado.

4.9. MODELAGEM MULTINIVEL ATRAVES DO SAS

Um dos pacotes estatísticos mais respeitados no meio científico é o Statistic Analysis Sistem, SAS, e a partir de 1996 vem trabalhando com a estimação de modelos multiníveis complexos e a apresentação de inúmeras estatísticas de ajuste através do procedimento MIXED.

O SAS é executado em ambiente Windows, e sua apresentação gráfica tem apresentado grandes melhorias a partir da versão 8, recentemente implementada.

O procedimento MIXED ajusta uma grande variedade de modelos mistos lineares. Um modelo linear misto é uma generalização de modelos lineares padrões que permite que os dados exibam correlações e variabilidade não constante, característica muito comum em dados hierárquicos.

Os modelos mistos fornecem uma flexibilidade de modelagem tanto para a média dos dados quanto para as variâncias e covariâncias. Os efeitos fixos deste modelo estão associados às variáveis explicativas, que podem ser quantitativas ou qualitativas. Uma ampla variedade de estruturas de covariâncias estão disponíveis neste procedimento e são utilizadas para modelar as estruturas de covariâncias dos parâmetros de efeitos aleatórios. As variâncias desses efeitos são geralmente conhecidas como componentes de variância e tornam-se parâmetros de uma particular estrutura.

O procedimento MIXED calcula várias estatísticas de ajuste diferentes e permite a execução de testes de hipóteses e intervalos de confiança. Os métodos de estimação implementados através do algoritmo de Newton-Raphson são os de máxima verossimilhança restrita (REML) e máxima verossimilhança (ML), dentre outros. Além disso, sua acessibilidade é muito fácil o que permite ajustar inúmeros modelos e posteriormente compará-los entre si, mesmo com a presença de desbalanceamento dos dados.

Utilizando o PROC MIXED pode-se ajustar três tipos comuns de modelos multiníveis, (1) modelos de "efeitos de Escola" com dois fatores, para dados com indivíduos agrupados dentro de hierarquias; (2) modelos de "crescimento individual" com dois níveis, para explorar dados longitudinais ao longo do tempo; (3) modelos com três níveis que combinam estas duas características.

A lógica atrás da sintaxe do PROC MIXED pode ser entendida considerando modelos de médias incondicionais para um conjunto de dados clássico de efeitos de escola com dois níveis. Suponha que você tenha dados com dois níveis de hierarquia – ou seja, estudantes dentro de escolas – e você gostaria de examinar o comportamento do nível 1 resultando Y_{ij}. A Análise geralmente começa com um modelo de médias incondicionais com nenhum preditor, que expressem Y_{ij} como a soma do nível 2 "intercepto"(β_{0j}) e um nível 1 aleatório de erro(r_{ij}) associado com o i-ésimo estudante da j-ésima escola. O nível 2 expressa os interceptos de nível-escola como uma soma de uma média geral (γ_{00}) e uma série de desvios aleatórios da média (u_{oj}). Mais adiante assumiremos que $r_{ij} \sim N(0, \sigma^2)$ e $u_{oj} \sim N(0, \sigma^2_{u0})$. O modelo pode ser apresentado como

$$Y_{ij} = \beta_{0j} + r_{ij}$$

$$\beta_{0j} = \gamma_{00} + u_{oj}$$

que podem ser reescritos como $Y_{ij} = \gamma_{00} + u_{oj} + r_{ij}$ (35)

A significância combinada em (35) é mais útil para especificar um modelo no PROC MIXED. Como destacado pelo parêntesis, o modelo é composto de duas partes, uma fixa que contém um único termo γ_{00} (para o intercepto) e um para a parte aleatória, que contém dois termos (para o intercepto u_{0j} e para os resíduos dentro da escola r_{ij}). A sintaxe do PROC MIXED reflete esta partição:

```
proc mixed;
class school;
model y = /solution;
random intercept/sub=school;
```

Depois de solicitar o procedimento e identificando as variáveis (usando **class**), o enunciado **MODEL** especifica os efeitos fixos e o enunciado **RANDOM** especifica os efeitos aleatórios. Este modelo parece estranho pois parece como se não tivesse preditores. Na verdade ele tem um preditor implícito, o vetor de médias, sendo composto por valores um, referente ao intercepto, incluído por *default*. (O intercepto pode ser suprimido adicionando a opção **/NOINT** no enunciado **MODEL**) a opção **/SOLUTION** permite ao SAS imprimir as estimativas para os efeitos fixos.

O enunciado **RANDOM** é crucial e sua especificação é geralmente uma parte delicada no uso do **PROC MIXED**. Como a maioria dos programas de regressão, o procedimento assume um efeito aleatório para o menor nível residual (aqui, dentro de escola), r_{ij}. Especificando o intercepto no enunciado aleatório, nós indicamos a presença de um segundo efeito aleatório que implica que o **INTERCEPT** deveria ser tratado não somente como um efeito fixo (γ_{00}), mas também como um efeito aleatório(τ_{00}). A opção **SUB=** no enunciado **RANDOM** especifica uma estrutura multinível, indicando como as unidades no nível 1 estão divididas dentro das unidades do nível 2 . Aqui os subgrupos são designados pela variável de classificação **SCHOOL**. Sem o enunciado **RANDOM** , o modelo ajustado seria

$$Y_{ij} = \gamma_{00} + r_{ij} \qquad (36)$$

O resultado apresentado é compreendido nas quatro seções de resultados, a primeira é a sessão Histórico de Iteração que fornece a informação de convergência. Pelo fato do programa ser eficiente, a convergência é geralmente rápida. Com modelos mais complexos ou com forte desbalanceamento em grandes conjuntos de dados, a convergência se torna mais lenta. Outra seção refere-se a seção Parâmetros de Covariância que fornece estimativas e testes de hipóteses para os parâmetros de covariância para os efeitos aleatórios. A seção Ajuste do Modelo fornece informações úteis para a comparação do ajuste do modelo entre múltiplos modelos com os mesmos efeitos fixos, mas diferentes efeitos aleatórios. Dois critérios provavelmente mais úteis são o critério de Informação de Akaike e o critério Bayesiano de Schwarz. A seção final apresenta as estimativas dos parâmetros e os testes de hipóteses para os efeitos fixos.

A inclusão de preditores do nível 1 e nível 2 não se torna muito complicada quando a sintaxe básica é entendida. Então, fica fácil incluir os preditores de nível 1 e nível 2. Efeitos fixos adicionais são especificados no enunciado **MODEL**; efeitos aleatórios adicionais são especificados no enunciado **RANDOM**. Interações entre preditores (no mesmo nível ou em níveis diferentes) são especificadas usando a sintaxe **VAR1*VAR2**. Preditores categóricos podem ser especificados usando cada um uma variável dummy para definição de usuário ou adicionando-as no enunciado **CLASS**.

Muitos pesquisadores querem preditores para "centros" qualquer média-geral ou média contextual. Diferente de alguns programas multiníveis que permitem o usuário centralizar quando da especificação do modelo, o usuário do **PROC MIXED** deve centrar em uma etapa anterior no **PROC SUMMARY,** que calcula a média geral e contextual, facilmente.

Para ilustrar, considere dois preditores: **MEANSES**, a média da escola **SES** e **CSES**, e a diferença entre cada **SES** dos estudantes e a média da escola. Nós ajustamos um modelo de interceptos e inclinação como efeitos aleatórios, da seguinte forma:

```
proc mixed;
class school;
model y = cses meanses cses*meanses/solution;
random intercept cses/
sub=school type=un;
```

Note a similaridade entre esta sintaxe para um modelo de médias incondicionais. Três efeitos fixos são adicionados ao **MODEL** – um para cada preditor e um terceiro para a interação. Um único efeito aleatório, **CSES**, é adicionado no **RANDOM** para indicar que as inclinações CSES de estudantes são permitidas variar através das escolas. **TYPE=UN** indica que a matriz de variância e covariância para os efeitos aleatórios é completamente geral, chamando o SAS para estimar o componente de variância para os **INTERCEPTs** e para a inclinação de CSES bem como para a covariância entre eles.

Também Modelos de crescimento individual podem ser ajustados usando o enunciado **RANDOM**, que imita uma análise de efeito de escola ou usando o enunciado **REPEATED**, que imita a análise de variância de medidas repetidas. De qualquer modo, você deve primeiro criar um conjunto de dados com um período de

tempo em que cada indivíduo tenha um registro para cada período do tempo que ele foi observado.

Considere o modelo de crescimento individual básico:

$$\begin{aligned} Y_{ij} &= \pi_{0j} + \pi_{1j}(TIME)_{ij} + r_{ij} \\ \pi_{0j} &= \gamma_{00} + u_{0j} \\ \pi_{1j} &= \gamma_{10} + u_{1j} \end{aligned} \qquad (37)$$

onde nós assumimos que $r_{ij} \sim N(0,\sigma^2)$ e $\begin{pmatrix} u_{0j} \\ u_{1j} \end{pmatrix} \sim \begin{bmatrix} \tau_{00} & \tau_{01} \\ \tau_{10} & \tau_{11} \end{bmatrix}$ (38)

O modelo em (37), pode ser escrito como :

$$Y_{ij} = [\gamma_{00} + \gamma_{10}TIME] + [u_{0j}u_{ij}\pi_{1j}TIME_{ij} + r_{ij}] \qquad (39)$$

e pode ser ajustado usando a seguinte linguagem de programação:

```
proc mixed;
class person;
model y = time/solution;
random intercept time/
subject=person type=un;
```

Note a similaridade da sintaxe para a análise de efeito escola. A variável **CLASS** muda de escola (**SCHOOL**) para pessoa (**PERSON**) para indicar que os dados representam múltiplas observações ao longo do tempo para os indivíduos. Usando esta variável **CLASS** no enunciado **RANDOM** (com **SUBJECT=** opção) permite-se que intercepto e inclinação variem de pessoa a pessoa.

Este modelo estabelece uma suposição comum, mas não realística, sobre o comportamento de r_{ij}, o resíduo de dentro de pessoa ao longo do tempo. Ajustando um modelo em que somente os interceptos variem entre pessoas, nós assumimos uma matriz de covariância componente simétrico para os erros, para cada pessoa. Ajustando um modelo em que a inclinação possa variar também, nós introduzimos a heterocedasticidade dentro desta matriz. Um dos recursos do PROC MIXED é permitir

que o usuário compare diferentes estruturas para a matriz de covariância dos erros. Em vez de um modelo de crescimento individual em (37)-(39), considere a seguinte alternativa:

$$
\begin{aligned}
&Y_{ij} = \pi_{0j} + \pi_{1j}(\text{TIME})_{ij} + r_{ij} \\
&\pi_{0j} = \gamma_{00} \qquad\qquad \pi_{1j} = \gamma_{10} \\
&r_{ij} \sim N(0, \Sigma)
\end{aligned}
\tag{40}
$$

Aqui, o intercepto e as taxas de crescimento são assumidos constantes entre as pessoas. Mas o modelo introduz um tipo diferente de complexidade: as observações residuais dentro de pessoas (depois de controlar pelo efeito linear de **TIME**) são correlacionadas através da matriz de variância e covariância dos erro Σ dentro de pessoas. Especificando estruturas para Σ alternativas (que são de modo ideal derivadas da teoria)o usuário pode comparar o ajuste de diferentes modelos.

O SAS tem dezenas de estruturas de covariância dos erros alternativas. Se existem somente três variações nos dados vale a pena explorar poucas estruturas, pois há poucos dados para cada pessoa. Com observações adicionais por pessoa, explorando estas alternativas para Σ (chamadas **R** no **PROC MIXED**) pode ser eficaz.

Este programa ainda possibilita a analise de modelos com três níveis ou mais. O **PROC MIXED** pode ser usado para ajustar modelos de três níveis ou mais. No caso da análise de "efeito escola", você pode incluir um enunciado **RANDOM** a mais, com a especificação de agrupamento (*nested)* adequada dada em **SUB**=opção. Por exemplo, se você tem dados de estudantes dentro de professores e dentro de escolas, você pode ajustar um modelo de médias incondicionais com a sintaxe:

```
proc mixed;
class teacher school;
model y = /solution;
random intercept/sub=school;
random intercept/
sub=teacher(school);
```

Neste caso a análise longitudinal que segue indivíduos aninhados dentro de grupos, e especificações de efeito escola podem ser adicionadas na especificação de modelo de crescimento individual.

Outras opções estão disponíveis para usuários interessados em modelos mistos mais complexos. Heterogeneidade na matriz de variância e covariância dos erros pode ser introduzida usando a opção **/GROUP** no comando **RANDOM**. Amostras baseadas na análise bayesiana pode ser conduzida usando o comando **PRIOR** que permite uma variedade de especificações de distribuições para a densidade a priori dos parâmetros de componentes de variância. O SAS também fornece duas macros – **GLMMIX e NLINMIX** que podem ser usadas para ajustes de modelos lineares generalizados mistos e modelos não lineares mistos que não envolvem uma conseqüência de normalidade contínua como tratado aqui.

Esta flexibilidade, torna prático o uso do **PROC MIXED**. Mais, ele não é escrito com a intenção de ser utilizado em modelos multiníveis. Uma vez entendida a sintaxe básica, diversos tipos de modelos podem ser ajustados.

4.10. OUTROS PROGRAMAS

Especializado em calcular o poder em delineamentos em dois níveis, o Pin T pode ser usado com o objetivo de se obter uma estimação a priori dos erros padrões dos coeficientes fixos.

O Mplus é outro programa com várias opções de estruturas de covariâncias, e permite a análise de dados multiníveis univariados e multivariados segundo metodologia hierárquica quanto metodologia de "Path Analysis" (ou análise de trajetórias), análise fatorial e modelos de equações estruturais (Snidjers & Bosker, 1999).

Outra alternativa é a utilização do software MLA, que calcula estimativas para modelos com dois níveis através de métodos de reamostragem. Várias implementações bootstrap e "jackknife" são incluídas.

Um programa que possibilita a estimação de modelos hierárquicos combinados com modelos para equações estruturais e medidas de erro é o BUGS. Ele utiliza também um procedimento de simulação para cálculo de estimativas Bayesianas conhecido como Gibbs sampling. Deve-se tomar cuidado pois este programa requer dados balanceados.

5. APLICAÇÃO EM DADOS LONGITUDINAIS

5.1. CARACTERIZAÇÃO

Dados com medidas repetidas surgem em vários contextos nas mais diversas áreas do conhecimento, tais como Epidemiologia, Agronomia, Farmacologia, Indústria, Economia, Ciências Sociais entre outras. O uso dessas medidas se dá quando o objetivo do estudo é avaliar o comportamento de uma variável resposta ao longo de uma dimensão específica, como o tempo ou o espaço. Define-se dados de medidas repetidas como dados gerados observando um número de unidades de investigação repetidamente sobre diferentes condições de avaliação, assumindo-se que as unidades de investigação constituem uma amostra aleatória da população de interesse. Um tipo de medida repetida são os dados longitudinais, obtidos quando as observações são ordenadas no tempo ou na posição do espaço.

Geralmente três tipos de análises estatísticas são aplicadas quando há casos de medidas repetidas, a Análise de Variância Univariada, Análise de Variância Multivariada e Métodos baseados em Modelos Mistos. No entanto pode-se abordar a análise de medidas repetidas segundo a ótica de modelos multiníveis. Estas medidas podem ser organizadas segundo uma forte estrutura hierárquica uma vez que as ocasiões de medida em estudo são agrupadas em unidades de nível mais baixo e os sujeitos pertencem a unidades de nível mais elevados, gerando uma maior variação entre indivíduos do que entre ocasiões dentro de um mesmo indivíduo. Dessa forma, correlações entre as observações nos diferentes níveis de hierarquia são calculadas e avaliadas pois influenciam muito no resultado do ensaio.

Nos modelos multiníveis, as unidades de cada nível são vistas como tendo um efeito aleatório, ou seja, são amostras aleatórias de uma população dessas unidades. Esses efeitos aleatórios serão responsáveis no modelo por coeficientes aleatórios, que levam em conta a variabilidade entre essas unidades, seja por aleatoriedade no intercepto ou nos coeficientes de regressão.

Uma vantagem do uso de modelos multiníveis com dados de medidas repetidas em relação aos demais modelos, exceto os modelos mistos, é que estes não exigem que haja um mesmo número de ocasiões de medidas por indivíduo, como ocorre muitas vezes em estudos longitudinais, quando a perda de uma ou mais ocasiões de medidas é muito comum.

Na análise de medidas repetidas, um modelo, clássico, muito utilizado é o modelo com estrutura de covariância componente simétrico, que é equivalente a um modelo de intercepto aleatório. Muitas vezes este modelo é chamado de modelo de efeitos aleatórios ou modelo misto, muito embora seja uma simples especificação, das múltiplas possibilidades dos modelos de efeitos aleatórios ou modelos mistos. O modelo com estrutura de covariância componente simétrico não possui variáveis explicativas, exceto as ocasiões de medidas. Assim, este delineamento é um delineamento puramente dentro de sujeitos, e o valor esperado para a medida na ocasião t pode ser denotado por μ_t e seu modelo expresso por:

$$Y_{ti} = \mu_t + U_{0i} + R_{ti} \qquad (41)$$

A suposição deste modelo é que os termos U_{0i} e R_{ti} sejam independentes e normalmente distribuídos com média zero e variâncias iguais a σ_U^2 e σ^2 respectivamente. Cabe salientar que a parte fixa deste modelo não possui um termo constante, mas é baseado em *m* variáveis *dummies* (denotadas por d_{hti}, com h=1,...m) para as *m* ocasiões de medidas, onde

$$d_{hti} = \begin{cases} 1, & t = h \\ 0, & t \neq h \end{cases} \qquad (42)$$

Estas variáveis d_{hti} assumem valor um quando t=h e são iguais a zero caso contrário. Assim, a parte fixa do modelo pode ser escrita como:

$$\mu_t = \sum_{h=1}^{m} \mu_h d_{hti} \qquad (43)$$

e o modelo com estrutura componente simétrico pode ser formulado como a forma comum da parte fixa de um modelo linear hierárquico:

$$Y_{ti} = \sum_{h=1}^{m} \mu_h d_{hti} + U_{0i} + R_{ti} \qquad (44)$$

Considerando um estudo onde as medidas foram coletadas em três ocasiões distintas de tempo, a equação (44) seria escrita da seguinte maneira:

$$Y_{ti} = \mu_1 d_{1ti} + \mu_2 d_{2ti} + \mu_{31} d_{3ti} + U_{0i} + R_{ti} \qquad (45)$$

Ainda considerando a estrutura de um modelo componente simétrico, é possível aprimorar o seu ajuste através da inclusão de variáveis explicativas na parte fixa do mesmo, relativas tanto ao indivíduo quanto às ocasiões de medidas. Se houver a inclusão de variáveis que trazem alguma informação sobre características individuais dos sujeitos, segundo a ótica da modelagem multinível, diz-se que elas pertencem ao segundo nível de hierarquia e cada uma é chamada de *variável entre-sujeitos*. Já se as variáveis incluídas descreverem o tempo entre duas ocasiões de medidas elas são ditas *variáveis dentro de sujeitos*. As variáveis entre e dentro de sujeitos são efeitos principais do modelo ajustado, no entanto é possível que haja efeitos de interação entre elas, caracterizando uma interação de níveis cruzados (Snijders & Bosker, 1999).

O modelo componente simétrico, que exige que a matriz de covariâncias seja constante, ou seja, todas as variâncias são iguais e também todas as covariâncias são iguais, é um modelo muito restritivo, pois acredita-se, em princípio, que medidas mais próximas são mais correlacionadas que medidas mais afastadas no tempo. Em certas ocasiões esta suposição pode ser relaxada com a simples inclusão de coeficientes de inclinação aleatórios no modelo.

Um modelo com intercepto aleatório e um coeficiente para cada instante t de avaliação é dado por:

$$Y_{ti} = \text{parte fixa} + U_{0i} + U_{1i}(t - t_0) + R_{ti} \qquad (46)$$

Este modelo possui um componente aleatório U_{1i} dependente do indivíduo, e os desvios U_{0i} que afetam todos os valores de Y_{ti} do mesmo modo. O efeito aleatório de tempo pode também ser descrito como uma interação aleatória tempo*individuo. As variáveis U_{1i} e U_{0i} são assumidas como independentes e possuírem uma distribuição normal bivariada com média zero e variâncias σ^2_{U0} e σ^2_{U1} e covariância igual a σ_{U10}.

As variâncias e covariâncias das medidas Y_{ti}, condicionado a variáveis explicativas são dadas por:

$$\begin{aligned} VAR(Y_{ti}) &= \sigma_0^2 + 2\sigma_{01}(t - t_0) + \sigma_1^2(t - t_0)^2 + \sigma^2 \\ COV(Y_{ti}, Y_{si}) &= \sigma_0^2 + \sigma_{01}\{(t - t_0) + (s - t_0)\} + \sigma_1^2(t - t_0)(s - t_0) \end{aligned} \qquad (47)$$

onde t é diferente de s. Estas fórmulas expressam o fato de que as variâncias e covariâncias são variáveis ao longo do tempo, caracterizando assim a heterocedasticidade. Logo, a correlação entre as medidas de tempo depende dos seus espaçamentos, bem como das suas posições.

Quando há a inclusão de mais que um coeficiente aleatório, há uma melhoria no ajuste da parte aleatória do modelo e dessa forma, a análise passa a ser vista como uma análise de tendência polinomial (Snijders & Bosker, 1999).

Em certas situações, um modelo com estrutura de covariância componente simétrico torna-se inadequado, especialmente quando há grande número de observações. Nestes casos, é recomendado utilizar um modelo multivariado completo, que não possui restrições

quanto a matriz de covariâncias, e conseqüentemente possui mais parâmetros a serem estimados.

O modelo multivariado completo (48) torna-se uma expansão do modelo componente simétrico, e é reconhecido como um modelo linear hierárquico através do uso de variáveis *dummy* (indicando as ocasiões de medida) definidas em (42)

$$\begin{aligned} Y_{ti} &= parte\ fixa + U_{ti} \\ Y_{ti} &= parte\ fixa + \sum_{h=1}^{m} U_{hi} d_{hti} \end{aligned} \qquad (48)$$

As variáveis U_{ti} para t = 1, ...,m são variáveis aleatórias do segundo nível, com esperança zero e matriz de covariâncias não estruturada, ou seja, permite acomodar quaisquer estimativas de variâncias e covariâncias dos dados. Dessa forma, este modelo não possui uma parte aleatória no primeiro nível, pois a matriz de covariâncias do vetor completo dos dados, condicionada às variáveis explicativas, é idêntica a matriz de covariâncias dos ($U_{1i}, ..., U_{pi}$). Diz-se, então, que a parte aleatória deste modelo é *saturada*, uma vez que proporciona um ajuste perfeito para a matriz de covariâncias.

Em situações que envolvam casos que não sejam de dados de medidas repetidas também pode-se ajustar um modelo multinível multivariado, conforme definido anteriormente, substituindo as m ocasiões de medidas por m variáveis diferentes que possuam aproximadamente uma distribuição normal multivariada. Em softwares de análise multinível, esta opção aparece como uma análise multivariada de dados incompletos, considerando que os valores faltantes são aleatórios. (Snijders & Bosker, 1999).

A utilização de modelos multivariados completos na análise de medidas repetidas, com ocasiões fixas de delineamento, como um modelo linear hierárquico não encontra problemas, quanto a formulação matemática. Para a estimação dos parâmetros de modelos que não possuem partes aleatórias no primeiro nível, já estão disponíveis algoritmos e softwares que calculam as estimativas de ML e REML para distribuições normais

multivariadas com dados incompletos e com conjuntos de variáveis explicativas que são diferentes para diferentes variáveis dependentes.

Através da utilização de variáveis *dummy* tem-se uma facilidade na obtenção da matriz de covariâncias, obtida para os coeficientes de regressão aleatórios, que é a matriz de covariâncias para os vetores de dados completos. Para permitir coeficientes de regressão aleatórios para outras variáveis é preciso somente que elas sejam linearmente independentes, dependendo somente da ocasião de medida e não do indivíduo. (Snijders & Bosker, 1999)

Assim, cada ajuste do modelo para a parte aleatória com alguma restrição na matriz de covariâncias, pode ser visto como um submodelo do modelo multivariado completo. O ajuste de diferentes estruturas de covariâncias em um mesmo conjunto de dados pode ser testado através do teste da razão de verossimilhança, também conhecido com o teste da estatística de *deviance*.

Uma técnica correspondente a modelagem hierárquica multivariada é a análise multivariada de variância (MANOVA) e a análise de regressão multivariada, que trabalham com conjunto de dados completos. Estas técnicas produzem testes exatos enquanto que os resultados para a modelagem hierárquica são aproximados.

Uma aplicação da metodologia de modelos multiníveis em medidas repetidas foi realizada utilizando dados referentes ao nível de glicose registrado em pacientes gestantes participantes do Estudo Brasileiro de Diabetes Gestacional, medido em três ocasiões distintas.

5.2. DESCRIÇÃO DO ESTUDO

O Estudo Brasileiro do Diabetes Gestacional – EBDG é um estudo de coorte conduzido em seis capitais brasileiras: Porto Alegre, São Paulo, Rio de Janeiro, Salvador,

Fortaleza e Manaus. Seu início foi em maio de 1991, com término em agosto de 1995 (Branchtein et al.,2000). A descrição do estudo com detalhes é feita em Carballo (2002).

Seu principal objetivo é estudar o diabetes e a intolerância à glicose gestacional em grávidas, cujo atendimento obstétrico era feito junto ao SUS, frente a alguns fatores de risco, prevalência, dentre outros desfechos. Em sua fase inicial, realizou-se o arrolamento das gestantes, entrevistas e medições antropométricas e um teste de tolerância à glicose (TTG). Em seguida, as gestantes foram acompanhadas através de uma revisão dos prontuários até o parto, quando foram obtidas as informações referentes ao recém-nascido e ao parto. Por fim, foi avaliada a morbi-mortalidade do recém-nascido feita através de um acompanhamento de até 28 dias após o parto.

5.2.1. CARACTERIZAÇÃO DA AMOSTRA

As gestantes participantes deste estudo que apresentaram-se para consulta em um dos seis centros, possuíam idade superior a 20 anos, encontravam-se entre a 21 e 28 semana de gestação e não possuíam histórico do Diabetes de Mellitus. O estudo foi de natureza observacional, sendo que as decisões relativas ao controle de hiperglicemia foram deixadas para os clínicos que avaliaram as pacientes durante o atendimento obstétrico. As informações sobre a terapia e a dieta de insulina foram obtidas através da revisão dos prontuários ou da carteira do pré-natal. A idade gestacional das pacientes foi avaliada por três parâmetros: Ecografia obstétrica, data da última menstruação e última altura uterina, medida pelo obstetra no dia do arrolamento ou no dia imediatamente anterior ao arrolamento.

Um total de 5564 mulheres participaram inicialmente do estudo, no entanto somente 4998 concluíram o cronograma final. Para focalizar a avaliação de risco de escala de glicemia, a qual fora da gravidez é considerada prejudicada pela tolerância a glicose, 21 mulheres foram excluídas da análise, pois apresentavam critérios de diabetes (glicose de jejum> 7mmol/l ou glicose a 2 horas > 11,1 mmol/l).

Para a construção de um modelo que explique a relação entre os níveis de glicemia registrados nas pacientes, sua altura, anos de educação, histórico do diabetes na família, número de gestações anteriores, foram excluídos casos que apresentavam dados de "missing" nas medidas dos níveis de glicose de jejum, 1h e 2 h, idade, peso , altura, razão cintura–quadril. Além disso, excluiu-se também um caso especifico, determinado pelos coordenadores da pesquisa. Desta forma as análises apresentadas neste estudo foram realizadas com base nas informações de 4968 pacientes.

5.2.2. MÉTODOS DE COLETA

Um questionário estruturado aplicado por pesquisadores especialmente treinados foi aplicado a todas as mulheres. Além disso, juntamente com o arrolamento, foi realizado um teste de tolerância a glicose (TTG) de duas horas com 75 gramas entre a 24ª e 28ª semana de gestação. As pacientes foram acompanhadas até a chegada ao hospital e durante o pós-parto através da revisão do prontuário.

O teste de tolerância à glicose utilizou procedimentos padrões. Após 12 a 14 horas de jejum foram coletado em cada uma das pacientes medidas de glicose, em seguida, foi administrada uma carga de 75 gramas de glicose anihidoxica e registrou-se repetidamente os níveis de glicose uma e duas horas depois. As amostras foram obtidas da veia antecubital e coletadas em tubos contendo fluido e mantidas a uma temperatura de 4°C ate a centrifugação. Nas medidas de plasma, utilizando métodos de glicose, adotou-se um coeficiente de variação menor que 5%.

A definição de Diabetes de Mellitus Gestacional foi definida de acordo com as novas recomendações da ADA para 2 horas 75g OGTT (Oral Glucose Tolerance Test) de no mínimo dois valores maiores que a glicose de jejum de 5,5 mmol/l, a 1 hora de 10 mmol/l ou a 2 horas de 8,6 mmol/l. Os critérios utilizados estão de acordo com os sugeridos pela World Health Organization (WHO).

5.3. ESTRATÉGIA DE ANÁLISE

A estrutura multinível presente neste estudo considera como unidades pertencentes ao primeiro nível de hierarquia as medidas de glicose repetidas em três ocasiões distintas (em jejum, 1 hora e 2 horas depois das pacientes ingerirem a glicose). Como unidades do segundo nível temos as gestantes, ou seja, temos medidas agrupadas em gestantes. Poder-se-ia pensar ainda em um terceiro nível de hierarquia, uma vez que as pacientes ainda encontram-se agrupadas em centros de estudos (um em cada capital estudada). A figura 8 representa a estrutura multinível presente nos dados para cada um dos centros.

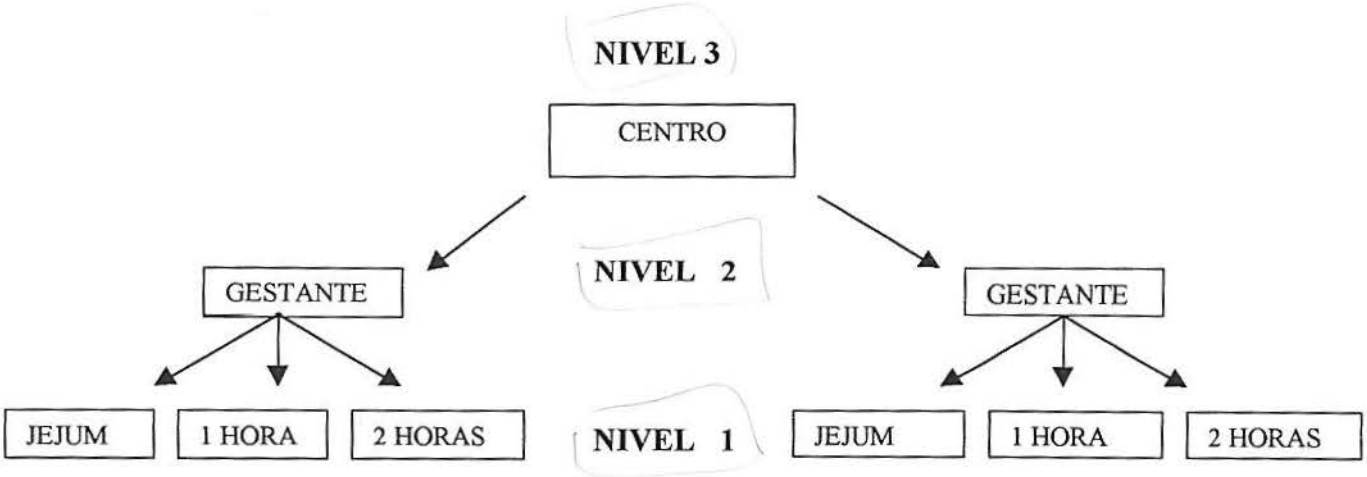

Figura 8. Estrutura de hierarquia presente nos dados.

A análise considerou todas as 4968 observações em um mesmo banco de dados. Inicialmente ajustou-se um modelo considerando o efeito do terceiro nível de hierarquia, representado pelos centros, no entanto este não se apresentou significativo, e por isso será retirado nas análises seguintes.

Foram ajustados modelos considerando o intercepto aleatório, que é equivalente ao modelo misto com estrutura de covariância componente simétrico. Na seqüência, foram acrescentadas variáveis explicativas, uma a uma, avaliando-se constantemente o valor da estatística de *deviance* e os resíduos.

5.4. RESULTADOS

A variabilidade entre os seis centros estudados não foi muito alta, e o coeficiente de correlação intra-classe, que neste caso mede o efeito de centro na variabilidade do nível de glicose das pacientes foi de 1.143% (variância estimada para o intercepto $\hat{\sigma}_U^2$ =11.3036 e variância estimada para o erro é $\hat{\sigma}^2$ =766.76), sugerindo, com o objetivo de simplificar o modelo, que não se faz necessário utilizar este nível hierárquico nas análises seguintes. Dessa forma o modelo considerado será constituído de dois níveis, o primeiro representado pelas medidas repetidas e o segundo pelas gestantes. Com esta estrutura de hierarquia ajustou-se um modelo com estrutura de covariância componente simétrico que possibilita a estimação de uma equação para cada um das pacientes, pois é definido da mesma maneira que um modelo de coeficientes aleatórios, com o intercepto aleatório.

5.4.1. Modelo Inicial

O modelo apresentado em (49) supõe que a média é igual para cada uma das medidas de glicose , ou seja, assume-se que as três ocasiões de medidas (Jejum, 1 hora e 2 horas) possuem a mesma média populacional.

$$\hat{\text{Glicose}}_{ti} = 102.43 + U_{ti} + R_{ti} \qquad (49)$$

A variância estimada para o intercepto é igual a $\hat{\sigma}_U^2$ =114.83 e a variância estimada para o erro é $\hat{\sigma}^2$ =654.53. O coeficiente de correlação intraclasse para este modelo é obtido dividindo-se a variância do nível 2 , ou seja a variância da gestante ($\hat{\sigma}_U^2$ =114.83) pela variância total, obtida pela soma das variâncias dos nível 1, de tempo, ($\hat{\sigma}^2$ =654.53) e do

nível 2, de gestante ($\hat{\sigma}_U^2 + \hat{\sigma}^2 = 769.36$), produzindo um valor de 0.1493 ou 14.93%. Logo, a estimativa da correlação intra-classe, para este caso, informa que 14.93% da variabilidade total dos níveis de glicose é devido a variação entre pacientes. O valor da *deviance* para este modelo foi igual a 141032.8 e servirá como padrão para comparação com os demais modelos ajustados posteriormente.

5.4.2. Modelo Incluindo o Efeito Fixo de Tempo

Parametrizando passo a passo, inicialmente é razoável estimar um modelo que permita que as médias de glicose variem ao longo do tempo. Os resultados obtidos com o ajuste deste modelo são apresentados na Tabela 2.

Tabela 2. Estimativas dos efeitos para o modelo que permite diferentes médias de glicoseao longo do tempo.

Efeito	Estimativa	Erro Padrão
Intercepto	103.80	0.3177
Jejum (0)	-22.0598	0.3190
1 hora (1)	17.9599	0.3190
2 horas (2)	0	-

Logo, o modelo estimado tem sua equação apresentada em (50)

$$\hat{\text{Glicose}}_{ti} = 103.80 + (-22.0598 * d_{1ti}) + (17.9599 * d_{2ti}) + U_{ti} + R_{ti} \qquad (50)$$

As estimativas para a variância do intercepto e para a variância do erro, são respectivamente iguais a $\hat{\sigma}_U^2 = 248.75$ e $\hat{\sigma}^2 = 252.49$. O coeficiente de correlação intraclasse para este modelo indica que 49.63% da variação dos níveis de glicose é devida a variação entre as pacientes .

menor deviance
melhor modelo

A estatística de *deviance* para o primeiro modelo (49) é igual a 141032.8 e para o segundo (50) é igual a 131579.2. Comparado (50) com (49) verifica-se que a diferença é significativa, pois a diferença desses valores segue aproximadamente uma distribuição Qui-Quadrado com 1 grau de liberdade (141032.8 - 131579.2 = 9453.6) e esta se apresenta muito superior ao valor tabelado ($p<0.0001$). Isto reforça o pensamento de que utilizar um modelo de coeficientes aleatórios, que permita a variação das médias ao longo do tempo, é mais adequado nesta situação.

As médias registradas em cada uma das ocasiões de medida (apresentadas na Tabela 3), sugerem que há diferenças significativas (Tabela 4) entre o nível de glicose dos tempos avaliados, quando comparados dois a dois (p-value<0.0001).

Tabela 3. Médias de glicose em cada ocasião de medida (Tempo).

Tempo	Média	Erro Padrão
Jejum (0)	81.7369	0.3177
1 hora (1)	121.76	0.3177
2 horas (2)	103.80	0.3177

Tabela 4. Diferenças entre as médias.

Tempo		Diferença Média	Erro Padrão	Pr> |t|
0	**1**	-40.0197	0.3190	<0.0001
0	**2**	-22.0598	0.3190	<0.0001
1	**2**	17.9599	0.3190	<0.0001

Os resultados das médias estimadas na Tabela 4, sugerem uma tendência quadrática para o nível médio de glicose e com isso não se justificaria ajustar uma tendência linear e utilizar uma inclinação aleatória.

Para minimizar a variação entre pacientes pretende-se, então, adicionar covariáveis ao nível de paciente que expliquem parte dessa variabilidade. Na etapa de inclusão das

demais variáveis explicativas que auxiliarão na explicação da variabilidade entre pacientes expressa pelo coeficiente de correlação intra-classe (49.63%), procedeu-se uma análise univariada para cada uma das covariáveis de interesse e em seguida ajustou-se o modelo global.

5.4.3. Altura das gestantes

A altura média das pacientes ficou em torno de 155.66 cm, com desvio de 6.48 cm. Os centros de Manaus e Fortaleza foram aqueles que apresentaram menor estatura média das gestantes, 151.63 cm e 152.56 cm respectivamente. As gestantes dos demais centros possuem altura média de aproximadamente 156 cm, exceto São Paulo que registrou altura média de 158cm.

Conforme relatado por Branchtein, et. al. (2000), mulheres com menos de 151cm possuem um aumento de 60% no risco relativo de ter Diabetes de Mellitus Gestacional, independente da clínica pré natal, idade, obesidade global, histórico familiar de diabetes entre outras características.

Dessa forma, optou-se por estratificar as mulheres em faixas de alturas determinadas pelos quartis das alturas das gestantes entrevistadas. A primeira faixa inclui todas as mulheres com altura até 151 cm. A segunda faixa contempla as pacientes cuja altura vai de 151.1 cm a 155.4 cm. Na terceira faixa temos as pacientes que possuem entre 155.5 cm e 159.8 cm e por fim na quarta e última faixa estão as mulheres que possuem acima de 160 cm.

Os resultados para os efeitos fixos considerando o efeito de altura e o da interação Tempo*Altura incorporados no modelo são apresentados na Tabela 5.

Tabela 5. Testes para os efeitos fixos do modelo.

Efeito	F	Pr>|t|
Tempo	7941.77	<0.0001
Altura	11.76	<0.0001
Tempo * Altura	10.98	<0.0001

As Médias de mínimos quadrados estimadas em função do tempo e da altura das gestantes são apresentadas na Tabela 6.

Tabela 6. Estimativas das Médias.

Efeito	Altura	Tempo	Estimativa	Erro Padão	Pr> |t|
Altura*Tempo	1	0	81.8744	0.6335	<.0001
Altura*Tempo	1	1	124.35	0.6335	<.0001
Altura*Tempo	1	2	107.28	0.6335	<.0001
Altura*Tempo	2	0	81.6794	0.6360	<.0001
Altura*Tempo	2	1	122.40	0.6360	<.0001
Altura*Tempo	2	2	105.11	0.6360	<.0001
Altura*Tempo	3	0	81.9788	0.6504	<.0001
Altura*Tempo	3	1	121.20	0.6504	<.0001
Altura*Tempo	3	2	102.41	0.6504	<.0001
Altura*Tempo	4	0	81.4445	0.6154	<.0001
Altura*Tempo	4	1	119.21	0.6154	<.0001
Altura*Tempo	4	2	100.52	0.6154	<.0001

As estimativas para a variância do intercepto e para a variância do erro, são respectivamente iguais a $\hat{\sigma}_U^2$ =247.10 e $\hat{\sigma}^2$ = 251.28. Para este modelo o teste de deviance estatística (141032.8 – 131467.9 = 9564.9; p-value<0.0001) sugere que a variável altura e sua interação com o tempo contribuem significativamente na explicação da variabilidade do nível de glicose em gestantes. A tendência, muito embora com intensidade levemente diferenciada para cada altura, evidencia-se quadrática para todos os níveis de altura.

5.4.4. Índice de Massa corporal antes da gravidez (BMIa).

O Índice de massa corporal é calculado dividindo o peso da paciente dividido pela altura ao quadrado e multiplicando o resultado por cem e sua unidade de medida é expressa em Kg/m^2 . Em média, as pacientes apresentam um índice de massa corporal antes da gravidez de 23.39 Kg/m^2, com desvio padrão de 4.05 Kg/m^2

A Tabela 7 apresenta os resultados para os testes de efeitos fixos do modelo acrescentado a variável Índice de massa corporal no modelo anterior.

Tabela 7. Testes para os efeitos fixos incluindo o Índice de Massa Corporal.

Efeito	F	Pr>|t|
Tempo	7596.94	<0.0001
Altura	10.36	<0.0001
Tempo * Altura	10.00	<0.0001
BMIA	188.15	<0.0001

A variância estimada para o intercepto foi de 232.59 e para o resíduo, 252.10. Comparado ao modelo anterior através dos valores das estatísticas de deviance temos que o modelo após acrescentar a covariável BMIa apresenta-se significativo, pois a diferença desses valores segue aproximadamente uma distribuição Qui-Quadrado com 1 grau de liberdade (131467.9 - 126024.8= 8443.1) e esta se apresenta muito superior ao valor tabelado ($p<0.0001$). Assim, a variável BMIa será considerada no modelo final como uma covariável que vai atuar ao nível de gestante a fim de auxiliar na explicação da variabilidade entre os níveis de glicose das pacientes.

5.4.5. Idade da Gestante (IDADE)

A idade gestacional média das pacientes deste estudo ficou em torno de 27.85 anos com desvio padrão de 5.46 anos.

Os testes de efeitos fixos deste modelo são apresentados na Tabela 8 e indicam que há significância na inclusão da covariável Idade da Gestante no modelo.

Tabela 8. Efeitos Fixos para o modelo considerando a covariável Idade

Efeito	F	Pr>|t|
Tempo	7941.77	<0.0001
Altura	12.54	<0.0001
Tempo*Altura	10.98	<0.0001
Idade	213.23	<0.0001

A variância estimada para o termo aleatório do intercepto foi de 233.54 e para o termo de erro do modelo foi de 251.28. Este modelo apresentou-se significativo, uma vez que a diferença entre as deviances dos modelos que incluem ou não a covariável idade apresentam-se significativamente diferentes (131467.9 - 131263.4 = 204.5) comparada ao valor de uma estatística qui-quadrado com 1 GL mostrou-se significativa ($p<0.0001$). Logo, a variável idade também participará do modelo final como uma covariável.

5.4.6. Histórico de Diabetes na Família (HFDM2)

Em 14.95% dos casos foi registrado a presença de histórico de diabetes na família da gestante, e partiu-se então para a análise desta informação como uma covariável ao nível de paciente. A tabela 9 apresenta os resultados dos testes de efeitos fixos para o modelo que considera a covariável Histórico de Diabetes na Família.

Tabela 9. Testes para os efeitos fixos do modelo considerando a covariável HFDM2.

Efeito	F	Pr>|t|
Tempo	7416.66	<.0001
Altura	11.50	<.0001
Altura* Tempo	10.02	<.0001
HFDM2	52.55	<.0001

As estimativas para os parâmetros de variância de U e R são respectivamente 246.26 e 253.89. O valor da deviance para este modelo foi de 123860.1 e quando comparado a deviance do modelo que não inclui nenhuma covariável, exceto a altura, temos como significativa ($p<0.0001$) a inclusão da variável HFDM2 no modelo (131467.9 - 123860.1=7607.8) .

5.4.7. Número de gravidezes anteriores (GRAVICAT)

O número de gravidezes anteriores variou de zero a quinze sendo que a maioria das pacientes nunca haviam engravidado anteriormente (27,27%).

Optou-se por classificar esta variável em 4 categorias, determinadas pelos quartis. A primeira contém todas as pacientes que nunca tiveram nenhum filho, a segunda é representada por aquelas que já apresentaram uma gravidez anteriormente, a terceira faixa é representada pelas gestantes que já tiveram duas gestações e a quarta e última classe está representada pelas mulheres que já tiveram 3 ou mais gestações anteriormente.

Os testes para os efeitos fixos do modelo incluindo a covariável Número de Gestações anteriores são apresentados na Tabela 10. Os resultados sugerem que há fracas evidências do efeito do número de gravidezes anteriores no nível de glicose das gestantes (p-value = 0.0961).

Tabela 10. Testes para os efeitos fixos do modelo considerando a covariável GRAVICAT.

Efeito	F	Pr>|t|
Tempo	7416.66	<.0001
Altura	11.50	<.0001
Altura* Tempo	10.02	<.0001
GRAVICAT	2.12	0.0961

Não fez deviance?

5.4.8. Temperatura Ambiental (TEMP5F)

Devida a grande variação da temperatura entre os centros optou-se por trabalhar as temperaturas divididas em faixas, de amplitude 5°C, conforme a Tabela 11.

Tabela 11. Faixas de temperaturas utilizadas na análise.

Temperatura	Faixa
Até 15°C	1
15°C a 20°C	2
20°C a 25°C	3
25°C a 30°C	4
Acima de 30°C	5

As estimativas dos testes de efeitos fixos do modelo apresentam significância na inclusão desta covariável no modelo, conforme os resultados da Tabela 12.

Tabela 12. Testes para os efeitos fixos do modelo considerando a covariável TEMP5F.

Efeito	F	Pr>|t|
Tempo	6774.03	<.0001
Altura	2.40	0.0655
Altura* Tempo	8.74	<.0001
TEMP5F	38.28	<.0001

não tem problema, é menor que 0,1

O valor da deviance para este modelo mostrou-se inferior aos modelos anteriormente ajustados, 117470.2, e quando comparado ao modelo que considera somente a altura, este apresenta-se altamente significativo (p-value<0.0001).

5.4.9. Modelo final

Após a análise considerando a inclusão de uma covariável de cada vez, foi ajustado um modelo incluindo todas aquelas que individualmente foram significativas (p-value<0.1).

Fazem parte do modelo então além da Altura, o Índice de Massa Corporal, anterior a Gestação, a Idade da Gestante, o Histórico de Diabetes na Família, a Temperatura Ambiental, e o Número de Gestações Anteriores.

Os testes para os efeitos fixos do modelo final são apresentados na Tabela 13 e as estimativas das médias de mínimos quadrados são apresentadas na Tabela 14.

Tabela 13. Testes para os efeitos fixos do Modelo final.

Efeito	F	Pr>|t|
Tempo	6076.32	<.0001
Altura	2.88	0.0348
Altura* Tempo	8.03	<.0001
BMIA	112.78	<.0001
IDADE	156.02	<.0001
HFDM2	22.81	<.0001
TEMP5F	52.11	<.0001
GRAVICAT	5.92	0.0005

Tabela 14 Estimativas das Médias.

Efeito	Altura	Tempo	Estimativa	Erro Padrão
Altura*Tempo	1	0	81.7608	0.8145
Altura*Tempo	1	1	124.02	0.8145
Altura*Tempo	1	2	107.31	0.8145
Altura*Tempo	2	0	82.1468	0.8097
Altura*Tempo	2	1	123.17	0.8097
Altura*Tempo	2	2	106.63	0.8097
Altura*Tempo	3	0	83.1797	0.8198
Altura*Tempo	3	1	122.70	0.8198
Altura*Tempo	3	2	104.39	0.8198
Altura*Tempo	4	0	83.1711	0.7748
Altura*Tempo	4	1	120.69	0.7748
Altura*Tempo	4	2	102.84	0.7748

A variância estimada para o intercepto é igual a $\hat{\sigma}_U^2$ =210.60 e a variância estimada para o erro é $\hat{\sigma}^2$ =267.00. A correlação intra-classe para o modelo final é de 0.44095, ou seja, excluindo a influência das covariáveis utilizadas, aproximadamente 44.1% da variabilidade total dos dados é devida a variação entre pacientes.

A tendência quadrática das médias prevalece para todos as categorias de altura, muito embora nas categorias de altura mais baixa o patamar de curvatura é mais acentuado.

O Gráfico dos Resíduos do modelo final não apresenta nenhum padrão definido, sugerindo a adequação do modelo.

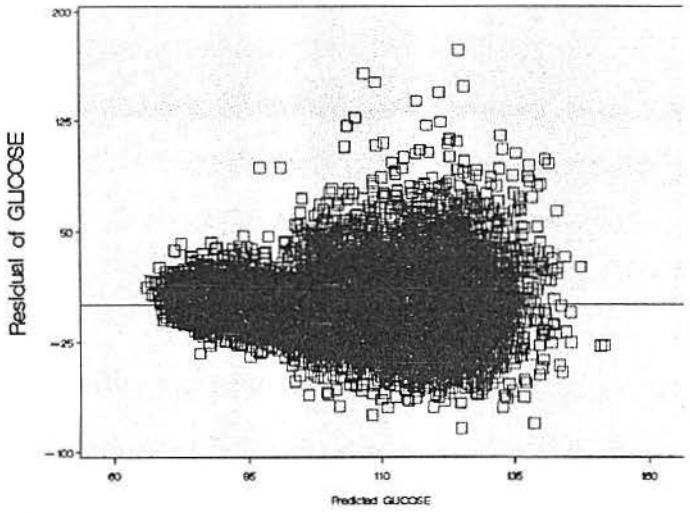

Figura 10. Resíduos versus preditos pelo modelo final.

6. CONCLUSÕES

O presente trabalho teve por objetivo caracterizar a metodologia de modelos multiníveis e ilustrar um enfoque das inúmeras aplicações.

A Análise Multinível é um procedimento sofisticado de modelagem estatística, sendo que sua utilização e divulgação encontra-se em plena expansão. Atualmente a maioria das aplicações econcentram-se em áreas de pesquisas educacionais, no entanto é possível utilizar esta metodologia nas mais diferentes áreas de investigação científica, e em especial na medicina e na epidemiologia, onde estudos multicêntricos e a meta-análise são muito utilizados e reconhecidos.

A modelagem multinível possibilita ajustar de forma mais realista a variação presente nestes dados, nos diferentes níveis de hierarquia, o que propicia análises mais fidedignas, ao não assumir erroneamente o pressuposto de independência entre as unidades, pertencentes a um nível maior de hierarquia. Logo, são obtidas estimativas mais eficientes dos coeficientes de regressão e de seus erros padrões, sendo possível também, maior flexibilidade na estrutura de covariâncias, permitindo a análise de dados com heterogeneidade de variâncias e presença de correlações. Além disso, permite ainda quantificar a variabilidade da variável resposta em cada um dos níveis de hierarquia presentes no estudo, de tal forma que a proporção da variabilidade possa ser comparada diretamente entre os níveis.

Na aplicação apresentada foi utilizado a modelagem multinível em um estudo epidemiológico multicêntrico com estrutura de medidas repetidas. Dos resultados obtidos pode-se afirmar que a modelagem multinível mostrou ser um procedimento eficiente e útil em estudos desta natureza.

A teoria subjacente a Modelos Multiníveis é relativamente complexa e em decorrência disto, são poucos os aplicativos computacionais disponíveis para tais análises, e consequentemente poucos são os que dominam e aplicam esta técnica. No entanto, com o desenvolvimento computacional, novos aplicativos estão sendo disponibilizados para auxiliarem na aplicação da modelagem multinível

proporcionando assim uma maior utilização desta metodologia, propiciando a necessária maior divulgação.

Acredita-se que este trabalho é consistente, e pode servir como um referencial introdutório, tanto sob o prisma teórico quanto aplicado, para os interessados na promissora e importante área de investigação científica que constituem os modelos multiníveis.

7. BIBLIOGRAFIA

BRANCHTEIN, L. et. Al. Short Stature and gestacional diabetes in Brasil. **Diabetologia.** Springer- Verlag, 43: 848-851. 2000.

BRYK, A. S. and RAUDENBUSH, S. W. Methodology for cross-level organizational research. **In: Research in Organizational Behavior**.1988.

BRYK, A. S. and RAUDENBUSH, S. W. **Hierarchical Linear Models: Applications and Data Analysis Methods**. Sage Publications, Newbury Park, CA.1992.

CARBALLO, M.T. **Predição da Macrossomia Fetal Através da Regressão Logística e de Redes Neurais Artificiais**. Porto Alegre, 107 p. (Monografia do curso de Bacharelado em Estatística, UFRGS) 2002.

DEMÉTRIO, C. G. B. Modelos Lineares Generalizados em Experimentação Agronômica. **46ªReunião Anual da RBRAS e 9°SEAGRO.** Piracicaba, SP, 2001.

GOLDSTEIN, H. **Multilevel Statistical Models.** 2ª edição. Edward Arnold. Londres. 1995.

HILDEN-MINTON, J. A. **Multilevel diagnostics for mixed and Linear Hierarquical Models**. University of California. Los Angeles. 1995

HOX, J. J. **Applied Multilevel Analysis**. Amsterdam: TT-Publickaties. LL. 1995

JENNRICH, R. and SCHLUCHTER, M. Unbalanced Repeated Measure Models With Structured Covariance Matrices. **Biometrics, 42**, 805-820. 1986.

KREFT, I. G.; de LEEUW, J. & KIM, K. **Comparing four different statistical packages for hierarchical linear regresion: Genmod, HLM, ML2, and VARCL**. Statistical Series n° 50. Los Angeles: University of California at Los Angeles. 1990.

KREFT, I. G.; de LEEUW, J. & der LEEDEN , R. V. Review of Five Multilevel Analysis Programs: BMDP-5V, GENMOD, HLM, ML3, VARCL. **The American Statistician**, 48, nº 4. 1994.

KREFT, I.& LEEUW, J.**Introducing Multilevel Modeling**.SAGE Publication. First Pulication. 1998.

LAPLANTE, B. **An introduction to the use of linear models with correlated data**. London Ontario.1999.

LITTLE, R. J. A. e RUBIN, D. B. **Statistical Analysis with Missing Data.** New York. Willey. 1987.

LONGFORD, N. T. **VARCL. Software for Variance Component Analysis of Data with Nested Random Effets (Maximum Likelihood).** Educational Testing Service, Princeton, NJ. 1990.

MASON, W. W., WONG, G.Y. and ENTWISLE, B. Contextual Analysis Trough the Multilevel Linear Model. **Sociology Methodology**, 72,-103.1984.

NELDER, J. A.; WEDDERBURN, R. W. M. Generalized LinearModels. **Journal of the Royal Statistical Society A, 135,** 3, p.370-84, 1972.

RASBASH, J. BROWNE, et. al . A User´s Guide to MlwiN. Version 2.1c. .Center for multilevel Model. Institute of Education. 2000.

SNIJDERS, T. A. B. & BOSKER, R. J. **Multilevel Analysis: An introduction to basic and advanced multilevel modeling.** SAGE Publications. L.1999.

8. ANEXOS

As análises foram feitas utilizando o aplicativo ANALYST e o procedimento PROC MIXED do Software Estatístico SAS, Versão 8.2. A linguagem de programação utilizada para o ajuste dos modelos é apresentada a seguir.

8.1. Modelo de intercepto aleatório para centros

Considerando um intercepto aleatório para cada um dos centros participantes do estudo, o modelo ajustado é apresentado a seguir:

```
/*modelo com intecepto aleatorio*/
proc mixed  data=final.sub  METHOD=REML CL ALPHA=.05 COVTEST;
class registro centroa;
model glicose=/ HTYPE=3 DDFM=SATTERTH
SOLUTION CL ALPHA=.05;
random intercept/ subject= centroa;
run;
```

8.2. Modelo de intercepto aleatório para pacientes

Uma vez especificado o intercepto como efeito aleatório, no comando **random,** uma equação para cada uma das gestantes é gerada. No programa a seguir considera-se que as médias das medidas são todas iguais.

```
proc mixed  data=banco_fi.banco_final_  METHOD=REML CL ALPHA=.05
COVTEST;
class registro;
model glicose=/ HTYPE=3 DDFM=SATTERTH
SOLUTION CL ALPHA=.05;
random intercept/subject=registro;
RUN;
```

8.3. Modelo de intercepto aleatório para pacientes considerando efeito fixo de tempo

Considerando que pode haver diferenças entre as médias de cada ocasião de medida, ou seja, considerando o tempo como variável classificatória, ainda dentro da concepção de modelo de coeficientes aleatórios, temos o seguinte programa:

```
proc mixed data=banco_fi.banco_final_ METHOD=REML CL ALPHA=.05 COVTEST;
class registro tempocat;
model glicose= tempocat/ HTYPE=3 DDFM=SATTERTH
SOLUTION CL ALPHA=.05;
random intercept/subject=registro;
lsmeans tempocat/pdiff CL ALPHA=.05;
run;
```

8.4. Modelo de intercepto aleatório para pacientes considerando os efeitos fixos de tempo, altura e interação entre eles

```
/*modelo com intecepto aleatorio*/
proc mixed   data=FINAL.banco_final_  METHOD=REML CL ALPHA=.05 COVTEST;
class registro altcat tempocat ;
model glicose= tempocat altcat tempocat*altcat / HTYPE=3 DDFM=SATTERTH
SOLUTION CL ALPHA=.05 ;
random intercept/ subject= registro;
lsmeans tempocat*altcat/pdiff CL ALPHA=.05;
run;
```

8.5. Modelo de intercepto aleatório para pacientes considerando os efeitos fixos de tempo, altura e interação entre eles incluindo as covariáveis

Para a inclusão da covariável Índice de Massa Corporal anterior a gestação, a programação utilizada está a seguir sendo semelhante para as demais covariáveis.

```
/*modelo com intecepto aleatorio*/
proc mixed   data=FINAL.banco_final_  METHOD=REML CL ALPHA=.05 COVTEST;
class registro altcat tempocat ;
model glicose= tempocat altcat tempocat*altcat bmia / HTYPE=3
DDFM=SATTERTH
SOLUTION CL ALPHA=.05 ;
random intercept/ subject= registro;
lsmeans tempocat*altcat/pdiff CL ALPHA=.05;
run;
```

8.6. Modelo Final

O programa utilizado para a estimação do modelo final foi:

```
/*modelo com intecepto aleatorio*/
```

```
proc mixed   data=FINAL.banco_final_  METHOD=REML CL ALPHA=.05 COVTEST;
class registro altcat tempocat hfdm2 temp5f gravicat ;
model glicose= tempocat altcat tempocat*altcat bmia idade hfdm2 temp5f
gravicat/ HTYPE=3 DDFM=SATTERTH
SOLUTION CL ALPHA=.05 OUTPM=work._prdmn;
random intercept/ subject= registro;
lsmeans tempocat*altcat/pdiff CL ALPHA=.05;
run;
```

8.7. Programação para o gráfico

```
OPTIONS NOTATE NONUMBER;

/*modelo com intecepto aleatorio*/
proc mixed   data=FINAL.banco_final_  METHOD=REML CL ALPHA=.05 COVTEST;
class registro altcat tempocat hfdm2 temp5f gravicat ;
model glicose= tempocat altcat tempocat*altcat bmia idade hfdm2 temp5f
gravicat/ HTYPE=3 DDFM=SATTERTH
SOLUTION CL ALPHA=.05 OUTPM=work._prdmn;
random intercept/ subject= registro;
lsmeans tempocat*altcat/pdiff CL ALPHA=.05;
run;

*** Plots ***;
goptions ftext=SWISS ctext=BLACK htext=1 cells
         gunit=pct htitle=6;
axis1 major=(number=5) label=(a=90 h=4) width=1;
axis2 offset=(10 pct) label=(h=4) width=1;
axis3 major=(number=5) offset=(5 pct) width=1;
** Residual Plots **;
goptions reset=symbol;
symbol1 color=BLUE height=4 value=SQUARE;

proc gplot data=work._prdmn ;
   label RESID = "Residual of GLICOSE";
   label PRED = "Predicted GLICOSE";
   plot RESID * PRED /
      vaxis=axis1 vminor=0 haxis=axis3 hminor=0 vref=0
      cframe=CXF7E1C2 caxis=BLACK
      description = "Residual by predicted GLICOSE" name="RESID";
run;
quit;
goptions reset=symbol ftext= ctext= htext=;
axis1; axis2; axis3;
goptions reset=all device=WIN     ;
proc delete data=work._prdmn;
run;
```