INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA – IBGE
ESCOLA NACIONAL DE CIÊNCIAS ESTATÍSTICAS – ENCE

Modelos multiníveis em pesquisas amostrais complexas – uma aplicação à valoração de aluguéis de imóveis residenciais segundo suas características/atributos

JAILSON MANGUEIRA ASSIS

RIO DE JANEIRO
Outubro de 2005

Modelos multiníveis em pesquisas amostrais complexas – uma aplicação à valoração de aluguéis de imóveis residenciais segundo suas características/atributos

JAILSON MANGUEIRA ASSIS

Dissertação apresentada à Escola Nacional de Ciências Estatísticas para obtenção do título de Mestre em Estudos Populacionais e Pesquisas Sociais.

Área de Concentração: Estatística Social
Orientador: Prof. Dr. Pedro Luis do Nascimento Silva
Co-orientador: Prof. Dr. Fernando Antônio da Silva Moura

RIO DE JANEIRO
Outubro de 2005

A848m ASSIS, Jailson Mangueira
Modelos multiníveis em pesquisas amostrais complexas – uma aplicação à valoração de aluguéis de imóveis residenciais segundo suas características/ atributos / Jailson Mangueira Assis
Rio de Janeiro : J. M. Assis, 2005.
114 fl. : il.
Orientador: Pedro Luis do Nascimento Silva
Dissertação (Mestrado) - Escola Nacional de Ciências Estatísticas. Programa de Pós-Graduação em Estudos Populacionais e Pesquisas Sociais.
Inclui bibliografia.
1. Modelos lineares (estatística). 2. Probabilidades. 3. Simulação - Métodos. 4. Pesquisa sobre Padrões de Vida. I. SILVA, Pedro Luis do Nascimento. II. Escola Nacional de Ciências Estatísticas. Pós-Graduação em Estudos Populacionais e Pesquisas Sociais. III.Título.

CDU 519.22

À minha mãe Josefa e à Marcela.

AGRADECIMENTOS

Durante o período de elaboração desta dissertação tive o apoio e incentivo de diversas pessoas. Mesmo correndo o risco de omitir o nome de algumas destas, seria uma falha ainda maior não citá-las nominalmente. Aqueles nomes, que por ventura tenham sido omitidos, recebem também meus sinceros agradecimentos.

Primeiramente, agradeço ao meu orientador, o Professor Pedro Luis do Nascimento Silva, o qual mereceu minha admiração e respeito desde o primeiro curso de Estatística Básica. Sem seus conhecimentos, paciência e bom senso não seria possível concluir com êxito esta dissertação.

Agradeço ao meu co-orientador Professor Fernando Moura, que me proporcionou os primeiros conhecimentos em Inferência Bayesiana e técnicas de simulação estocástica.

A Escola Nacional de Ciências Estatísticas (ENCE) e alguns de seus professores que tive a felicidade de conhecer no período do meu curso também foram de fundamental importância. Com eles obtive conhecimentos e experiência para realizar este projeto, aprendendo técnicas mais avançadas e métodos do pensamento científico voltados para pesquisa.

Agradeço em especial à Elisa Lustosa Caillaux, que foi gerente da Pesquisa sobre Padrões de Vida (PPV) do IBGE e prontamente me forneceu a documentação e os dados da pesquisa para que eu pudesse realizar a aplicação. Atualmente tenho a sorte de trabalhar em sua equipe e sou grato por seu constante apoio e incentivo. Lembro também aqui os colegas da Coordenação de População e Indicadores Sociais que torceram por ver este trabalho concluído.

Aos colegas da Coordenação de Métodos e Qualidade do IBGE, Marcos Paulo, Sônia e Vermelho que me ajudaram a obter as bases de dados da PPV e do Censo Demográfico de 1991.

Em relação aos métodos computacionais, agradeço às colegas Solange Corrêa e Viviane Quintaes. Solange me cedeu o programa em SAS que calcula estimativas pelo procedimento de MQGIPP. Viviane foi quem me deu as primeiras dicas sobre o *software* WinBUGS.

Aos amigos da minha turma de mestrado Humberto, Juliana, Kleber, Luciano, Maurício, Rita e todos aqueles que tornaram nossa convivência neste período algo que deve ser lembrado com alegria.

Finalmente agradeço a minha família da qual sempre recebi conforto nos momentos difíceis e apoio incondicional, compreendendo minha ausência durante esta empreitada.

RESUMO

Dados provenientes de pesquisas domiciliares geralmente possuem algum tipo de estrutura hierárquica, como pessoas agrupadas em domicílios, domicílios agrupados em setores censitários e assim por diante. Tais dados são freqüentemente obtidos por meio de pesquisas com esquemas amostrais complexos, o que pode tornar o plano amostral informativo ou não ignorável. Quando isto ocorre, não considerar o planejamento amostral no processo de modelagem pode resultar em inferências viesadas e análises equivocadas dos dados.

Nesta dissertação são comparados três métodos de ajuste de modelos lineares hierárquicos de dois níveis para os dados da Pesquisa sobre Padrões de Vida – PPV – realizada pelo Instituto Brasileiro de Geografia e Estatística – IBGE. O primeiro método não considera as informações do processo de amostragem no ajuste do modelo. O segundo método emprega o procedimento de ponderação pelas probabilidades de inclusão na amostra, proposto por Pfeffermann *et al*. (1998a). O terceiro método, envolve o ajuste de um modelo de distribuição amostral, obtido em função do modelo populacional e das probabilidades de inclusão de primeira ordem. Este ajuste é implementado com a aplicação do paradigma Bayesiano pelo uso de simulação via Monte Carlo com Cadeias de Markov (MCMC), conforme proposto por Pfeffermann, Moura e Silva (2001). Uma aplicação destes três métodos é feita para modelar a relação entre o valor mensal estimado de aluguel do domicílio com vários de seus atributos, incluindo acesso a serviços e considerando um possível efeito de grupo do setor censitário onde o domicílio se localiza.

Palavras-chave: Modelos hierárquicos, Amostragem informativa, Ponderação pelas probabilidades, Simulação via MCMC.

ABSTRACT

Frequently there are hierarchical structures in data collected through household surveys, such as people clustered in households, households clustered in census tracts, and so on. These surveys are generally based on complex sampling, often leading to informative or non-ignorable sample designs. In such cases, modelling data ignoring the sample design may yield biased inferences and distort the analysis.

The main objective of this work is to compare tree approaches to fit hierarchical linear models of two levels using data from the Brazilian Living Standards Measurement Survey. The first method fits the model ignoring the sample design. In the second approach, the model is fitted using the probability weighting procedure proposed by Pfeffermann *et al.* (1998a). On the third method, a sample distribution model is fitted for the sample data as a function of a population model and the first order sample inclusion probabilities. In order to implement this method, the full Bayesian paradigm through Markov Chain Monte Carlo (MCMC) simulations was used, as discussed by Pfeffermann, Moura and Silva (2001). An application of these three methods was made to model the relationship between the estimate of household's rental value and its characteristics and access to services, considering a possible group effect of the census tract where the household is located.

Key words: Hierarchical linear models, Informative sampling, Probability weighting, Simulation by MCMC.

SUMÁRIO

LISTA DE TABELAS

LISTA DE QUADROS

LISTA DE FIGURAS

LISTA DE EXEMPLOS

1 – INTRODUÇÃO

Em diversas populações é comum que os dados investigados possuam uma estrutura de agrupamento ou hierarquia. Até o final de década de 80, grande parte dos modelos ajustados não levava em conta esta estrutura. Em parte, isto devia-se à falta de métodos e *softwares* que viabilizassem o tratamento destes dados, forçando os analistas a escolherem a unidade sobre a qual o seu estudo iria incidir. Caso fosse escolhido o nível 2, por exemplo o desempenho de escolas, as medidas coletadas dos alunos (unidades de nível 1) seriam agregadas por médias ou proporções de forma a se ter uma medida para cada escola. Este procedimento implica em perda de informação referente à variabilidade intra-escola. Por outro lado, caso o analista optasse por analisar os dados no nível de alunos, as informações referentes às escolas seriam repetidas para cada aluno da escola, resultando em estimativas de erro padrão das estimativas dos parâmetros incorretas, interferindo na inferência a ser realizada. Modelos mais adequados para a análise de dados com estes padrões de variabilidade complexos são os chamados modelos hierárquicos, também conhecidos como modelos multiníveis.

Dados provenientes de pesquisas socioeconômicas ou da área médica freqüentemente possuem algum tipo de estrutura hierárquica. Quando estas pesquisas são realizadas por meio de levantamento amostral com desenho não ignorável, mais uma fonte de variação é introduzida no processo. Nestes casos os procedimentos utilizados para ajustar os modelos hierárquicos usuais não são mais adequados. Pfeffermann *et al.* (1998a) desenvolveram um método que considera as probabilidades de inclusão na amostra para corrigir as estimativas dos parâmetros. Este método é uma adaptação do método de Mínimos Quadrados Generalizados Iterativo.

Quando as probabilidades de seleção da amostra são relacionadas com o valor da variável resposta, mesmo após condicionar nas covariáveis do modelo, o processo amostral se

torna informativo e o modelo para os dados da amostra é diferente do modelo para a população. Pfeffermann, Moura e Silva (2001) apresentam uma abordagem que consiste em derivar o modelo da distribuição amostral com base na distribuição populacional correspondente e nas probabilidades de seleção da amostra. O ajuste para este método foi feito através do paradigma Bayesiano com uso da técnica de Monte Carlo com Cadeias de Markov – MCMC (do inglês *Markov Chain Monte Carlo*).

Com o objetivo de comparar os resultados das diferentes abordagens na estimação de modelos hierárquicos, foi feita uma aplicação com dados da Pesquisa sobre Padrões de Vida – PPV do IBGE para um modelo hierárquico de dois níveis. A aplicação consistiu em modelar o valor do aluguel de imóveis residenciais em função de suas características físicas e de sua localização. Aqui a hierarquia nos dados se baseia na hipótese de que imóveis localizados em um mesmo setor censitário tendem a ter aluguéis mais semelhantes do que imóveis em outros setores censitários, mesmo depois de controladas as características do imóvel. Desta forma, as unidades de nível 1 seriam formadas pelos domicílios e as unidades de nível 2 pelos setores censitários.

Foram utilizadas variáveis explicativas para os dois níveis do modelo. No nível 1 foram consideradas todas as questões que constavam na PPV referentes à estrutura, dimensão, e acesso a serviços dos domicílios. Para o nível de setor censitário, foram utilizadas algumas medidas resumo calculadas a partir do Censo Demográfico de 1991. A inclusão destas variáveis procurou principalmente descrever o efeito da localização do imóvel.

Esta dissertação está organizada em seis capítulos. No capítulo 2 são apresentados e discutidos os dados utilizados na aplicação e a descrição do problema aqui estudado. Além disto, o plano amostral da Pesquisa sobre Padrões de Vida é apresentado de forma detalhada.

Nos capítulos 3 e 4 são discutidas as metodologias das diferentes abordagens utilizadas no ajuste dos modelos hierárquicos. No capítulo 3 são introduzidos o modelo

hierárquico e o método de estimação por Mínimos Quadrados Generalizados Iterativo (MQGI). No capítulo 4 são apresentados os métodos de Mínimos Quadrados Generalizados Iterativo Ponderado pelas Probabilidades (MQGIPP) e o do modelo da distribuição amostral que será ajustado via Monte Carlo com Cadeias de Markov (MCMC).

O capítulo 5 contém os resultados dos modelos ajustados, assim como uma análise sobre a comparação das diferentes abordagens.

No capítulo 6 são apresentadas as conclusões, recomendações de extensões e sugestões de trabalhos futuros, que poderão melhorar as estimativas de modelos hierárquicos ajustados com dados provenientes de amostras complexas.

2 – DESCRIÇÃO DO PROBLEMA E FONTE DE DADOS

2.1 – Descrição do Problema

O estudo do mercado habitacional brasileiro possui grande importância nas áreas econômica e social. O setor de moradias responde por grande parte do número de empregos da construção civil. Por outro lado, o contingente populacional sem acesso a moradia adequada ainda é muito vasto. O conhecimento das especificidades e da lógica do funcionamento do mercado habitacional é de fundamental importância para um desenho eficiente das políticas públicas, sendo particularmente útil na construção de índices de preços ou em estudos de distribuição e desigualdade do estoque de riqueza.

Um aspecto necessário para o entendimento deste mercado é o fato de que o bem habitação é marcadamente heterogêneo, no sentido de que seu preço é diferenciado segundo suas características de localização, acesso a serviços ou tipo de construção. Santos e Cruz (2000, p. 6) dizem que:

> Um enfoque possível para lidar com essa dificuldade consiste em tratar o bem *habitação* como um bem composto por um conjunto de outros bens (que seriam as características individuais de cada habitação, tais como sua localização, número de cômodos, amenidades na vizinhança, etc.), cujos preços implicitamente contribuem para a formação do preço de mercado de cada habitação. Esses preços são definidos como preços hedônicos ou preços implícitos das características do bem *habitação*.

Desta forma, o preço de mercado da habitação poderia ser decomposto no preço de suas características/atributos. O método do preço hedônico viabiliza, com base no preço de aluguel do imóvel, estimar o quanto cada característica/atributo contribui para o valor do aluguel do imóvel e então prever ou imputar um valor de aluguel para os imóveis que não são alugados, de acordo com suas características/atributos.

Reis, Tafner e Reiff (2001) sugerem em seu trabalho a utilização do método do preço hedônico para analisar a distribuição do estoque de riqueza habitacional entre as famílias brasileiras no período de 1992 a 1999. Os autores utilizaram os dados da Pesquisa Nacional por Amostra de Domicílios (PNAD) do IBGE para o referido período.

Nesta dissertação utiliza-se a abordagem do preço hedônico para estudar a relação do valor do aluguel mensal com os atributos físicos e de localização do imóvel, com base nos dados da Pesquisa sobre Padrões de Vida (PPV) do IBGE. Os modelos aqui ajustados consideraram as mesmas variáveis explicativas sugeridas no artigo de Reis, Tafner e Reiff (2001), acrescentando-se outras variáveis disponíveis na PPV.

2.2 – Descrição da Pesquisa sobre Padrões de Vida

A Pesquisa sobre Padrões de Vida (PPV) foi um projeto piloto realizado pelo IBGE em convênio com o Banco Mundial, tendo como objetivo principal fornecer informações sobre o bem-estar e condições de vida da população em diferentes grupos sociais. Desta forma, a pesquisa foi pensada para captar diversas dimensões, tais como: características sobre moradia, tendências demográficas, atividades econômicas, acesso aos serviços de saúde e educação, nutrição, antropometria e ainda uma avaliação opinativa, por parte do entrevistado, sobre as suas condições de vida.

Os temas investigados pela pesquisa, embora em grande número, não foram tratados com a mesma profundidade que em pesquisas mais específicas. No entanto, os resultados deveriam resumir de forma muldimensional o bem-estar dos entrevistados e permitir estudar as interações entre os diversos temas.

A pesquisa foi realizada por meio de um levantamento amostral domiciliar aplicado nas regiões Nordeste e Sudeste do país. Seu período de realização se deu entre março de 1996

e março de 1997, visando captar fenômenos sazonais. Os questionários foram aplicados em duas visitas ao mesmo domicílio num intervalo de duas semanas, buscando-se assim obter um maior controle da qualidade dos dados. Maiores detalhes da pesquisa podem ser encontrados em Caillaux (1996).

2.2.1 – Plano Amostral da Pesquisa sobre Padrões de Vida

O levantamento das informações da Pesquisa sobre Padrões de Vida (PPV) foi feito através da execução de um plano amostral probabilístico em dois estágios, com as unidades primárias de amostragem (setores) estratificadas e selecionadas com probabilidades proporcionais ao tamanho (PPT) e as unidades secundárias (domicílios) selecionadas com eqüiprobabilidade dentro de cada setor selecionado no primeiro estágio. A dimensão da amostra foi definida de acordo com a disponibilidade de recursos.

A pesquisa foi realizada nas Regiões Nordeste e Sudeste do país considerando 10 estratos geográficos, a saber: Região Metropolitana de Fortaleza, Região Metropolitana de Recife, Região Metropolitana de Salvador, restante da área urbana do Nordeste, restante da área rural do Nordeste, Região Metropolitana de Belo Horizonte, Região Metropolitana do Rio de Janeiro, Região Metropolitana de São Paulo, restante da área urbana do Sudeste e restante da área rural do Sudeste.

Foi utilizado um critério estatístico para definir uma estratificação adicional das unidades primárias de amostragem (setores), com base na renda média mensal por setor dos chefes de domicílio. Esta informação foi investigada no Censo Demográfico de 1991 para todos os domicílios. Foram criados três estratos de renda (Baixa, Média e Alta) para cada um dos dez estratos geográficos. A combinação entre os estratos geográficos e os estratos de renda resultou num total de 30 estratos.

Inicialmente o tamanho da amostra foi fixado em 480 domicílios por estrato geográfico. Nos estratos denominados "restante da área rural do Nordeste" e "restante da área rural do Sudeste", foi estabelecido que seriam pesquisados 30 setores e 16 domicílios em cada setor. Para cada um dos estratos geográficos urbanos e regiões metropolitanas foram selecionados 60 setores e 8 domicílios por setor. Tal diferença na distribuição dos tamanhos de amostra de 1º e 2º estágios se deveu à dificuldade de acesso, e conseqüente aumento de custo para pesquisar setores localizados em áreas rurais. O tamanho total da amostra foi inicialmente fixado em 540 setores censitários e 4.800 domicílios.

A distribuição da amostra de setores entre os estratos de renda de cada estrato geográfico foi feita por alocação proporcional, tendo como base o total de domicílios particulares permanentes ocupados em cada estrato, isto é:

$$n_r = n.\frac{M_r}{M} \tag{2.1}$$

onde

n_r: número de setores a serem selecionados no estrato de renda *r*, em determinado estrato geográfico;

M_r: número de domicílios particulares permanentes ocupados na população do estrato de renda *r*, obtido pelo Censo Demográfico de 1991;

n: número total de setores a serem selecionados no estrato geográfico especificado;

M: número total de domicílios particulares permanentes ocupados no estrato geográfico que contém o estrato de renda *r*.

Após a alocação da amostra nos estratos de renda, o tamanho da amostra final ficou em 554 setores censitários e 4.940 domicílios, devido à aplicação de procedimentos de

arredondamento e da condição de que houvesse ao menos uma amostra de dois setores dentro de cada estrato. A distribuição da amostra segundo os estratos pode ser vista na tabela 2.1.

Tabela 2.1 Distribuição da amostra por estrato geográfico e estrato de renda

Estrato Geográfico	Estrato de Renda	Número de setores na amostra	Número de domicílios na amostra
Região Metropolitana de Fortaleza	Baixa	46	368
	Média	11	88
	Alta	5	40
Região Metropolitana de Recife	Baixa	45	358
	Média	9	72
	Alta	7	54
Região Metropolitana de Salvador	Baixa	46	368
	Média	9	72
	Alta	6	48
Restante do Nordeste Urbano	Baixa	47	376
	Média	11	88
	Alta	3	24
Restante do Nordeste Rural	Baixa	26	416
	Média	5	80
	Alta	2	32
Região Metropolitana de Belo Horizonte	Baixa	45	360
	Média	11	88
	Alta	6	48
Região Metropolitana do Rio de Janeiro	Baixa	44	352
	Média	10	80
	Alta	7	56
Região Metropolitana de São Paulo	Baixa	44	352
	Média	12	96
	Alta	5	40
Restante do Sudeste Urbano	Baixa	33	264
	Média	20	160
	Alta	8	64
Restante do Sudeste Rural	Baixa	19	304
	Média	9	144
	Alta	3	48
Total	Baixa	395	3.518
	Média	107	968
	Alta	52	454
	---	---	---
	Total	554	4.940

Para obter estimativas para o universo investigado usando os dados da amostra da PPV é necessário usar os pesos ou fatores de expansão amostrais. O peso relacionado a um domicílio qualquer selecionado na amostra é dado pelo inverso da probabilidade de inclusão deste domicílio na amostra. A obtenção dos pesos se deu através do estimador natural para estimar o total de uma característica y qualquer de acordo com o desenho utilizado na PPV,

ou seja, amostragem estratificada e conglomerada em dois estágios, com seleção PPT com reposição de setores no primeiro estágio (os setores são as unidades primárias de amostragem – UPAs) e seleção equiprovável dos domicílios em cada setor, dado por:

$$\hat{Y} = \sum_{h=1}^{H} \frac{1}{n_h} \sum_{i=1}^{n_h} \frac{M^*_{hi}}{p_{hi}} \frac{1}{m_{hi}} \sum_{j=1}^{m_{hi}} y_{hij} \qquad (2.2)$$

onde

n_h: número de setores selecionados para a amostra no estrato h;

M^*_{hi}: número de domicílios particulares permanentes ocupados do i-ésimo setor do estrato h, obtido pela listagem[1];

p_{hi}: probabilidade de seleção, num sorteio, do i-ésimo setor do estrato h;

$$p_{hi} = \frac{M_{hi}}{M_h}$$

M_{hi}: número de domicílios particulares permanentes ocupados do i-ésimo setor do estrato h, obtido pelo Censo Demográfico de 1991;

M_h: número total de domicílios particulares permanentes ocupados do estrato h, obtido pelo Censo Demográfico de 1991;

m_{hi}: número de domicílios com entrevista realizada no i-ésimo setor do estrato h.

A expressão (2.2) pode ser re-escrita como:

$$\hat{Y} = \sum_{h=1}^{H} \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} \frac{M_h}{n_h M_{hi}} \frac{M^*_{hi}}{m_{hi}} y_{hij} = \sum_{h=1}^{H} \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} w_{hij} y_{hij} \qquad (2.3)$$

[1] A operação de listagem foi feita para atualizar o cadastro de domicílios nos setores selecionados para a amostra. Esta operação foi realizada para cada setor no trimestre em que a pesquisa foi realizada naquele setor.

onde

$$w_{hij} = \frac{M_h}{n_h M_{hi}} \frac{M^*_{hi}}{m_{hi}} \tag{2.4}$$

é o peso do j-ésimo domicílio do i-ésimo setor do estrato h. Note que os pesos para os domicílios de um mesmo setor são constantes, variando apenas entre domicílios de setores diferentes.

2.3 – Descrição dos Dados

A Pesquisa sobre Padrões de Vida (PPV) teve um bloco do questionário destinado a levantar as características do domicílio pesquisado. A princípio, todas as variáveis relativas à estrutura do imóvel e aquelas que indicassem acesso a serviços de infra-estrutura que constaram deste bloco do questionário fizeram parte do modelo considerado para explicar o aluguel estimado. Em adição a estas, foram também incluídas variáveis em nível de setor censitário, buscando mensurar o efeito de localização do imóvel. A relação de todas as variáveis explicativas do modelo inicialmente considerado consta do Apêndice 1.

As variáveis no nível de setor censitário foram construídas com base na amostra do Censo Demográfico de 1991 da forma como foram definidas em Reis *et al*. (2001) e estão dispostas no Apêndice 1. Estas variáveis poderiam ser construídas também com a base de dados da PPV, assim como Reis *et al*. construíram as de seu estudo com base na PNAD. Entretanto como a idéia é ter medidas para descrever o setor censitário, as observações do Censo Demográfico poderiam levar a melhores estimativas, uma vez que a amostra de domicílios do Censo Demográfico em cada setor censitário é maior do que a da PPV.

Entretanto, quando foram calculadas as estimativas pelo Censo Demográfico não foi possível estimar a mediana do logaritmo da renda domiciliar em um setor censitário, quando não havia informação disponível de renda domiciliar para nenhum domicílio daquele setor. Para não perder as informações deste setor nos ajustes dos modelos, optou-se por deflacionar a renda domiciliar da PPV para o período de referência do Censo Demográfico e imputar, apenas para aquele setor, a mediana do logaritmo da renda domiciliar estimada pela PPV. O deflacionamento teve por base o Índice Nacional de Preços ao Consumidor (INPC) no período de agosto de 1991 a agosto de 1996.

A variável resposta para a modelagem – aluguel estimado – merece atenção especial. Ao invés de usar apenas os valores dos aluguéis pagos para os imóveis alugados, optou-se por utilizar uma variável construída na PPV que depende crucialmente da forma de ocupação do domicílio, a qual foi pesquisada da forma apresentada no Quadro 2.1.

Quadro 2.1 Descrição da condição de ocupação do domicílio

Condição de ocupação	Situação pesquisada
Alugado	Domicílio com aluguel pago por morador(es), ainda que parcialmente.
Próprio em aquisição	Domicílio de propriedade, total ou parcial, de um ou mais moradores e ainda não integralmente pago.
Próprio já pago	Domicílio de propriedade, total ou parcial, de um ou mais moradores e já integralmente pago.
Cedido por empregador	Domicílio cedido por empregador (particular ou público) de qualquer um dos moradores, ainda que mediante uma taxa de ocupação (impostos, condomínio, etc.) ou de conservação. Inclui domicílios cujo aluguel integral é pago, direta ou indiretamente, pelo empregador de um dos moradores.
Cedido de outra forma	Domicílio cedido gratuitamente por pessoa que não seja moradora ou por instituição que não seja empregadora de algum dos moradores, ainda que mediante uma taxa de ocupação (impostos, condomínio, etc.) ou de conservação. Inclui domicílios cujo aluguel integral é pago, direta ou indiretamente, por pessoa que não seja moradora ou por instituição que não seja empregadora de algum morador.
Invasão	Domicílio ocupado de forma ilegal.

Na Tabela 2.2 apresenta-se a distribuição dos domicílios segundo a condição de ocupação.

Tabela 2.2 Distribuição da condição de ocupação do domicílio, considerando ou não a ponderação, freqüência relativa, desvio padrão da freqüência relativa estimada e Efeito do Plano Amostral (EPA)

Condição de ocupação	Sem ponderação	Com ponderação			
	Freqüência Simples	Freqüência Simples	Freqüência Relativa (%)	Desvio Padrão Freq. Relativa	EPA Freq. Relativa
Alugado	733	3.932.994	14,6	0,82	2,69
Próprio em aquisição	333	1.342.309	5,0	0,70	5,06
Próprio já pago	3.072	17.448.164	64,9	1,25	3,39
Cedido por empregador	296	1.120.129	4,2	0,47	2,68
Cedido outra forma	434	2.601.191	9,7	0,71	2,85
Invasão	72	439.483	1,6	0,45	6,11
Total	4.940	26.884.270	100,0	0	-

Observa-se que a amostra foi relativamente pequena (4.940 domicílios) para estimar uma população com quase 27 milhões de domicílios em duas grandes regiões geográficas do Brasil. A grande maioria (64,9%) dos domicílios pesquisados foi do tipo próprio já pago e apenas 72 domicílios, menos de 2% da amostra, foram declarados como sendo invadidos. A coluna referente aos efeitos do plano amostral ampliado (EPA) indica que se estaria incorrendo em subestimação da variância das estimativas caso o desenho de amostragem fosse considerado aleatório simples para efeito de estimação e análise dos dados. Os casos mais agudos de subestimação da variância ocorreriam para as estimativas das freqüências relativas dos domicílios invadidos ou dos próprios em aquisição.

Para os domicílios alugados (14,6%) foi perguntado o valor do aluguel pago nos últimos 30 dias. Para os domicílios próprios em aquisição, foi levantado o valor da prestação paga nos últimos 30 dias. Para os domicílios que se enquadravam nas demais formas de ocupação, foi perguntado o valor estimado do aluguel do imóvel. Desta forma, a variável

resposta considerada neste estudo – aluguel estimado – foi construída com base no valor do aluguel, da prestação paga ou da estimativa de aluguel, dependendo da forma de ocupação do domicílio.

Dos 4.940 domicílios pesquisados, 16 não possuíam informação sobre a estimativa do valor do aluguel do imóvel, sendo 11 próprios já pagos e 5 cedidos de outra forma, restando 4.924 domicílios com valores válidos na variável "aluguel estimado". A Tabela 2.3 apresenta a média, mínimo, máximo e quartis do aluguel estimado segundo a forma de ocupação do imóvel, sendo as estimativas dos quartis e da média calculadas considerando os pesos amostrais.

Tabela 2.3 Estimativas da média, mínimo, máximo e quartis do aluguel segundo a forma de ocupação do imóvel

Condição de ocupação	Média	Mínimo	1° Quartil	Mediana	3° Quartil	Máximo
Alugado	242,6	0,0	92,6	181,4	299,5	4.500,0
Próprio em aquisição	180,6	0,0	29,3	64,2	218,2	3.000,0
Próprio já pago	291,1	5,0	58,1	148,9	342,3	10.000,0
Cedido por empregador	145,7	5,0	26,9	48,5	97,4	2.200,0
Cedido outra forma	170,7	5,0	46,9	95,6	194,2	1.500,0
Invasão	119,7	10,0	48,4	94,2	133,1	600,0
Todos	258,0	0,0	50,0	149,4	295,4	10.000,0

Observa-se que a média do aluguel estimado é maior para os imóveis próprios já pagos, seguida pela média dos imóveis alugados. Ao se analisar os quartis, o mínimo e o máximo, verifica-se que a distribuição do aluguel estimado é fortemente assimétrica à direita para qualquer forma de ocupação do domicílio. A Figura 2.1 apresenta a forma da distribuição da variável aluguel estimado, observando-se ser necessário aplicar uma transformação na variável resposta.

Figura 2.1 Histograma do aluguel estimado

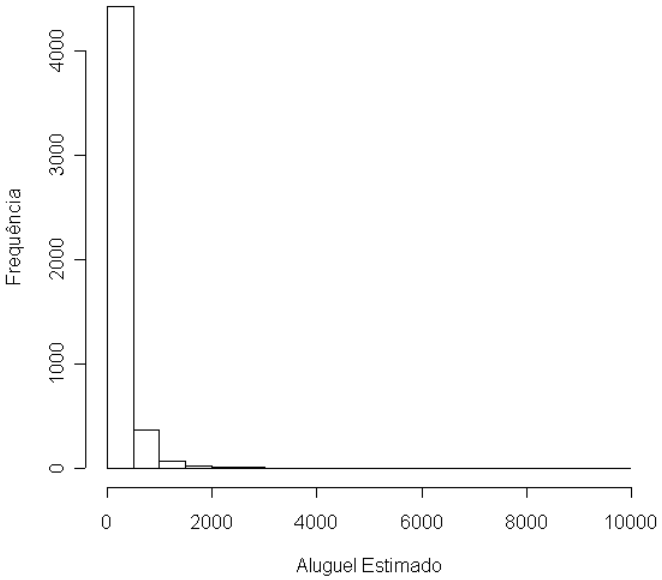
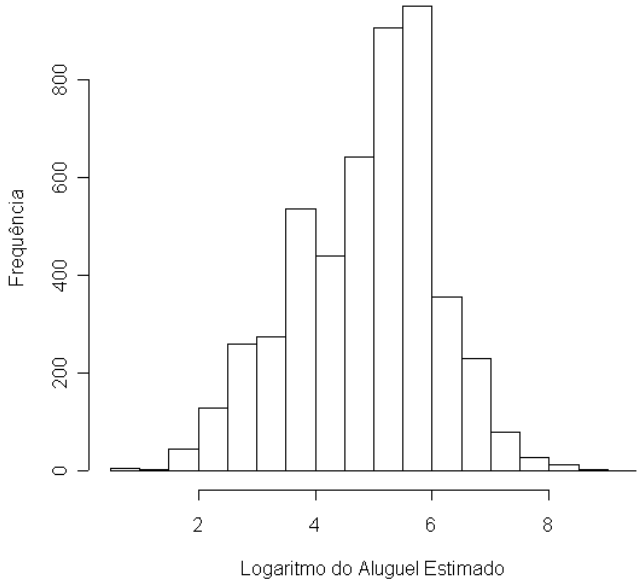


Figura 2.2 Histograma do logaritmo do aluguel estimado

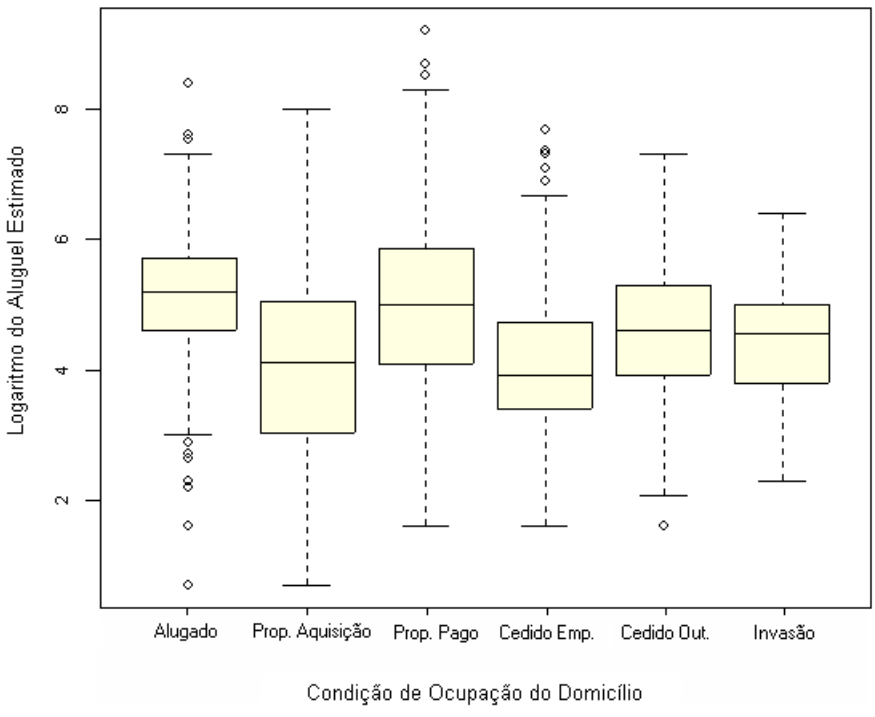


Aplicando-se uma transformação logarítmica ao valor do aluguel estimado, chega-se a uma distribuição quase simétrica, como apresentado na Figura 2.2, que considera os pesos amostrais. Neste procedimento foram perdidos mais 22 domicílios para a análise. Destes, 3 informaram que não pagavam nada no aluguel de suas residências e 19 informaram que o

valor da prestação do imóvel em aquisição era zero. O total de domicílios perdidos foi de 38, o que representa menos de 1% dos 4.940 domicílios na amostra. Como a proporção de perda foi muito pequena e não houve perda total para nenhum setor censitário, não foi adotado nenhum procedimento para o tratamento da não-resposta, prosseguindo a análise com a amostra de casos completos, com 4.902 domicílios.

A Figura 2.3 apresenta os diagramas de caixas para o logaritmo do aluguel estimado segundo a forma de ocupação do imóvel, considerando os pesos amostrais no cálculo das estimativas das juntas. Nesta figura evidencia-se que a transformação logarítmica tornou as distribuições relativamente simétricas para todas as classes. Chama a atenção o valor relativamente elevado da mediana do aluguel estimado para os domicílios invadidos em relação às outras formas de ocupação. Em contrapartida, o valor mediano dos imóveis cedidos por empregador é relativamente baixo.

Figura 2.3 Diagramas de caixas do logaritmo do aluguel estimado segundo a forma de ocupação do imóvel



3 – MODELOS HIERÁRQUICOS USUAIS

3.1 – Modelos Hierárquicos: Uma Introdução

Em diversas populações é comum que os dados investigados numa pesquisa possuam uma estrutura de agrupamento ou hierarquia. Um exemplo clássico para este tipo de situação é quando se deseja analisar o rendimento escolar de alunos, onde é natural questionar se as condições de ensino na turma ou escola podem interferir de alguma forma nos resultados alcançados pelos alunos. Neste caso existe uma estrutura hierárquica de três níveis, com os alunos sendo as unidades de nível 1, as turmas as unidades de nível 2 e as escolas as unidades de nível 3. Situações similares podem ser encontradas em relações como funcionários em empresas, valor de aluguel de imóveis residenciais em diferentes bairros, etc. As unidades de nível 1 também podem ser chamadas de unidades elementares e as de nível 2 (ou superior) de grupos.

Um caso especial de dados com estrutura de grupo é quando se trabalha com estudos longitudinais, em que observações repetidas (unidades de nível 1) são realizadas sobre determinados indivíduos (unidades de nível 2). Dados obtidos por amostragem em dois estágios também possuem estrutura hierárquica, uma vez que existe dependência das unidades do segundo estágio selecionadas dentro das mesmas unidades do primeiro estágio. O mesmo raciocínio é válido para desenhos de amostras em mais de dois estágios.

Na maioria dos casos as unidades elementares podem ser independentes para grupos diferentes, mas possuem características semelhantes dentro de um determinado grupo. Esta dependência existente entre as observações dentro de um mesmo grupo é, por vezes, encarada como um problema para a análise dos dados e ignorá-la pode levar a conclusões equivocadas. Em certas situações é possível agregar as unidades de nível 1 para o nível de grupo, entretanto

as relações que forem verificadas nesse nível só poderão ser utilizadas para fazer afirmações sobre os grupos e não para fazer afirmações sobre as unidades elementares.

Por outro lado, também há problemas em tratar os dados apenas no nível de unidades elementares. Os modelos de regressão tradicionais não seriam válidos, pois o pressuposto de independência dos resíduos de observações distintas não seria respeitado. Além disto, se as unidades de nível 2 têm alguma relação significativa com o problema em estudo não se pode simplesmente desprezá-las, como analisar o desempenho escolar de alunos esquecendo-se da qualificação ou dos métodos de ensino dos professores. Outro problema ocorre quando os dados são pesquisados através de amostragem, pois considerar as unidades de nível 1 como sendo observações independentes geralmente implica em subestimar a margem de erro das estimativas, se o plano amostral for conglomerado.

A metodologia adequada para a análise de dados com padrões de variabilidade complexos como estes onde há uma estrutura de grupos nas observações se dá através do uso de modelos multiníveis, também conhecidos como modelos hierárquicos, modelos de efeitos mistos, modelos de coeficientes aleatórios e modelos de componentes de variância (Snijders e Bosker, 1999). Estes modelos consideram que alguns coeficientes são fixos e outros são aleatórios. Apesar de serem mais complexos, apresentam resultados melhores para modelar dados com estrutura de agrupamento que modelos usuais de regressão.

Apenas modelos lineares hierárquicos com dois níveis serão objeto de análise nesta dissertação. Para estes, a variável resposta deve estar no nível 1 e ter relação com variáveis preditoras de ambos os níveis de variação. É um tipo de modelo de regressão ou de análise de variância, diferindo dos tradicionais por possuir um termo de erro para cada nível da hierarquia. Outros modelos com esta estrutura são obtidos através de experimentos do tipo *split-plot*.

3.2 – Modelos com Interceptos Aleatórios

O modelo hierárquico com dois níveis mais simples é o modelo de análise de variância (ANOVA) definido por:

modelo de nível 1:

$$y_{ij} = \beta_{0i} + \varepsilon_{ij} \tag{3.1}$$

onde

y_{ij} : variável resposta com $i = 1, \ldots, n$, $j = 1, \ldots, m_i$;

n: tamanho da amostra de unidades de nível 2;

m_i: tamanho da amostra de unidades de nível 1 na unidade de nível 2 i;

β_{0i} : média da unidade de nível 2 i;

$\varepsilon_{ij} \sim N\left(0, \sigma^2\right)$;

$Cov\left(\varepsilon_{ij}, \varepsilon_{kl}\right) = 0$, $\forall$ par ij diferente do par kl .

modelo de nível 2:

$$\beta_{0i} = \gamma_{00} + u_{0i} \tag{3.2}$$

onde

γ_{00} : resposta média esperada para a população como um todo;

u_{0i} : efeito aleatório associado ao i-ésimo grupo;

$u_{0i} \sim N\left(0, \tau_0\right)$;

$Cov\left(u_{0i}, u_{0k}\right) = 0$, $\forall\ k \neq i$;

$Cov\left(\varepsilon_{ij}, u_{0k}\right) = 0$, $\forall\ ij$ e k .

O modelo descrito nas equações (3.1) e (3.2) não possui nenhuma variável explicativa, apenas uma média geral γ_{00}, um termo de erro para o nível de grupo u_{0i} e outro para o nível de indivíduos ε_{ij}. A variância da variável resposta y_{ij} pode ser decomposta como a soma da variância entre os grupos τ_0 e da variância intra-grupos σ^2:

$$Var(y_{ij}) = \tau_0 + \sigma^2. \tag{3.3}$$

Uma medida da semelhança entre as unidades de nível 1 em um mesmo grupo é o coeficiente de correlação intraclasse, definido por:

$$\rho_I = \frac{\tau_0}{\tau_0 + \sigma^2}. \tag{3.4}$$

Esta medida representa a proporção da variância da resposta explicada pela variabilidade das unidades de nível 2. Ela mede a homogeneidade dos grupos ou ainda o grau de dependência dos indivíduos no mesmo grupo. Se a correlação intraclasse for não nula, isto é, se $\rho_I \neq 0$, então o efeito de grupo estará presente e os modelos de regressão tradicionais não serão mais adequados, dado que a condição de independência dos resíduos nestes modelos não seria satisfeita. Caso esse coeficiente seja igual a zero, pode-se usar os modelos tradicionais de regressão.

Para tentar melhorar a qualidade do ajuste de um modelo hierárquico é possível adicionar variáveis explicativas em ambos os níveis de variação, sendo que as variáveis do nível 2 podem ser médias, somas ou outras agregações de variáveis do nível 1. Uma formulação geral para o modelo linear hierárquico de intercepto aleatório é definida por:

modelo de nível 1:

$$y_{ij} = \beta_{0i} + \sum_{p=1}^{P} \beta_p x_{pij} + \varepsilon_{ij} \tag{3.5}$$

onde

x_{pij} : valor da p-ésima variável explicativa do nível 1, para a unidade j do grupo i, com

$p = 1, \ldots, P$, $i = 1, \ldots, n$, $j = 1, \ldots, m_i$;

β_p : parâmetro fixo associado à p-ésima variável explicativa de nível 1;

$\varepsilon_{ij} \sim N(0, \sigma^2)$;

$Cov(\varepsilon_{ij}, \varepsilon_{kl}) = 0$, $\forall$ par ij diferente do par kl.

modelo de nível 2:

$$\beta_{0i} = \gamma_{00} + \sum_{q=1}^{Q} \gamma_{0q} z_{qi} + u_{0i} \tag{3.6}$$

onde

z_{qi} : valor da q-ésima variável explicativa do nível 2 para o grupo i, com $q = 1, \ldots, Q$,

$i = 1, \ldots, n$;

γ_{0q} : parâmetro fixo associado à q-ésima variável explicativa de nível 2;

$u_{0i} \sim N(0, \tau_0)$;

$Cov(u_{0i}, u_{0k}) = 0$, $\forall\ k \neq i$;

$Cov(\varepsilon_{ij}, u_{0k}) = 0$, $\forall\ ij$ e k.

3.3 – Modelos com Interceptos e Coeficientes Aleatórios

Este é o caso geral para modelos lineares hierárquicos de dois níveis. Neste modelo não apenas os interceptos variam aleatoriamente entre os grupos, mas também as inclinações do nível 1. Quando o número de inclinações aleatórias não for pequeno, o número de parâmetros se torna muito grande e a interpretação fica mais complicada. A definição do modelo com o intercepto e uma inclinação aleatória é dada por:

modelo de nível 1:

$$y_{ij} = \beta_{0i} + \beta_{1i} x_{1ij} + \varepsilon_{ij} \tag{3.7}$$

onde

β_{0i} : média da unidade *i* de nível 2;

β_{1i} : inclinação relativa à variável explicativa x_{1ij}, associada à unidade *i* de nível 2;

$\varepsilon_{ij} \sim N\left(0, \sigma^2\right)$;

$Cov\left(\varepsilon_{ij}, \varepsilon_{kl}\right) = 0$, $\forall$ par ij diferente do par kl.

modelo de nível 2:

$$\begin{aligned} \beta_{0i} &= \gamma_{00} + \sum_{q=1}^{Q} \gamma_{0q} z_{qi} + u_{0i} \\ \beta_{1i} &= \gamma_{10} + \sum_{q=1}^{Q} \gamma_{1q} z_{qi} + u_{1i} \end{aligned} \tag{3.8}$$

onde

γ_{00} : valor esperado dos interceptos na população para $z_{qi} = 0$;

γ_{10} : valor esperado das inclinações na população para $z_{qi} = 0$;

u_{0i} : efeito aleatório no intercepto associado ao i-ésimo grupo;

u_{1i} : efeito aleatório na inclinação associado ao i-ésimo grupo;

$u_{0i} \sim N(0, \tau_0)$ e $Cov(u_{0i}, u_{0k}) = 0$, $\forall\ k \neq i$;

$u_{1i} \sim N(0, \tau_1)$ e $Cov(u_{1i}, u_{1k}) = 0$, $\forall\ k \neq i$;

$Cov(u_{0i}, u_{1i}) = \tau_{01}$, $\forall\ i$;

$Cov(u_{0i}, u_{1k}) = 0$, $\forall\ i \neq k$;

$Cov(u_{0i}, \varepsilon_{kj}) = Cov(u_{1i}, \varepsilon_{kj}) = 0$, $\forall\ i, j, k$.

3.4 – Estimação

Os métodos mais utilizados para estimar os parâmetros de um modelo linear hierárquico são o de Máxima Verossimilhança (MV), o de Máxima Verossimilhança Restrita (MVR) e o de Mínimos Quadrados Generalizados Iterativo (MQGI). A diferença entre os dois primeiros é que o de Máxima Verossimilhança Restrita estima os componentes de variância (σ^2 e τ's) levando em conta a perda de graus de liberdade para estimar os parâmetros de regressão (β's e γ's), o que o método de Máxima Verossimilhança não faz, implicando que as estimativas obtidas por Máxima Verossimilhança tenham um viés para menos. Entretanto, segundo Snijders e Bosker (1999), as estimativas por Máxima Verossimilhança devem ser usadas quando há interesse em fazer teste de *deviance* para comparar modelos que tenham a parte fixa diferente. Sob hipótese de normalidade, Goldstein (1995) afirma que o método MQGI é equivalente ao método MV.

3.4.1 – Estimação MQGI para Modelo de Intercepto Aleatório com Uma só Variável Explicativa de Nível 1

O método dos Mínimos Quadrados Generalizados Iterativo (MQGI) é o que será utilizado neste trabalho, pois constitui a base do procedimento desenvolvido por Pfeffermann *et al*. (1998a) para obter estimativas que consideram o plano amostral em um modelo hierárquico. Para ilustrar como funciona este método considere, por simplicidade, o modelo hierárquico de componentes de variância de 2 níveis, conforme apresentado em (3.9):

$$y_{ij} = \gamma_{00} + \beta_1 x_{1ij} + u_{0i} + \varepsilon_{ij} \tag{3.9}$$

A estrutura de covariâncias da variável resposta y_{ij} pode ser escrita como uma matriz bloco diagonal da forma:

$$\Sigma = \begin{bmatrix} \Sigma_1 & \cdots & \mathbf{0} \\ \vdots & \vdots & \vdots \\ \mathbf{0} & \cdots & \Sigma_n \end{bmatrix} \tag{3.10}$$

onde

$$\Sigma_i = V\left(u_{0i} + \varepsilon_{ij}\right) = \tau_0 \begin{bmatrix} 1 & \cdots & 1 \\ \vdots & \vdots & \vdots \\ 1 & \cdots & 1 \end{bmatrix} + \sigma^2 \begin{bmatrix} 1 & \cdots & 0 \\ \vdots & \vdots & \vdots \\ 0 & \cdots & 1 \end{bmatrix}, \quad i = 1,\ldots,n\,.$$

Se τ_0 e σ^2 fossem conhecidos, as estimativas dos efeitos fixos γ_{00} e β_1 poderiam ser obtidas por Mínimos Quadrados Generalizados (MQG), e seriam dadas por:

$$\hat{\boldsymbol{\theta}} = \left(X'\Sigma^{-1}X\right)^{-1}X'\Sigma^{-1}Y = \begin{bmatrix}\hat{\gamma}_{00}\\ \hat{\beta}_1\end{bmatrix} \tag{3.11}$$

onde

$$Y' = \begin{bmatrix} y_{11} & \cdots & y_{1m_1} & y_{21} & \cdots & y_{2m_2} & \cdots & y_{n1} & \cdots & y_{nm_n}\end{bmatrix};$$

$$X' = \begin{bmatrix} 1 & \cdots & 1 & 1 & \cdots & 1 & \cdots & 1 & \cdots & 1\\ x_{111} & \cdots & x_{11m_1} & x_{121} & \cdots & x_{12m_2} & \cdots & x_{1n1} & \cdots & x_{1nm_n}\end{bmatrix}.$$

Sob a hipótese de que τ_0 e σ^2 sejam conhecidos, $\hat{\boldsymbol{\theta}}$ em (3.11) é o melhor estimador linear não-viesado para $\boldsymbol{\theta} = \begin{bmatrix}\gamma_{00}\\ \beta_1\end{bmatrix}$.

Geralmente os valores de τ_0 e σ^2 não são conhecidos, o que torna necessário um procedimento para estimá-los, juntamente com os efeitos fixos γ_{00} e β_1. O MQGI é um procedimento iterativo para resolver este problema, alternativo aos métodos de MV e MVR. Neste método, os valores iniciais são estimativas dos efeitos fixos obtidas por mínimos quadrados ordinários (MQO), ou seja, considerando $\tau_0^{(0)} = 0$ e $\sigma^{2(0)} = 1$. Os resíduos decorrentes do ajuste deste modelo podem ser escritos como:

$$\hat{r}_{ij} = y_{ij} - \hat{\gamma}_{00}^{(1)} - \hat{\beta}_1^{(1)}x_{1ij} \tag{3.12}$$

onde $\hat{\gamma}_{00}^{(1)}$ e $\hat{\beta}_1^{(1)}$ são as estimativas de γ_{00} e β_1 obtidas sob a hipótese de que $\tau_0 = 0$ e $\sigma^2 = 1$.

Escrevendo (3.12) em forma matricial, tem-se:

$$\hat{R} = Y - X\hat{\boldsymbol{\theta}}^{(1)} \tag{3.13}$$

onde $\hat{\boldsymbol{\theta}}^{(1)} = \begin{bmatrix} \hat{\gamma}_{00}^{(1)} \\ \hat{\beta}_1^{(1)} \end{bmatrix}$.

Note-se que sob a hipótese de que $\tau_0 = 0$ e $\sigma^2 = 1$, têm-se $E\left(\hat{R}\hat{R}'\right) = \Sigma$. Esta matriz pode ser re-escrita em forma de vetor $vec\left(\hat{R}\hat{R}'\right)$, dispondo as colunas umas embaixo das outras e similarmente com a matriz Σ, construindo um vetor $vec(\Sigma)$. A relação entre $vec\left(\hat{R}\hat{R}'\right)$ e $vec(\Sigma)$ pode ser formulada em termos de um modelo linear da seguinte maneira:

$$vec\left(\hat{R}\hat{R}'\right) = vec(\Sigma) + \boldsymbol{\eta} = \begin{bmatrix} 1 & 1 \\ 1 & 0 \\ \vdots & \vdots \\ 1 & 1 \end{bmatrix} \begin{bmatrix} \tau_0 \\ \sigma^2 \end{bmatrix} + \boldsymbol{\eta} \tag{3.14}$$

em forma matricial

$$vec\left(\hat{R}\hat{R}'\right) = H\boldsymbol{\lambda} + \boldsymbol{\eta} \tag{3.15}$$

onde

H é a matriz de zeros e uns associada aos componentes de variância τ_0 e σ^2;

$\boldsymbol{\lambda} = \begin{bmatrix} \tau_0 \\ \sigma^2 \end{bmatrix}$ é o vetor de parâmetros aleatórios a serem estimados;

$\boldsymbol{\eta}$ é o vetor de erros aleatórios, com $E(\boldsymbol{\eta}) = \mathbf{0}$.

O vetor de parâmetros $\boldsymbol{\lambda}$ pode ser estimado por mínimos quadrados generalizados, com a matriz de covariâncias de $vec\left(\hat{R}\hat{R}'\right)$ dada por $\Sigma^* = 2\left(\Sigma^{-1} \otimes \Sigma^{-1}\right)$, onde $\otimes$ é o produto de Kronecker. A estimativa do vetor $\boldsymbol{\lambda}$ é dada por:

$$\hat{\boldsymbol{\lambda}} = \left(H'\Sigma^{*^{-1}} H\right)^{-1} H'\Sigma^{*^{-1}} vec\left(\hat{R}\hat{R}'\right) \tag{3.16}$$

Uma vez obtida a estimativa $\hat{\boldsymbol{\lambda}}$ para $\boldsymbol{\lambda}$, pode-se obter novas estimativas dos parâmetros fixos, considerando agora os valores estimados $\hat{\tau}_0$ e $\hat{\sigma}^2$ obtidos usando (3.16). Repete-se o processo recursivamente até que a convergência estipulada pelo usuário seja alcançada.

De forma resumida o método de Mínimos Quadrados Generalizados Iterativo (MQGI) pode ser descrito pelo algoritmo a seguir.

Passo 0: Assumir $\hat{\tau}_0^{(0)} = 0$ e $\hat{\sigma}^{2(0)} = 1$, e portanto fazer $\Sigma^{(0)} = I$.

Para $t = 1,2,...$ repetir os passos 1 e 2 a seguir.

Passo 1: Estimar os parâmetros dos efeitos fixos considerando que os parâmetros de efeitos aleatórios sejam conhecidos, ou seja, $\hat{\boldsymbol{\theta}}^{(t)} = \hat{\boldsymbol{\theta}}\left(\hat{\tau}_0^{(t-1)}, \hat{\sigma}^{2(t-1)}\right) = \left[X'\left(\Sigma^{(t-1)}\right)^{-1} X\right]^{-1} X'\left(\Sigma^{(t-1)}\right)^{-1} Y$;

Passo 2: Estimar os parâmetros dos efeitos aleatórios considerando que os parâmetros de efeito fixo sejam conhecidos, da seguinte maneira:

a) $\Sigma^{*(t)} = 2\left[\left(\Sigma^{(t-1)}\right)^{-1} \otimes \left(\Sigma^{(t-1)}\right)^{-1}\right]$

b) Obter $vec\left(\hat{R}^{(t)}\hat{R}^{(t)'}\right)$, onde $\hat{R}^{(t)} = Y - X\hat{\boldsymbol{\theta}}^{(t)}$

c) Calcular $\hat{\boldsymbol{\lambda}}^{(t)} = \begin{bmatrix} \hat{\tau}_0^{(t)} \\ \hat{\sigma}^{2(t)} \end{bmatrix} = \left[H' \left(\Sigma^{*(t)} \right)^{-1} H \right]^{-1} H' \left(\Sigma^{*(t)} \right)^{-1} vec\left(\hat{R}^{(t)} \hat{R}^{(t)'} \right)$

Os passos acima devem ser repetidos até que a convergência seja alcançada. O critério de convergência usual é dado por:

$$\max \left| \frac{\hat{\omega}_l^{(t)} - \hat{\omega}_l^{(t-1)}}{\hat{\omega}_l^{(t-1)}} \right| < 10^{-K} \tag{3.17}$$

onde

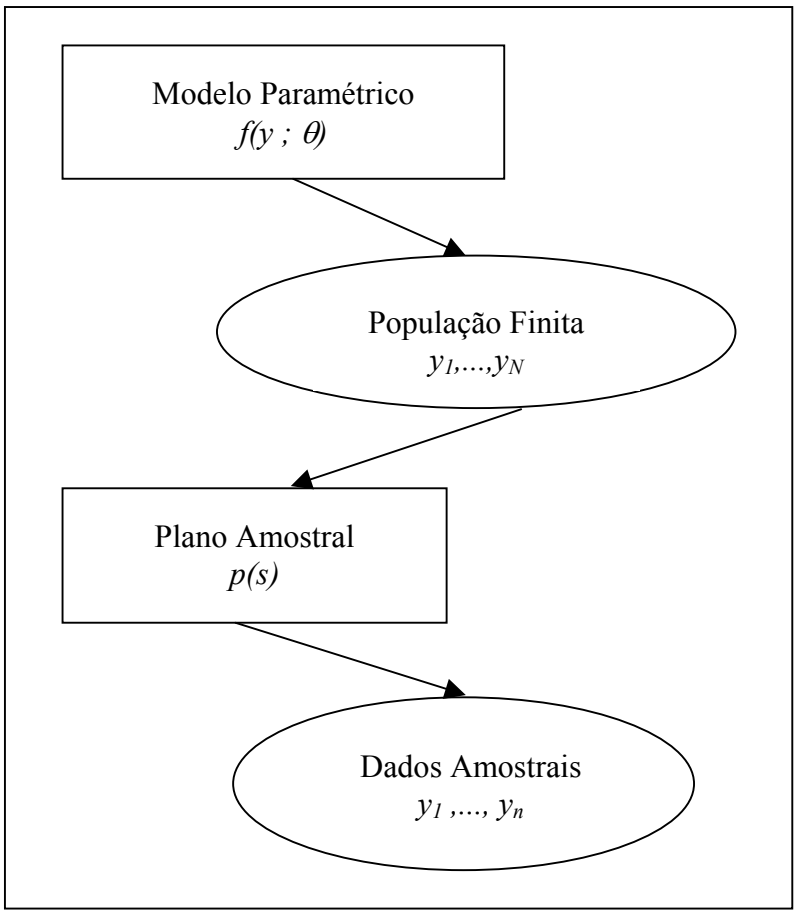
K é fixado pelo usuário;

$\hat{\boldsymbol{\omega}}^{(t)} = \left[\omega_1^{(t)}, \ldots, \omega_L^{(t)} \right] = \left[\hat{\boldsymbol{\theta}}^{(t)'}, \hat{\boldsymbol{\lambda}}^{(t)'} \right] = \left[\hat{\gamma}_{00}^{(t)}, \hat{\beta}_1^{(t)}, \hat{\tau}_0^{(t)}, \hat{\sigma}^{2(t)} \right].$

4 – MODELOS HIERÁRQUICOS COM DADOS DE AMOSTRAS COMPLEXAS

Uma abordagem comumente utilizada para analisar dados provenientes de amostras complexas é a de *modelagem de superpopulação*. Nesta abordagem, um modelo paramétrico $f(y;\theta)$ é formulado para descrever as observações $y_1,\ldots,y_N$ na população. Como a população não será completamente observada, a inferência para o parâmetro θ do modelo terá que ser feita com base nos dados de uma amostra (sem perda de generalidade representada por $y_1,\ldots,y_n$) selecionada utilizando um plano amostral $p(s)$.

Figura 4.1 Modelagem de Superpopulação



A Figura 4.1 foi retirada de Pessoa e Silva (1998) e apresenta um esquema gráfico da abordagem de modelagem de superpopulação. Os modelos apresentados nesta dissertação serão ajustados por meio desta abordagem.

4.1 – Planos Amostrais Informativos e Ignoráveis

Diversas pesquisas por amostragem utilizam esquemas complexos de seleção. Freqüentemente, os modelos hierárquicos ajustados aos dados destas pesquisas não levam em consideração o planejamento amostral nas estimativas dos parâmetros e das variâncias dos estimadores. Nestes casos, os dados são tratados como provenientes de uma amostra em dois estágios com seleção aleatória simples nos dois estágios e as conclusões com base no ajuste deste modelo podem ser equivocadas. Desta forma, é necessário o uso de procedimentos para corrigir as estimativas e fazer com que o modelo reflita adequadamente o comportamento da população.

O tipo de tratamento dispensado a dados de amostras é diferenciado de acordo com o desenho amostral. Quando o esquema de seleção for do tipo aleatório simples com reposição, o modelo na amostra é o mesmo que o da população e, neste caso, o plano amostral é dito *ignorável*. Entretanto, este tipo de esquema não é muito utilizado em pesquisas por razões de eficiência e de custo, preferindo-se esquemas amostrais complexos que incluam estratificação, conglomeração, probabilidades de seleção desiguais, etc. Para estes desenhos, os modelos na amostra e na população podem ser muito diferentes (plano amostral *não-ignorável*). Segundo Skinner *et al*. (1989), ignorar nestes casos as informações do plano amostral na análise pode acarretar vícios na inferência.

Exemplo 4.1 Plano Amostral Ignorável

Sob uma abordagem usual, os dados da amostra observada poderiam ser modelados por $Y_1,\ldots,Y_n$ v.a. IID $\sim f(y;\theta)$, com $l_s(\theta \mid \mathbf{y}) = \prod_{i=1}^{n} f(y_i;\theta)$.

Sob a abordagem de superpopulação, o modelo populacional seria $Y_1,\ldots,Y_N$ v.a. IID $\sim f(y;\theta)$, com $l_U(\theta \mid \mathbf{y}) = \prod_{i=1}^{N} f(y_i;\theta)$.

Aplicando um esquema de amostragem aleatória simples com reposição teria-se $l_s(\theta \mid \mathbf{y}) = \prod_{i=1}^{n} f(y_i;\theta) = l_U(\theta \mid \mathbf{y})$. Quando o esquema amostral não for aleatório simples com reposição esta igualdade não é mais válida.

Em planos amostrais onde a seleção da amostra depende do valor de variáveis de interesse da pesquisa, o plano amostral é dito *informativo*, com $P(i \in s \mid y_i) \neq 0$. Pessoa e Silva (1998) citam como exemplo deste tipo de plano um estudo de caso-controle. Nesta situação a amostra é selecionada de forma que existam *casos* (presença de determinada condição) e *controles* (ausência desta condição), havendo o interesse de modelar este indicador (presença ou ausência da condição) em função de outras variáveis da pesquisa. Note que este indicador é considerado no mecanismo de seleção da amostra. Planos amostrais informativos são não-ignoráveis.

Exemplo 4.2 Plano Amostral Informativo

Um exemplo simples de esquema amostral informativo é apresentado por Pfeffermann (2002) no contexto de estimação em pequenas áreas. Suponha que todas as respostas y_{ij} de uma dada área i sejam binárias com $P(y_{ij} = 1) = p$ e

uma amostra seja selecionada com probabilidades $P(j \in s_i \mid y_{ij} = k) = \pi_k$, para $k = 0, 1$. Com a aplicação da regra de Bayes, tem-se que $P(y_{ij} = 1 \mid j \in s_i) = \pi_1 p / [\pi_1 p + \pi_2 (1 - p)] = p_s$. Neste caso, $p \neq p_s$, ao menos que $\pi_1 = \pi_2$, e a inferência em *p* que ignora o esquema de seleção, na verdade significa inferência em p_s.

Quando há o intuito de ajustar modelos usando dados de amostras obtidas com planos *não-ignoráveis* devem ser considerados estimadores especiais, ou considerar modelos mais complexos que permitam acomodar aspectos da estrutura da população e do plano amostral. Para modelos paramétricos de apenas um nível o método da Máxima Pseudo-Verossimilhança (MPV) considera os pesos amostrais na estimação dos parâmetros, como descrito por Skinner (1989). Este método se baseia em uma idéia central: o modelo de interesse é proposto para descrever os dados da população, não os dados da amostra. Trata-se de um exemplo de modelagem de superpopulação (ver Pessoa e Silva, 1998, capítulo 5).

Formulando o modelo para a população, é possível escrever a função de verossimilhança correspondente, supondo que os valores das variáveis de interesse são conhecidos para todas as unidades da população (isto equivaleria a ter realizado um censo da população). Para um modelo que considera que os valores da população finita são independentes, as funções de verossimilhança e de log-verossimilhança populacionais seriam dadas respectivamente por:

$$l_U(\boldsymbol{\theta}) = \prod_{j \in U} f(y_j; \boldsymbol{\theta}) \tag{4.1}$$

e

$$L_U(\boldsymbol{\theta}) = \sum_{j \in U} \log\left[f(y_j;\boldsymbol{\theta})\right] \ . \tag{4.2}$$

As equações de verossimilhança populacionais correspondentes são dadas por:

$$\sum_{j \in U} \mathbf{u}_j(\boldsymbol{\theta}) = \mathbf{0} \tag{4.3}$$

onde $\mathbf{u}_j(\boldsymbol{\theta}) = \partial \log\left[f(y_j;\boldsymbol{\theta})\right] / \partial \boldsymbol{\theta}$ é o vetor de escores do elemento j, $j \in U$.

Na hipótese de um censo, o estimador de Máxima Verossimilhança (MV) de $\boldsymbol{\theta}$ seria a solução das "equações de verossimilhança do censo" dadas em (4.3). Mas no caso de dados obtidos por pesquisas amostrais estas equações não são conhecidas, entretanto podem ser estimadas, ao notar que a soma no primeiro termo de (4.3) é um total populacional cuja estimação a partir da amostra é imediata.

O estimador de Máxima Pseudo-Verossimilhança (MPV) para um vetor de totais populacionais $\mathbf{T} = \sum_{j \in U} \mathbf{u}_j(\boldsymbol{\theta})$ é obtido por uma ponderação das funções escores da forma $\hat{\mathbf{T}} = \sum_{j \in s} w_j \mathbf{u}_j(\boldsymbol{\theta})$, onde w_j são pesos amostrais definidos de maneira apropriada. O estimador MPV será a solução das equações de Pseudo-Verossimilhança dadas por $\hat{\mathbf{T}} = \sum_{j \in s} w_j \mathbf{u}_j(\boldsymbol{\theta}) = \mathbf{0}$.

Em um modelo hierárquico as observações da população finita não são independentes, e não é mais possível escrever o logaritmo da verossimilhança como uma soma de logaritmos de densidades para cada unidade da população finita, implicando que os parâmetros do modelo hierárquico não podem mais ser estimados por uma simples ponderação das funções escores das observações da amostra.

Uma abordagem alternativa para o ajuste de modelos hierárquicos com dados de amostras complexas foi desenvolvida por Pfeffermann *et al*. (1998a). O procedimento é uma adaptação do método de mínimos quadrados generalizados iterativo (MQGI) que procura incorporar as informações do plano amostral no ajuste, buscando compensar o efeito de diferentes probabilidades de seleção das unidades da amostra.

4.2 – Efeito do Plano Amostral Ampliado (EPA)

Uma medida utilizada para dimensionar o efeito de não considerar as informações do desenho amostral nas estimativas de parâmetros de um modelo, conforme definido em Pessoa e Silva (1998), é dada por:

$$EPA\left(\hat{\theta},\upsilon_0\right)=\frac{V_{VERD}\left(\hat{\theta}\right)}{E_{VERD}\left(\upsilon_0\right)} \tag{4.4}$$

onde

$\upsilon_0 = \hat{V}_{IID}\left(\hat{\theta}\right)$ é um estimador da variância de $\hat{\theta}$ calculado sob a hipótese de que as observações são IID;

$E_{VERD}\left(\upsilon_0\right)$ é o valor esperado da distribuição de υ_0 sob o plano amostral verdadeiro;

$V_{VERD}\left(\hat{\theta}\right)$ é a variância de $\hat{\theta}$ sob o plano amostral verdadeiro.

A medida definida em (4.4) é denominada de Efeito do Plano Amostral Ampliado (EPA) e avalia a tendência de υ_0 a subestimar ou superestimar a verdadeira variância de $\hat{\theta}$. Valores de EPA maiores que 1 indicam que ignorar as informações do plano amostral

implicam em subestimação da variância verdadeira de $\hat{\theta}$. Analogamente, valores menores que 1 implicam em superestimação da variância verdadeira de $\hat{\theta}$.

4.3 – Definição do Modelo Populacional e Pressupostos do Desenho Amostral

Considere uma população em dois níveis, com N unidades de nível 2 (correspondentes às unidades primárias de amostragem) e M_i unidades de nível 1 na i-ésima unidade de nível 2 $(i = 1,..., N)$. Suponha que na população a variável resposta y_{ij} seja gerada pelo modelo de dois níveis a seguir:

$$y_{ij} = \mathbf{x}_{ij}^{'}\boldsymbol{\beta} + \mathbf{z}_i^{'}\mathbf{u}_i + z_{0ij}\varepsilon_{ij} \qquad (4.5)$$

onde

$\mathbf{x}_{ij}$ e $\mathbf{z}_i$ são vetores de covariáveis de dimensões p e q respectivamente com $i = 1,\ldots, N$ e $j = 1,\ldots, M_i$.

$\boldsymbol{\beta}$ é um vetor de parâmetros fixos de dimensão p;

$\mathbf{u}_i$ é um vetor de erros aleatórios associado à unidade i de nível 2 de dimensão q;

ε_{ij} é um termo de erros aleatórios associado à unidade j do grupo i;

$\mathbf{u}_i$ e ε_{ij} são mutuamente independentes, com $\mathbf{u}_i \sim N(\mathbf{0}, \boldsymbol{\Omega})$ e $\varepsilon_{ij} \sim N(0, \sigma^2)$;

z_{0ij} é um escalar correspondente à unidade j do grupo i.

Na maioria das aplicações o termo z_{0ij} é igual a 1 para todo *i* e *j*. Entretanto, valores desiguais de z_{0ij} permitem representar padrões conhecidos de heteroscedasticidade dentro dos conglomerados.

Assume-se um esquema amostral de dois estágios, com *n* unidades de nível 2 selecionadas no primeiro estágio com probabilidades de inclusão $\pi_i (i = 1,..., N)$. No segundo estágio são selecionadas m_i unidades de nível 1 na *i*-ésima unidade de nível 2 com probabilidades de inclusão $\pi_{j|i}$. Portanto, a probabilidade não condicional de inclusão da unidade elementar *j* do conglomerado *i* na amostra é dada por $\pi_{ij} = \pi_{j|i}\pi_i$. Tal esquema pode ser informativo se as probabilidades π_i e $\pi_{j|i}$ forem relacionadas aos termos de erro $\mathbf{u}_i$ e ε_{ij}, logo também à variável resposta y_{ij}.

4.4 – Estimação por Mínimos Quadrados Generalizados Iterativo Ponderado pelas Probabilidades (MQGIPP)

Para estimar os parâmetros diretamente pelo método da máxima pseudo-verossimilhança (MPV) seria preciso escrever uma função de verossimilhança considerando que os dados são provenientes de um censo, estimar seu logaritmo e maximizar esta função numericamente. Por simplicidade na estimação e eficiência computacional, Pfeffermann *et al*. (1998a), optaram por usar uma adaptação do método MQGI por analogia ao método da MPV. O método MQGI alterna, iterativamente, entre a estimação de $\boldsymbol{\beta}$ e $(\boldsymbol{\Omega}, \sigma^2)$, sendo equivalente ao método de máxima verossimilhança no caso padrão, sob hipótese de normalidade (Goldstein, 1995).

Primeiramente são escritas as expressões dos estimadores de $\boldsymbol{\beta}$ e $(\boldsymbol{\Omega},\sigma^2)$ como se todos os elementos da população fossem observados e então estes estimadores são trocados por estimadores amostrais ponderados.

Considere o algoritmo de MQGI para o caso hipotético onde todos os valores (y_{ij}, $\mathbf{x}_{ij}$, $\mathbf{z}_i$ e z_{0ij}) são observados para todas as unidades da população *j=1,...,Mi* e *i=1,...,N*. Seja $Y_i = (y_{i1},...,y_{iM_i})'$, $X_i = (\mathbf{x}_{i1}^{'},...,\mathbf{x}_{iM_i}^{'})'$ e $\mathbf{e}_i = (e_{i1},...,e_{iM_i})'$, onde $e_{ij} = \mathbf{z}_i^{'}\mathbf{u}_i + z_{0ij}\varepsilon_{ij}$. O modelo em (4.5) pode ser definido em notação matricial como:

$$Y_i = X_i\boldsymbol{\beta} + \mathbf{e}_i, \qquad \mathbf{e}_i \sim N(\mathbf{0},V_i) \tag{4.6}$$

onde $V_i = Z_i\boldsymbol{\Omega}Z_i^{'} + \sigma^2 D_i$, $Z_i = (\mathbf{z}_{i1},...,\mathbf{z}_{iM_i})'$ e $D_i = diag(z_{0i1}^2,...,z_{0iM_i}^2)'$.

Seja $\boldsymbol{\theta} = (\theta_1,...,\theta_s)'$ um vetor de dimensão $s = \frac{q(q+1)}{2} + 1$ contendo os elementos distintos de $\boldsymbol{\Omega}$ e σ^2, com $\theta_s = \sigma^2$. Então V_i pode ser expresso como uma função linear de $\boldsymbol{\theta}$, da forma:

$$V_i = \sum_{k=1}^{s} \theta_k G_{ki}$$

onde

$G_{ki} = Z_i H_{ki} Z_i^{'} + \delta_{ks} D_i$, H_{ki} é uma matriz conhecida, de dimensão *q* x *q*, com valores 0 ou 1;

$$\delta_{ks} = \begin{cases} 0, & se\ k \neq s \\ 1, & se\ k = s \end{cases}.$$

Escrevendo $\mathbf{e}_i(\boldsymbol{\beta}) = (Y_i - X_i\boldsymbol{\beta})(Y_i - X_i\boldsymbol{\beta})'$, é possível mostrar que $\mathbf{e}_i(\boldsymbol{\beta})$ tem esperança V_i. O algoritmo MQGI para estimar $\boldsymbol{\beta}$, $\boldsymbol{\Omega}$ e σ^2 no modelo de superpopulação (4.5) envolve o cálculo de uma seqüência de estimadores para $\boldsymbol{\beta}$ e $\boldsymbol{\theta}$, imaginando que todas as unidades da população tivessem sido observadas como em um censo, denotados por $\boldsymbol{\beta}_C^{(t)}$ e $\boldsymbol{\theta}_C^{(t)}$ para *t=1,2,...*, como segue:

Passo 1: faça $\boldsymbol{\beta}_C^{(t)} = \left(P^{(t)}\right)^{-1} Q^{(t)}$

onde

$$P^{(t)} = \sum_{i=1}^{N} X_i' V_{it}^{-1} X_i \; ;$$

$$Q^{(t)} = \sum_{i=1}^{N} X_i' V_{it}^{-1} Y_i \; ;$$

$$V_{it} = V_i(\hat{\boldsymbol{\theta}}_C^{(t-1)}) \, .$$

Passo 2: faça $\boldsymbol{\theta}_C^{(t)} = R^{(t)-1} S^{(t)}$,

onde o *kl*-ésimo elemento da matriz $R^{(t)}$ de dimensão *s* x *s* é $\sum_{i=1}^{N} tr(V_{it}^{-1} G_{ki} V_{it}^{-1} G_{li})$ e o *k*-ésimo elemento do vetor $S^{(t)}$ de dimensão *s* é $\sum_{i=1}^{N} tr\{V_{it}^{-1} G_{ki} V_{it}^{-1} e_i(\hat{\boldsymbol{\beta}}^{(t)})\}$.

O processo é inicializado para algum valor $\boldsymbol{\theta}_C^{(0)}$ e, sob condições de regularidade (Pfeffermann *et al.*, 1998a), $\boldsymbol{\beta}_C^{(t)}$ e $\boldsymbol{\theta}_C^{(t)}$ convergem para $\boldsymbol{\beta}_C$ e $\boldsymbol{\theta}_C$ quando $t \to \infty$.

A partir daqui, como os dados disponíveis são de uma amostra, estas expressões devem ser substituídas por estimadores amostrais $(\hat{P}^{(t)}, \hat{Q}^{(t)}, \hat{R}^{(t)}, \hat{S}^{(t)})$, com $\boldsymbol{\beta}$ e $\boldsymbol{\theta}$ sendo estimados por $\hat{\boldsymbol{\beta}}$ e $\hat{\boldsymbol{\theta}}$, com:

Passo 1:

$$\hat{\boldsymbol{\beta}}^{(t)} = \left(\hat{P}^{(t)}\right)^{-1} \hat{Q}^{(t)} \quad \text{para} \quad t=1,2,... \tag{4.7}$$

Passo 2:

$$\hat{\boldsymbol{\theta}}^{(t)} = \hat{R}^{(t)-1} \hat{S}^{(t)} \quad \text{para} \quad t=1,2,... \tag{4.8}$$

Se $(\hat{P}^{(t)}, \hat{Q}^{(t)}, \hat{R}^{(t)}, \hat{S}^{(t)})$ são tomados como versões amostrais de $(P^{(t)}, Q^{(t)}, R^{(t)}, S^{(t)})$, $\hat{\boldsymbol{\beta}}$ e $\hat{\boldsymbol{\theta}}$ são estimadores de MQGI não ponderados. Tais estimadores ignoram o plano amostral, o que implica que devem ser adaptados para considerar as informações da amostra de modo a estimar consistentemente as quantidades da população finita.

A abordagem proposta consiste em substituir cada soma sobre as unidades da população de nível 2 por uma soma das unidades de nível 2 da amostra ponderada por $w_i = \pi_i^{-1}$. Analogamente substituir as somas sobre as unidades de nível 1 na população por somas das unidades da amostra de nível 1 ponderadas por $w_{j|i} = \pi_{j|i}^{-1}$. As quantidades π_i e $\pi_{j|i}$ correspondem às probabilidades de seleção das unidades de nível 2 e nível 1, respectivamente. Os estimadores resultantes são chamados de estimadores de mínimos quadrados generalizados iterativo ponderados pelas probabilidades (MQGIPP).

Para obter estes estimadores, inicialmente formula-se $(P^{(t)}, Q^{(t)}, R^{(t)}, S^{(t)})$ como funções de somas sobre *i* e *j*. Por simplicidade será considerado aqui o caso para *q = 1*, isto é,

para apenas um efeito aleatório no nível 2. Uma generalização para $q > 1$ pode ser encontrada em Pfeffermann *et al.* (1998a, apêndice A). Quando $q = 1$ tem-se que:

$$\begin{aligned} P^{(t)} &= \sum_{i=1}^{N} T_{i1} - a_i T_{i2} T_{i2}^{'} \\ Q^{(t)} &= \sum_{i=1}^{N} T_{i3} - a_i T_{i2} T_{i4} \end{aligned} \tag{4.9}$$

onde $T_{i1} = \sum_{j=1}^{M_i} \frac{\mathbf{x}_{ij}^{'}\mathbf{x}_{ij}}{z_{0ij}^2}$, $T_{i2} = \sum_{j=1}^{M_i} \frac{\mathbf{x}_{ij}^{'}\mathbf{z}_{i}^{'}}{z_{0ij}^2}$, $T_{i3} = \sum_{j=1}^{M_i} \frac{\mathbf{x}_{ij}^{'}y_{ij}}{z_{0ij}^2}$, $T_{i4} = \sum_{j=1}^{M_i} \frac{y_{ij}\mathbf{z}_{i}^{'}}{z_{0ij}^2}$, $a_i = \left(T_{i5} + \frac{\hat{\sigma}^2}{\hat{\tau}_0^2} \right)$,

$T_{i5} = \sum_{j=1}^{M_i} \frac{\mathbf{z}_i^{'}\mathbf{z}_i}{z_{0ij}^2}$, $\hat{\tau}_0^2$ e $\hat{\sigma}^2$ são os estimadores por MQGI da iteração $t-1$, considerando que os dados para toda a população sejam conhecidos.

Similarmente, tem-se:

$$R^{(t)} = \begin{pmatrix} \sum_{i=1}^{N} b_i^2 & \sum_{i=1}^{N} \frac{b_i^2}{T_{i5}} \\ \sum_{i=1}^{N} \frac{b_i^2}{T_{i5}} & \sum_{i=1}^{N} \left\{ \hat{\sigma}^{-4}(M_i - 1) + \frac{b_i^2}{T_{i5}^2} \right\} \end{pmatrix}, \; S^{(t)} = \begin{pmatrix} \sum_{i=1}^{N} b_i^2 \tilde{\mathbf{u}}_i^2 \\ \sum_{i=1}^{N} \left(\hat{\sigma}^{-4} T_{i6} + \frac{b_i^2 \tilde{\mathbf{u}}_i^2}{T_{i5}} \right) \end{pmatrix} \tag{4.10}$$

onde $b_i = \left(\hat{\tau}_0^2 + \frac{\hat{\sigma}^2}{T_{i5}} \right)^{-1}$, $T_{i6} = \sum_{j=1}^{M_i} \tilde{\varepsilon}_{ij}^2$, $\tilde{\mathbf{u}}_i = \left(\sum_{j=1}^{M_i} \frac{e_{ij}\mathbf{z}_i^{'}}{z_{0ij}^2} \right) \Big/ T_{i5}$, $\tilde{\varepsilon}_{ij} = \frac{\left(e_{ij} - \mathbf{z}_i^{'}\tilde{\mathbf{u}}_i \right)}{z_{0ij}}$ e

$e_{ij} = y_{ij} - \mathbf{x}_{ij}^{'}\hat{\boldsymbol{\beta}}^{(t)}$.

Os estimadores de MQGIPP são obtidos substituindo-se as somas na população da forma $\sum_{i=1}^{N} d_i$ e $\sum_{j=1}^{M_i} d_{ij}$ pelas respectivas somas na amostra $\sum_{i=1}^{n} w_i d_i$ e $\sum_{j=1}^{m_i} w_{j|i} d_{ij}$, onde w_i e $w_{j|i}$ são os pesos da unidade de nível 2 e os pesos condicionais das unidades de nível 1,

respectivamente. As somas amostrais ponderadas são não viesadas e consistentes para as correspondentes somas na população sob a distribuição de aleatorização induzida pelo processo de amostragem. A estimativa para M_i em (4.10) é dada por $\hat{M}_i = \sum_{j=1}^{m_i} w_{j|i}$, mesmo quando o valor é conhecido, pois em estudos de simulação feitos por Pfeffermann *et al.* (1998a), o uso de M_i leva a estimativas viesadas de $\hat{\sigma}^2$.

4.5 – Método do Modelo de Distribuição Amostral

O uso do procedimento de ponderação pelas probabilidades de inclusão na amostra (MQGIPP), proposto por Pfeffermann *et a.l.* (1998a) para ajustar o efeito de amostragem não ignorável na estimação de parâmetros num modelo multinível tem quatro limitações importantes:

1. As variâncias dos estimadores ponderados são geralmente maiores que as dos estimadores correspondentes não ponderados.
2. A inferência é quase que restrita à estimação pontual. Afirmações probabilísticas requerem pressupostos de normalidade assintótica, já que a distribuição exata dos estimadores pontuais ponderados geralmente não pode ser obtida.
3. O uso de pesos amostrais não permite, em geral, condicionar na amostra selecionada de grupos (unidades de 2º nível ou superior) ou valores das variáveis explicativas do modelo.
4. Não é claro como predizer os efeitos aleatórios das unidades de segundo nível sob amostragem informativa; por exemplo, como prever o escore médio de um conglomerado que não está na amostra. Sob amostragem informativa os

conglomerados não selecionados também formam uma amostra informativa com comportamento diferente do da população.

Uma abordagem alternativa para ajustar modelos multiníveis em amostragem informativa foi proposta por Pfeffermann, Moura e Silva (2001). O método parte da idéia de extrair o modelo hierárquico associado aos dados da amostra como função do modelo populacional e das probabilidades de inclusão de primeira ordem, para então ajustar o modelo amostral. Com uma especificação correta deste modelo, superam-se as limitações referentes ao método de ponderação pelas probabilidades.

Segundo Pfeffermann, Krieger e Rinnot (1998), a distribuição amostral é definida como a distribuição de alguma medida obtida a partir dos dados da amostra selecionada. Sob amostragem informativa, esta distribuição é diferente da correspondente distribuição populacional. Entretanto, geralmente ambas pertencem a uma mesma família de distribuições, diferindo apenas nos parâmetros que as caracterizam.

Em geral é possível aproximar a distribuição paramétrica de medidas amostrais e usar esta aproximação para fazer inferências na distribuição populacional correspondente, explorando a relação existente entre as duas distribuições.

4.5.1 – Modelo Populacional

Seja o modelo populacional definido em (4.5), mas apenas com interceptos aleatórios e representado por:

modelo de nível 1

$$y_{ij} \mid \beta_{0i} = \beta_{0i} + \mathbf{x}_{ij}'\boldsymbol{\beta} + \varepsilon_{ij} \;;\; \varepsilon_{ij} \sim N(0, \sigma^2),\; i = 1, \ldots, N,\; j = 1, \ldots, M_i \qquad (4.11)$$

modelo de nível 2

$$\beta_{0i} = \mathbf{z}_i'\boldsymbol{\gamma} + u_i \;;\; u_i \sim N(0, \tau_0),\; i = 1 \ldots N \tag{4.12}$$

As respectivas funções de densidade de probabilidades para os dois níveis são dadas por:

$$f_p(y_{ij} \mid \mathbf{x}_{ij}, \boldsymbol{\theta}_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{(y_{ij} - \beta_{0i} - \mathbf{x}_{ij}'\boldsymbol{\beta})^2}{2\sigma^2} \right\} \tag{4.13}$$

e

$$f_p(\beta_{0i} \mid \mathbf{z}_i, \boldsymbol{\lambda}) = \frac{1}{\sqrt{2\pi\tau_0}} \exp\left\{ -\frac{(\beta_{0i} - \mathbf{z}_i'\boldsymbol{\gamma})^2}{2\tau_0} \right\} \tag{4.14}$$

onde $\boldsymbol{\theta}_i = (\beta_{0i}, \boldsymbol{\beta}', \sigma^2)'$ e $\boldsymbol{\lambda} = (\boldsymbol{\gamma}', \tau_0)'$.

4.5.2 – Modelo Amostral

Considere a seleção de uma amostra por um procedimento em dois estágios. Neste processo, *n* unidades de nível 2 são selecionadas com probabilidade $\pi_i = \Pr(i \in s)$ no primeiro estágio e m_i unidades de nível 1 são selecionadas da unidade *i* de nível 2 no segundo estágio, com probabilidades $\pi_{j|i} = \Pr(j \in s_i \mid i \in s)$. O conjunto s_i define a amostra de unidades do nível 1 selecionadas da unidade *i* de nível 2. Segundo Pfeffermann, Moura e Silva (2001), as distribuições amostrais das unidades de primeiro e segundo níveis são dadas respectivamente por:

$$f_s(y_{ij} \mid \mathbf{x}_{ij}, \boldsymbol{\theta}_i) = f(y_{ij} \mid \mathbf{x}_{ij}, \boldsymbol{\theta}_i, j \in s_i \text{ e } i \in s) = \frac{E_p(\pi_{j|i} \mid y_{ij}, \mathbf{x}_{ij}, \boldsymbol{\theta}_i) f_p(y_{ij} \mid \mathbf{x}_{ij}, \boldsymbol{\theta}_i)}{E_p(\pi_{j|i} \mid \mathbf{x}_{ij}, \boldsymbol{\theta}_i)} \tag{4.15}$$

$$f_s(\beta_{0i} \mid \mathbf{z}_i, \boldsymbol{\lambda}) = f(\beta_{0i} \mid \mathbf{z}_i, \boldsymbol{\lambda}, i \in s) = \frac{E_p(\pi_i \mid \beta_{0i}, \mathbf{z}_i, \boldsymbol{\lambda}) f_p(\beta_{0i} \mid \mathbf{z}_i, \boldsymbol{\lambda})}{E_p(\pi_i \mid \mathbf{z}_i, \boldsymbol{\lambda})} \tag{4.16}$$

onde

$f_p(\cdot)$: distribuição populacional com esperança $E_p(\cdot)$;

$f_s(\cdot)$: distribuição amostral com esperança $E_s(\cdot)$.

O modelo apresentado nas equações (4.15) e (4.16) é associado aos dados da amostra e depende do modelo populacional e das probabilidades de seleção de primeira ordem das unidades de nível 1 e de nível 2. As esperanças que figuram neste modelo podem ser modeladas com base no conhecimento do processo de amostragem e dos dados da amostra. Na prática tais esperanças não possuem uma forma conhecida, particularmente quando as probabilidades de seleção dependem também de variáveis do desenho amostral que não estão entre as variáveis explicativas.

Ao assumir que as probabilidades de seleção sejam conhecidas pelo menos na amostra, a forma das esperanças pode ser identificada e estimada em princípio pelos dados da amostra. Segundo Pfeffermann, Krieger e Rinott (1998), em condições regulares tais esperanças podem ser aproximadas por polinômios de baixa ordem em y_{ij} (ou β_{0i}) e nos componentes de x_{ij}, ou ainda por exponenciais destes polinômios.

No caso específico da aplicação desta dissertação, a Pesquisa sobre Padrões de Vida – PPV foi realizada por meio de um levantamento amostral em dois estágios. No primeiro

estágio foram selecionados setores censitários com probabilidades proporcionais ao tamanho dos setores; no segundo estágio foram selecionados domicílios com eqüiprobabilidade, conforme descrito na seção 2.2.1. Desta forma, a especificação das esperanças em (4.15) não se faz necessária, uma vez que as unidades de nível 1 foram selecionadas por um plano amostral ignorável o que torna $E_p(\pi_{j|i} \mid y_{ij}, \mathbf{x}_{ij}, \boldsymbol{\theta}_i) = \Pr(j \in s_i \mid y_{ij}, \mathbf{x}_{ij}, \boldsymbol{\theta}_i) = \Pr(j \in s_i \mid \mathbf{x}_{ij}, \boldsymbol{\theta}_i) = E_p(\pi_{j|i} \mid \mathbf{x}_{ij}, \boldsymbol{\theta}_i)$. Neste caso a distribuição amostral em (4.15) será a mesma que a distribuição populacional em (4.13).

A especificação das esperanças em (4.16) foi feita de maneira análoga ao estudo de Pfeffermann, Moura e Silva (2001), considerando o tamanho dos setores se distribuindo como:

$$\log(M_i) \sim N[\alpha_0 + \alpha_1 \beta_{0i}, \sigma_M^2]. \tag{4.17}$$

Neste estudo, as esperanças condicionais em (4.16) foram modeladas como uma distribuição lognormal, uma vez que as unidades de nível 1 foram selecionadas com probabilidades proporcionais ao tamanho de setores censitários e estes tamanhos se distribuíam aproximadamente como uma lognormal na população. Desta forma as esperanças em (4.16) foram especificadas da forma:

$$E_p(\pi_i \mid \beta_{0i}, \mathbf{z}_i, \boldsymbol{\lambda}) \cong \frac{n}{NE_p(M_i)} \exp\left[\alpha_0 + \alpha_1 \beta_{0i} + \frac{\sigma_M^2}{2}\right] \tag{4.18}$$

$$E_p(\pi_i \mid \mathbf{z}_i, \boldsymbol{\lambda}) \cong \frac{n}{NE_p(M_i)} \exp\left[\alpha_0 + \alpha_1 \mathbf{z}_i' \boldsymbol{\gamma} + \frac{\alpha_1^2 \tau_0 + \sigma_M^2}{2}\right] \tag{4.19}$$

Note que o termo $\frac{n}{NE_p(M_i)}$ é cancelado em (4.16).

A principal vantagem de basear a inferência na abordagem do modelo de distribuição amostral é a possibilidade de usar procedimentos padrões de inferência, podendo estimar os parâmetros por máxima verossimilhança ou outros métodos. Nesta dissertação foram utilizados métodos de simulação com base no paradigma Bayesiano para a estimação dos parâmetros de interesse.

4.6 – Métodos Bayesianos de Simulação

Os recentes avanços na capacidade de processamento e de armazenamento computacional vêm permitindo o maior desenvolvimento e utilização de métodos de simulação. Tais métodos são particularmente úteis em análises Bayesianas devido à relativa facilidade de gerar amostras de uma distribuição de probabilidade a posteriori, particularmente quando a função de densidade não pode ser explicitamente integrada.

4.6.1 – Inferência Bayesiana

A análise de dados Bayesiana é caracterizada por um conjunto de métodos para fazer inferência usando modelos de probabilidade para quantidades que não são observadas, mas que se deseja conhecer, e para a relação destas com os dados observados. Segundo Gelman *et. al.* (1995), entende-se por inferência Bayesiana o processo de ajustar modelos de probabilidade a um conjunto de dados e sumariar os resultados por uma distribuição de probabilidade nos parâmetros e nas quantidades que ainda não observadas.

Os métodos Bayesianos têm como característica principal o uso explícito de probabilidade para quantificar incerteza em inferências. Seu processo de análise de dados pode ser dividido nos três passos seguintes:

1. Definição do modelo de probabilidade completo, ou seja, da distribuição de probabilidade conjunta para todas as quantidades observáveis e não observáveis no problema.
2. Cálculo da distribuição a posteriori das quantidades não observáveis com base nos dados observados.
3. Avaliação do ajuste, fazendo as modificações necessárias no modelo.

A liberdade de ajustar modelos complexos é uma vantagem do paradigma Bayesiano, que fornece métodos conceitualmente simples para lidar com vários parâmetros. As conclusões acerca dos parâmetros são feitas em termos de probabilidades. Para isto é necessário explicitar o modelo com a distribuição de probabilidade conjunta para os parâmetros θ e os dados observados y. A função de densidade pode ser escrita como o produto de duas densidades que são referidas como distribuição a priori $p(\theta)$ e distribuição amostral $p(y \mid \theta)$:

$$p(\theta, y) = p(\theta)p(y \mid \theta) \tag{4.20}$$

Aplicando a regra de Bayes chega-se à distribuição a posteriori dos parâmetros condicionada aos dados observados:

$$p(\theta \mid y) = \frac{p(\theta, y)}{p(y)} = \frac{p(\theta)p(y \mid \theta)}{p(y)} \tag{4.21}$$

onde

$p(y) = \sum_{\theta} p(\theta)p(y \mid \theta)$, se θ é variável aleatória discreta;

$p(y) = \int p(\theta)p(y \mid \theta)\, d\theta$, caso contrário.

A expressão (4.21) pode ser reescrita omitindo o fator $p(y)$ que não depende de θ, levando a uma distribuição a posteriori não normalizada:

$$p(\theta \mid y) \propto p(\theta)p(y \mid \theta) \tag{4.22}$$

Uma vez desenvolvido o modelo $p(\theta \mid y)$ deve-se executar os cálculos necessários para sumariar ou descrever $p(\theta \mid y)$ de maneira apropriada, calculando-se a média, intervalos de credibilidade e outras medidas que possam representar a distribuição inteira.

4.6.2 – Métodos de simulação

Muitos problemas em inferência Bayesiana são tratados com métodos de simulação, que tipicamente envolvem a geração de observações de uma distribuição de probabilidade e o uso da amostra resultante para aproximar funções e, desta forma, obter as medidas resumo de interesse. Para isto, é necessário um número razoavelmente grande de simulações para que tal aproximação convirja para a distribuição alvo. Em geral, as únicas razões que limitam o número de simulações são o tempo de processamento e a capacidade de armazenamento, o que vem se tornando um problema não significativo com os avanços computacionais.

É possível expressar a integral de uma função qualquer como uma média de uma distribuição a posteriori:

$$E(h(\theta) \mid y) = \int h(\theta) p(\theta \mid y) d\theta \tag{4.23}$$

Desta forma, é possível estimar a integral desta função pela estimativa da média da amostra simulada:

$$\hat{E}(h(\theta) \mid y) = \frac{1}{n} \sum_{t=1}^{n} h(\theta^t) \tag{4.24}$$

onde $\theta^1, \ldots, \theta^n$ são observações geradas da distribuição $p(\theta \mid y)$.

A idéia da simulação de cadeias de Markov é simular um caminho aleatório no espaço de θ, o qual converge através de um número suficiente de simulações para uma distribuição estacionária que é a distribuição a posteriori conjunta $p(\theta \mid y)$, a distribuição alvo. O método se caracteriza como uma cadeia de Markov, pois as amostras são geradas de forma seqüencial, com a distribuição de cada nova amostra gerada dependendo apenas do último valor gerado. Segundo Gelman *et al.* (1995), os dois métodos mais utilizados em simulações de cadeias de Markov são o algoritmo de Metropolis generalizado e o Amostrador de Gibbs.

4.6.2.1 – O algoritmo de Metropolis

Segundo Gelman *et. al.* (1995), o algoritmo de Metropolis foi introduzido no artigo de Metropolis *et al*. (1953). O método cria uma seqüência de pontos aleatórios $(\theta^1, \theta^2, ...)$ cujas distribuições convergem para a distribuição alvo $p(\theta \mid y)$. Cada distribuição pode ser considerada como um caminho aleatório cuja distribuição estacionária é $p(\theta \mid y)$. O algoritmo prossegue da seguinte forma:

1. Gerar um ponto de partida θ^0, para o qual $p(\theta^0 \mid y) > 0$, de uma distribuição inicial $p_0(\theta)$.
2. Para $t = 1,2,...$
   a) Amostrar um ponto candidato $\theta*$ da distribuição de salto no tempo t, $J_t(\theta* \mid \theta^{t-1})$. A distribuição de salto deve ser simétrica, isto é, $J_t(\theta_a \mid \theta_b) = J_t(\theta_b \mid \theta_a)$ para todo θ_a, θ_b e t.
   b) Calcular a razão das densidades:

$$r = \frac{p(\theta* \mid y)}{p(\theta^{t-1} \mid y)} \tag{4.25}$$

   c) Calcular a probabilidade de aceitação $\min(r,1)$ e gerar $u \sim U(0,1)$.
   d) Se $u \leq r$ aceitar o novo valor e fazer $\theta^t = \theta*$, caso contrário rejeitar o novo valor e fazer $\theta^t = \theta^{t-1}$.

4.6.2.2 – O algoritmo de Metropolis-Hastings

O algoritmo de Metropolis-Hastings é uma generalização do algoritmo de Metropolis, diferindo deste último em duas maneiras:

1. A distribuição de salto não necessariamente deve ser simétrica, isto é não há a necessidade de que $J_t(\theta_a \mid \theta_b) = J_t(\theta_b \mid \theta_a)$.
2. Para corrigir a assimetria na distribuição de salto, a razão r em (4.25) passa a ser:

$$r = \frac{p(\theta^* \mid y)/J_t(\theta^* \mid \theta^{t-1})}{p(\theta^{t-1} \mid y)/J_t(\theta^{t-1} \mid \theta^*)} \tag{4.26}$$

Ao permitir que a distribuição de salto seja assimétrica pode haver um aumento na velocidade do caminho aleatório.

A distribuição de salto ideal no algoritmo de Metropolis-Hastings é a própria distribuição alvo, isto é, $J_t(\theta^* \mid \theta) = p(\theta^* \mid y)$ para todo θ. Neste caso a razão r é sempre exatamente igual a 1 e as iterações θ^t são uma seqüência de amostras independentes de $p(\theta \mid y)$. Cabe ressaltar que o uso de métodos iterativos de simulação justifica-se em casos em que a geração não-iterativa for muito complicada.

4.6.2.3 – O amostrador de Gibbs

O amostrador de Gibbs é uma técnica que permite gerar variáveis aleatórias de uma distribuição sem precisar calcular sua densidade, utilizando apenas propriedades elementares de cadeias de Markov. Segundo Gelman *et. al.* (1995), a técnica surgiu mais popularmente com o artigo de Geman e Geman (1984) no estudo de modelos de processamento de imagens e mais recentemente, Gelfand e Smith (1990) provocaram novo interesse no amostrador de Gibbs explicitando seu uso potencial em diversos problemas estatísticos.

Não existe mecanismo de aceitação-rejeição no amostrador de Gibbs. Todos os saltos são aceitos, com a cadeia sempre se movendo para um novo valor. Suponha que θ pode ser decomposto em d componentes $\theta = (\theta_1, \ldots, \theta_d)$, então as transições são feitas de acordo com as distribuições condicionais completas $p(\theta_j \mid \theta_{-j}^{t-1}, y)$, onde θ_{-j}^{t-1} representa todos os outros componentes de θ exceto θ_j: $\theta_{-j}^{t-1} = (\theta_1^t, \ldots, \theta_{j-1}^t, \theta_{j+1}^{t-1}, \ldots, \theta_d^{t-1})$.

Nos casos em que é complicado amostrar diretamente de $p(\theta \mid y)$, mas com as distribuições condicionais completas conhecidas, pode-se usar o amostrador de Gibbs pelos seguintes passos:

1. Gerar o ponto de partida $\theta^0 = \left(\theta_1^0, \ldots, \theta_d^0\right)$.
2. Obter um novo valor de θ^t a partir de θ^{t-1} através da geração sucessiva dos valores:

$$
\begin{aligned}
\theta_1^t &\sim p\left(\theta_1 \mid \theta_2^{t-1}, \theta_3^{t-1}, \ldots, \theta_d^{t-1}, y\right) \\
\theta_2^t &\sim p\left(\theta_2 \mid \theta_1^{t-1}, \theta_3^{t-1}, \ldots, \theta_d^{t-1}, y\right) \\
&\vdots \\
\theta_d^t &\sim p\left(\theta_d \mid \theta_1^{t-1}, \theta_2^{t-1}, \ldots, \theta_{d-1}^{t-1}, y\right)
\end{aligned}
$$

Após a convergência os valores resultantes formam uma amostra de $p(\theta \mid y)$. Para proceder com este algoritmo é necessário que todas as distribuições condicionais completas sejam conhecidas. Para determinados problemas a amostragem para algumas, ou mesmo todas distribuições condicionais, não é possível, mas pode-se construir aproximações em que a amostragem é possível. Nestes casos o amostrador de Gibbs utiliza passos de Metropolis para realizar tal aproximação.

4.6.2.4 – Verificação de Convergência

Para se fazer inferência com base em métodos de simulação iterativa é necessário verificar se a convergência para a distribuição alvo foi atingida. Tal verificação pode ser feita informalmente através de análise gráfica, onde os valores simulados para uma determinada quantidade de interesse são plotados contra a iteração *t*. À medida que *t* aumenta os valores

simulados não devem variar muito. É difícil afirmar conclusivamente que uma cadeia convergiu, mas pode-se afirmar quando esta não converge.

Mesmo após a convergência, as iterações do início do processo são mais próximas aos valores iniciais do que da distribuição de interesse. Para diminuir o efeito da distribuição inicial, uma prática adotada é descartar a primeira metade das simulações de cada seqüência e fazer inferências usando somente a segunda metade. Neste caso há o pressuposto de que os valores simulados θ^t, para t suficientemente grande, são amostras independentes e identicamente distribuídas (iid) da distribuição alvo $p(\theta \mid y)$.

Além da análise gráfica, existem outros métodos para diagnosticar a convergência das cadeias. Um teste comumente utilizado é o proposto por Gelman *et al.* (1998), tendo como base a comparação de *C* diferentes seqüências simuladas de tamanho *T*. Se a variância entre as cadeias for menor que a variância dentro das cadeias então há indícios de convergência. O procedimento deve ser realizado para cada parâmetro de interesse θ separadamente. As variâncias entre as cadeias (V_E) e dentro das cadeias (V_D) são dadas respectivamente por:

$$V_E = \frac{T}{C-1}\sum_{c=1}^{C}\left(\bar{\theta}_{.c} - \bar{\theta}_{..}\right)^2 \tag{4.27}$$

onde

$\bar{\theta}_{.c} = \frac{1}{T}\sum_{t=1}^{T}\theta_{tc}$, com $t = 1,\ldots,T \quad c = 1,\ldots,C$;

$\bar{\theta}_{..} = \frac{1}{C}\sum_{c=1}^{C}\bar{\theta}_{.c}$; e

$$V_D = \frac{1}{C}\sum_{c=1}^{C}s_c^2 \tag{4.28}$$

onde $s_c^2 = \frac{1}{T-1}\sum_{t=1}^{T}\left(\theta_{tc} - \overline{\theta}_{.c}\right)^2$.

Uma estimativa da variância marginal a posteriori de θ pode ser obtida por uma média ponderada de V_E e V_D, dada por:

$$\hat{\text{var}}^{+}(\theta \mid y) = \frac{T-1}{T}V_D + \frac{1}{T}V_E. \tag{4.29}$$

O valor de $\hat{\text{var}}^{+}(\theta \mid y)$ em (4.29) superestima a variância marginal a posteriori assumindo que a cadeia não convergiu. Entretanto, este estimador é não-viesado se a distribuição inicial for igual a distribuição alvo ou quando $T \to \infty$. Uma medida para aferir a convergência de uma simulação é dada por um fator que reduz a escala da distribuição atual de θ quando $T \to \infty$. Este fator é dado por:

$$\sqrt{\hat{R}} = \sqrt{\frac{\hat{\text{var}}^{+}(\theta \mid y)}{V_D}} \tag{4.30}$$

O valor de $\sqrt{\hat{R}}$ tende para 1 quando $T \to \infty$. Se $\sqrt{\hat{R}}$ não estiver próximo de 1 para todos os parâmetros de interesse então as simulações devem continuar. Na prática o quão próximo $\sqrt{\hat{R}}$ está de 1 depende do problema em estudo, na maioria das vezes valores menores que 1,2 são aceitáveis.

No uso do *software* WinBUGS, Spiegelhalter *et al.* (2003) aconselham que quando a convergência for satisfeita, as simulações devem ser executadas até que o Erro de Monte Carlo seja menor que 5% do desvio padrão da amostra simulada. O Erro de Monte Carlo

fornece uma estimativa do erro padrão de Monte Carlo da média $\sigma/\sqrt{T}$. O método para estimar σ pode ser encontrado em Roberts[2] (1996, p.50 apud Spiegelhalter *et al.,* 2003).

4.7 – Método do Modelo de Distribuição Amostral via Monte Carlo com Cadeias de Markov (MCMC)

Fazendo $Y = \{y_{ij} : i = 1...n;\ j = 1...m_i\}$ os valores observados de y e $M = \{M_i : i = 1...n\}$ os tamanhos dos setores censitários na amostra, pode-se denotar os dados observados como $D_{obs} = (Y, M)$. As distribuições conjuntas das observações D_{obs}, dos parâmetros $(\beta_{0i}, \boldsymbol{\beta}, \tau_u, \tau_\varepsilon)$ que indexam o modelo populacional e dos parâmetros adicionais $(\boldsymbol{\alpha}, \tau_M)$ que aparecem no modelo amostral podem ser escritas como:

$$\begin{aligned} &f_s(D_{obs}, \{\beta_{0i}\}, \beta, \alpha, \tau_u, \tau_\varepsilon, \tau_M \mid X, Z) \cong \prod_{i=1}^{n} \prod_{j=1}^{m_i} f_s(y_{ij} \mid \mathrm{x}_{ij}, \beta_{0i}, \beta, \tau_\varepsilon). \\ &f_s(M_i \mid \mathbf{z}_i, \beta_{0i}, \boldsymbol{\alpha}, \tau_M).f_s(\beta_{0i} \mid \mathbf{z}_i, \boldsymbol{\gamma}, \boldsymbol{\alpha}, \tau_u).p(\boldsymbol{\beta}).p(\boldsymbol{\gamma})p(\boldsymbol{\alpha})p(\tau_u)p(\tau_\varepsilon)p(\tau_{\mathbf{M}}) \end{aligned} \tag{4.31}$$

onde $X = (\mathbf{x}_{11}, \ldots, \mathbf{x}_{nm_i})'$ e $Z = (\mathbf{z}_1, \ldots, \mathbf{z}_n)'$.

Para ficar em conformidade com a convenção do *software* WinBUGS as variâncias τ_0, σ^2 e σ_M^2 foram substituídas pelas funções $\tau_u = 1/\tau_0$, $\tau_\varepsilon = 1/\sigma^2$ e $\tau_M = 1/\sigma_M^2$, respectivamente. As distribuições amostrais $f_s(y_{ij} \mid \mathbf{x}_{ij}, \beta_{0i}, \boldsymbol{\beta}, \tau_\varepsilon)$ e $f_s(\beta_{0i} \mid \mathbf{z}_i, \boldsymbol{\gamma}, \boldsymbol{\alpha}, \tau_u)$ são definidas em (4.15) e (4.16), respectivamente. A distribuição amostral dos tamanhos dos setores censitários é definida similarmente à equação (4.15), como:

---

[2] ROBERTS, G.O. Markov chain concepts related to sampling algorithms. In W.R. Gilks, S Richardson and D.J.

$$f_s(M_i \mid \mathbf{z}_i, \beta_{0i}, \boldsymbol{\alpha}, \sigma_M^2) = \frac{E_p(\pi_i \mid M_i, \mathbf{z}_i, \beta_{0i}, \boldsymbol{\alpha}, \tau_M) f_p(M_i \mid \mathbf{z}_i, \beta_{0i}, \boldsymbol{\alpha}, \sigma_M^2)}{E_p(\pi_i \mid \mathbf{z}_i, \beta_{0i}, \boldsymbol{\alpha}, \sigma_M^2)} \tag{4.32}$$

Por Pfeffermann *et al.* (1998b), a distribuição condicional em (4.32) sob o desenho amostral da PPV é novamente lognormal e:

$$f_s[\log(M_i) \mid \mathbf{z}_i, \beta_{0i}, \boldsymbol{\alpha}, \sigma_M^2] = N[\alpha_0 + \alpha_1 \beta_{0i} + \sigma_M^2, \sigma_M^2] \tag{4.33}$$

Note que a única diferença entre a distribuição populacional na equação (4.17) e a amostral em (4.33) é a adição do termo de variância σ_M^2 à média.

As distribuições a posteriori condicionadas nos dados e nos valores dos parâmetros remanescentes, necessárias para a aplicação do método MCMC são obtidas da distribuição conjunta definida em (4.31). A seguir são detalhadas as distribuições a posteriori de cada parâmetro, usando a notação "Demais" para denotar os dados e os parâmetros remanescentes. A notação $p(\cdot)$ é usada para denotar as distribuições a priori.

$$f(\beta_{0i} \mid \text{Demais}) \propto \prod_{j=1}^{m_i} f_s(y_{ij} \mid \mathbf{x}_{ij}, \beta_{0i}, \boldsymbol{\beta}, \tau_\varepsilon) f_s(\mathsf{M}_i \mid \mathbf{z}_i, \beta_{0i}, \boldsymbol{\alpha}, \tau_M) f_s(\beta_{0i} \mid \mathbf{z}_i, \boldsymbol{\gamma}, \boldsymbol{\alpha}, \tau_u), \ i = 1...n$$

$$f(\boldsymbol{\beta} \mid \text{Demais}) \propto \prod_{i=1}^{n} \prod_{j=1}^{m_i} f_s(y_{ij} \mid \mathbf{x}_{ij}, \beta_{0i}, \boldsymbol{\beta}, \tau_\varepsilon) p(\boldsymbol{\beta})$$

$$f(\boldsymbol{\gamma} \mid \text{Demais}) \propto \prod_{i=1}^{n} f_s(\beta_{0i} \mid \mathbf{z}_i, \boldsymbol{\gamma}, \boldsymbol{\alpha}, \tau_u) p(\boldsymbol{\gamma})$$

$$f(\boldsymbol{\alpha} \mid \text{Demais}) \propto \prod_{i=1}^{n} f_s(\mathsf{M}_i \mid \mathbf{z}_i, \beta_{0i}, \boldsymbol{\alpha}, \tau_M) p(\boldsymbol{\alpha})$$

$$f(\tau_\varepsilon \mid \text{Demais}) \propto \prod_{i=1}^{n} \prod_{j=1}^{m_i} f_s(y_{ij} \mid \mathbf{x}_{ij}, \beta_{0i}, \boldsymbol{\beta}, \tau_\varepsilon) p(\tau_\varepsilon)$$

$$f(\tau_u \mid \text{Demais}) \propto \prod_{i=1}^{n} f_s(\beta_{0i} \mid \mathbf{z}_i, \boldsymbol{\gamma}, \boldsymbol{\alpha}, \tau_u) p(\tau_u)$$

Spiegelhalter (Eds.) Markov chain Monte Carlo in Practice. Chapman and Hall, London, UK. 1996.

$$f(\tau_{\mathbf{M}} \mid \text{Demais}) \propto \prod_{i=1}^{n} f_s(\mathsf{M}_i \mid \mathbf{z}_i, \beta_{0i}, \boldsymbol{\alpha}, \tau_M)\, p(\tau_{\mathbf{M}})$$

Foram utilizadas prioris pouco informativas para os parâmetros, como segue: $p(\boldsymbol{\beta}) = N(\mathbf{0};10^3\mathbf{I})$, $p(\boldsymbol{\gamma}) = N(\mathbf{0};10^3\mathbf{I})$, $p(\boldsymbol{\alpha}) = N(\mathbf{0};10^3\mathbf{I})$, $p(\tau_u) = Pa(10^{-2};1)$, $p(\tau_\varepsilon) = Pa(10^{-2};1))$ e $p(\tau_M) = Pa(10^{-2};1)$, onde $Pa(a,b)$ define a distribuição de *Pareto* com parâmetros *a* e *b*. Note que $\tau \sim Pa(10^{-2};1)$ é equivalente a $(1/\tau) = \sigma^2 \sim Uniforme(0,10^2)$.

5 – RESULTADOS

Neste capítulo são apresentados os resultados dos diversos ajustes, conforme as metodologias apresentadas nos capítulos 3 e 4. Como mencionado anteriormente, a aplicação consiste em modelar o valor do aluguel estimado em função das características/atributos do imóvel. Estas metodologias podem ser empregadas para imputar um valor de aluguel para imóveis que não são alugados, com base no valor que um inquilino pagaria por um imóvel com condições semelhantes de localização, vizinhança, dimensão e qualidade do imóvel.

Uma aplicação potencial para os métodos apresentados nesta dissertação é a imputação de aluguéis residenciais para os domicílios ocupados por seus proprietários. Esta informação é particularmente útil na composição do sistema de contas nacionais (SCN). Tal sistema consiste num conjunto integrado de contas macroeconômicas de forma a obter uma representação completa e clara do funcionamento da economia, ainda que simplificada. Um dos temas considerados no SCN é a atividade Aluguel de Imóveis, que abrange as famílias que alugam imóveis de sua propriedade e empresas especializadas na administração, locação e arrendamento de bens imóveis. A valoração do produto desta atividade também inclui os aluguéis imputados aos domicílios próprios.

5.1 – Ajuste do Modelo de Regressão Linear Múltipla

Com a finalidade de avaliar a relação entre o valor do aluguel estimado dos imóveis residenciais com seus atributos físicos e de localização, optou-se primeiramente por utilizar um modelo de regressão linear múltipla desconsiderando o fato dos dados provirem de uma amostra complexa. Este modelo servirá como base para a comparação dos resultados obtidos

por métodos de modelagem e estimação mais elaborados e adequados à estrutura e origem dos dados.

Para as variáveis explicativas observadas no nível de domicílio optou-se por trabalhar apenas com variáveis categóricas, desta forma foram categorizadas as variáveis apresentadas no Quadro 5.1 de acordo com sua freqüência.

Quadro 5.1 Variáveis que foram categorizadas

Variável original	Valores da variável categorizada
Número de quartos ou dormitórios.	1. Um ou nenhum 2. Dois 3. Três 4. Quatro ou mais
Número de cômodos menos quartos, banheiros e cozinha.	1. Nenhum 2. Um 3. Dois 4. Três ou mais
Número de cômodos usados permanentemente como dormitório.	1. Um 2. Dois 3. Três 4. Quatro ou mais
Número de cômodos usados exclusivamente para trabalho, estudo, etc.	1. Um ou mais 2. Nenhum

O ajuste do modelo de regressão linear múltipla foi feito através da PROC GENMOD do SAS, usando a função de ligação identidade e supondo a variável resposta (logaritmo do aluguel estimado) com distribuição normal. Como a quantidade de variáveis analisadas é relativamente alta, optou-se pelo critério de seleção *backward*, eliminando-se sucessivamente as variáveis que tivessem probabilidade de significância maior que 0,05, com base no teste de Wald. Desta forma foram removidas do modelo as variáveis listadas na Tabela 5.1.

Tabela 5.1 Variáveis não significativas ao nível de significância de 5%

Passo	Variável	Probabilidade de significância
1	Banheiro de uso exclusivo (Sim; Não).	0,8476
2	Número de cômodos usados permanentemente como dormitório.	0,6353
3	Localização do domicílio (Condomínio regularizado de casas e/ou apartamentos; Favelas ou conjuntos não regularizados; Casa-de-cômodos ou cortiço; Construção isolada).	0,3482
4	Outra forma de abastecimento de água (Rede geral; Poço na propriedade; Poço fora da propriedade; Bica pública; Carro pipa; Outra forma; Não utiliza).	0,1880
5	Outro tipo de iluminação (Elétrica; Gerador; Lampião; Vela; Não Utiliza).	0,1949
6	Número de cômodos usados exclusivamente para trabalho, estudo.	0,1557
7	Existência de cozinha (Sim; Não).	0,1245
8	Principal tipo de iluminação (Elétrica; Gerador; Lampião; Vela).	0,1051

Uma vez selecionadas as variáveis do modelo, foi feita uma análise de agrupamento dos níveis das variáveis, buscando agregar os níveis que não fossem significativamente diferentes entre si. Para maior facilidade na interpretação das estimativas dos parâmetros procurou-se hierarquizar de forma decrescente os níveis das variáveis categóricas de acordo com o valor médio do aluguel estimado, de modo que ao tomar a última linha como base espera-se que os sinais dos coeficientes estimados sejam positivos. Para avaliar a existência de diferença significativa entre os níveis, aplicou-se o teste de hipótese descrito a seguir.

H_0: $\beta_k - \beta_l = 0$, para $k \neq l$

H_1: $\beta_k - \beta_l \neq 0$, para $k \neq l$

A estatística de teste utilizada foi a Estatística de Wald apropriada, definida por:

$$W = \frac{(\hat{\beta}_k - \hat{\beta}_l)^2}{\hat{Var}(\hat{\beta}_k) + \hat{Var}(\hat{\beta}_l) - 2\hat{COV}(\hat{\beta}_k, \hat{\beta}_l)} \tag{5.1}$$

Sob a hipótese nula, a estatística W tem distribuição χ_1^2.

O procedimento utilizado consistiu em testar a igualdade dos parâmetros dos níveis das variáveis categóricas dois a dois e verificar se existia diferença entre eles ao nível de 5% de significância. Caso a hipótese nula fosse aceita então os dois níveis da variável seriam agrupados, repetindo-se este procedimento até que todos os níveis fossem significativamente diferentes entre si para todas as variáveis categóricas. Após esta etapa foi feita uma nova hierarquização decrescente das categorias das variáveis explicativas em função do valor médio do logaritmo do aluguel estimado.

O primeiro modelo proposto é explicitado na equação (5.2). A seguir, seus parâmetros são descritos no Quadro 5.2.

$$y_j = \sum_{p=0}^{P} \beta_p x_{pj} + \varepsilon_j \tag{5.2}$$

onde

y_j : logaritmo natural do aluguel estimado do j-ésimo domicílio da amostra;

x_{pj} : valor da p-ésima variável explicativa do j-ésimo domicílio da amostra;

β_p : parâmetro associado à p-ésima variável explicativa;

$\varepsilon_j \sim N\left(0, \sigma^2\right)$;

$Cov\left(\varepsilon_j, \varepsilon_l\right) = 0$, $\forall\ j \neq l$.

Quadro 5.2 Descrição dos Parâmetros e Variáveis Indicadoras Associadas às Variáveis Explicativas do Modelo (5.2)

Fonte/Fator	Parâmetro	Variável Indicadora
Intercepto	β_0	x_{0j}=1 em qualquer caso
Condição de ocupação do domicílio		
Próprio já pago, cedido por empregador ou invasão	β_1	x_{1j}=1 se o domicílio for próprio já pago, cedido por empregador ou invadido; x_{1j}=0 caso contrário.
Alugado ou cedido por outra forma	β_2	x_{2j}=1 se o domicílio for alugado ou cedido por outra forma; x_{2j}=0 caso contrário.
Próprio em aquisição	–	Categoria base; x_{1j}=x_{2j}=0
Tipo do domicílio		
Apartamento ou quarto/cômodo	β_3	x_{3j}=1 se o domicílio for apartamento ou quarto/cômodo; x_{3j}=0 caso contrário.
Casa	–	Categoria base; x_{3j}= 0
Material predominante nas paredes externas		
Alvenaria, madeira aparelhada, madeira aproveitada ou tijolo sem revestimento	β_4	x_{4j}=1 se as paredes externas forem de alvenaria, madeira aparelhada, madeira aproveitada ou tijolo sem revestimento; x_{4j}=0 caso contrário.
Taipa não-revestida ou outro material	–	Categoria base; x_{4j}= 0
Material predominante no piso		
Carpete	β_5	x_{5j}=1 se o piso for carpete; x_{5j}=0 caso contrário.
Madeira aparelhada, cerâmica, lajota, ardósia ou madeira aproveitada	β_6	x_{6j}=1 se o piso for de madeira aparelhada, cerâmica, lajota, ardósia ou madeira aproveitada; x_{6j}=0 caso contrário.
Cimento, terra ou outro material	–	Categoria base; x_{5j}=x_{6j}=0
Material predominante no teto		
Laje de concreto, madeira aparelhada ou madeira aproveitada	β_7	x_{7j}=1 se o teto for de laje de concreto, madeira aparelhada ou madeira aproveitada; x_{7j}=0 caso contrário.
Telha, zinco ou outro material	–	Categoria base; x_{7j}=0
Total de dormitórios		
Três ou mais	β_8	x_{8j}=1 se o domicílio possuir três ou mais dormitórios; x_{8j}=0 caso contrário.
Dois	β_9	x_{9j}=1 se o domicílio possuir dois dormitórios; x_{9j}=0 caso contrário.
Um ou nenhum	–	Categoria base; x_{8j}=x_{9j}=0
Total de cômodos menos quartos, banheiros e cozinha		
Três ou mais	β_{10}	x_{10j}=1 se o domicílio possuir três ou mais outros tipos de

Fonte/Fator	Parâmetro	Variável Indicadora
		cômodos; x_{10j}=0 caso contrário.
Um ou dois	β_{11}	x_{11j}=1 se o domicílio possuir um ou dois outros tipos de cômodos; x_{11j}=0 caso contrário.
Nenhum	–	Categoria base; x_{10j}=x_{11j}=0
Total de banheiros		
Três ou mais	β_{12}	x_{12j}=1 se o domicílio possuir três ou mais banheiros; x_{12j}=0 caso contrário.
Dois	β_{13}	x_{13j}=1 se o domicílio possuir dois banheiros; x_{13j}=0 caso contrário.
Um ou nenhum	–	Categoria base; x_{12j}=x_{13j}=0
Localização do banheiro		
Dentro do domicílio	β_{14}	x_{14j}=1 se o banheiro for dentro do domicílio; x_{14j}=0 caso contrário.
Fora do domicílio	–	Categoria base; x_{14j}= 0
Tipo de água usada para beber		
Mineral	β_{15}	x_{15j}=1 se a água utilizada para beber for mineral; x_{15j}=0 caso contrário.
Filtrada	β_{16}	x_{16j}=1 se a água utilizada para beber for filtrada; x_{16j}=0 caso contrário.
Fervida ou natural	–	Categoria base; x_{15j}=x_{16j}=0
Existência de calçada na frente do domicilio		
Sim	β_{17}	x_{17j}=1 se existir calçada na frente do domicílio; x_{17j}=0 caso contrário.
Não	–	Categoria base; x_{17j}= 0
Tipo de rua onde o domicílio se localiza		
Asfaltada	β_{18}	x_{18j}=1 se a rua for asfaltada; x_{18j}=0 caso contrário.
Paralelepípedo, terra/barro ou outro tipo	–	Categoria base; x_{18j}= 0
Estado de conservação do domicílio (entrevistador)		
Excelente	β_{19}	x_{19j}=1 se o domicílio estiver em excelente estado; x_{19j}=0 caso contrário.
Bom	β_{20}	x_{20j}=1 se o domicílio estiver em bom estado; x_{20j}=0 caso contrário.
Regular	β_{21}	x_{21j}=1 se o domicílio estiver em estado regular; x_{21j}=0 caso contrário.
Ruim	–	Categoria base; x_{19j}=x_{20j}= x_{21j}=0
Existência de água canalizada		
Sim	β_{22}	x_{22j}=1 se existir água canalizada no domicílio; x_{22j}=0 caso contrário.

Fonte/Fator	Parâmetro	Variável Indicadora
Não	–	Categoria base; $x_{22j}= 0$
Principal forma de abastecimento de água		
Rede geral ou poço na propriedade	β_{23}	$x_{23j}=1$ se o domicílio for abastecido por rede geral ou poço na propriedade; $x_{23j}=0$ caso contrário.
Carro pipa	β_{24}	$x_{24j}=1$ se o domicílio for abastecido por carro pipa; $x_{24j}=0$ caso contrário.
Poço fora da propriedade, bica pública ou outra forma	–	Categoria base; $x_{23j}=x_{24j}=0$
Tipo de escoadouro sanitário		
Fossa séptica	β_{25}	$x_{25j}=1$ se escoadouro sanitário for por fossa séptica; $x_{25j}=0$ caso contrário.
Rede coletora de esgoto, fossa rudimentar, vala ou outro tipo	β_{26}	$x_{26j}=1$ se o escoadouro sanitário for por rede coletora de esgoto, fossa rudimentar, vala ou outro tipo; $x_{26j}=0$ caso contrário.
Não tem	–	Categoria base; $x_{25j}=x_{26j}=0$
Principal combustível para cozinhar		
Eletricidade, gás botijão/canalizado, querosene ou outro	β_{27}	$x_{27j}=1$ se o principal combustível para cozinhar for eletricidade, gás botijão/canalizado, querosene ou outro; $x_{27j}=0$ caso contrário.
Carvão/lenha	–	Categoria base; $x_{27j}=0$
Outro tipo de combustível para cozinhar		
Eletricidade, querosene ou outro	β_{28}	$x_{28j}=1$ se o outro combustível para cozinhar for eletricidade, querosene ou outro; $x_{28j}=0$ caso contrário.
Gás botijão/canalizado ou não utiliza	β_{29}	$x_{29j}=1$ se o outro combustível para cozinhar for gás botijão/canalizado ou não utiliza; $x_{29j}=0$ caso contrário.
Carvão/lenha	–	Categoria base; $x_{28j}=x_{29j}=0$
Existência de telefone		
Sim	β_{30}	$x_{30j}=1$ se existir telefone no domicílio; $x_{30j}=0$ caso contrário.
Não	–	Categoria base; $x_{30j}=0$
Destino do lixo domiciliar		
Coletado	β_{31}	$x_{31j}=1$ se o lixo domiciliar for coletado; $x_{31j}=0$ caso contrário.
Queimado/enterrado, jogado em terreno baldio ou outro destino	β_{32}	$x_{32j}=1$ se o lixo domiciliar for queimado/enterrado, jogado em terreno baldio ou tiver outro destino; $x_{32j}=0$ caso contrário.
Jogado em rio, lagoa, etc.	–	Categoria base; $x_{31j}=x_{32j}=0$
Situação		
Urbana	β_{33}	$x_{33j}=1$ se o setor censitário estiver em área urbana; $x_{33j}=0$ caso contrário.
Rural	–	Categoria base; $x_{33j}= 0$

Fonte/Fator	Parâmetro	Variável Indicadora
Proporção de moradores brancos no setor censitário	β_{34}	–
Proporção de moradores com mais de 25 anos de idade e mais de 11 anos de estudo no setor censitário[3]	β_{35}	–
Idade mediana dos moradores do setor censitário	β_{36}	–
Densidade de moradores por dormitórios no setor censitário	β_{37}	–
Mediana do logaritmo da renda domiciliar no setor censitário	β_{38}	–

As estimativas dos parâmetros do modelo (5.2), seus respectivos erros padrões e os correspondentes efeitos do plano amostral estão apresentados na Tabela 5.2, que contém os resultados de dois ajustes: o primeiro considera que os dados foram obtidos a partir de uma amostra aleatória simples, enquanto que no segundo são utilizadas as informações do plano amostral de fato utilizado na PPV. Ao se comparar os resultados dos dois ajustes, percebe-se que todos os fatores são significativos ao nível de significância de 5% para o ajuste sem ponderação, enquanto que ao considerar as informações do plano amostral, alguns fatores deixam de ser significativos, conforme destacado na tabela.

Tabela 5.2 Estimativas dos parâmetros e de seus respectivos erros padrões para o modelo de regressão linear múltipla especificado na equação (5.2) e cujos parâmetros foram descritos no Quadro 5.2

Fonte/Fator	Parâmetro	Sem Ponderação		Com Ponderação		
		Estimativa	Erro Padrão	Estimativa	Erro Padrão	EPA
Intercepto	β_0	-3,528	0,252	-4,014	0,527	4,91
Condição de ocupação do domicílio						
Próprio já pago, cedido por empregador ou invasão	β_1	1,648	0,065	1,401	0,115	7,72

[3] Embora talvez fizesse mais sentido calcular esta proporção em relação aos moradores que tivessem mais de 25 anos, preferiu-se manter esta variável da mesma forma que foi indicada no artigo de Reis *et al*. (2001), ou seja, a proporção de moradores com mais de 25 anos de idade e com mais de 11 anos de estudo em relação ao total de moradores do setor censitário.

Fonte/Fator	Parâmetro	Sem Ponderação		Com Ponderação		
		Estimativa	Erro Padrão	Estimativa	Erro Padrão	EPA
Alugado ou cedido por outra forma	β_2	1,461	0,067	1,225	0,123	7,97
Próprio em aquisição	-	-	-	-	-	-
Tipo do domicílio						
Apartamento ou quarto/cômodo	β_3	-0,114	0,033	-0,182	0,051	2,63
Casa	-	-	-	-	-	-
Material predominante nas paredes externas						
Alvenaria, madeira aparelhada, madeira aproveitada ou tijolo sem revestimento	β_4	0,268	0,058	0,271	0,093	3,84
Taipa não-revestida ou outro material	-	-	-	-	-	-
Material predominante no piso						
Carpete	β_5	0,352	0,073	0,423	0,088	2,46
Madeira aparelhada, cerâmica, lajota, ardósia ou madeira aproveitada	β_6	0,228	0,023	0,276	0,036	2,36
Cimento, terra ou outro material	-	-	-	-	-	-
Material predominante no teto						
Laje de concreto, madeira aparelhada ou madeira aproveitada	β_7	0,136	0,021	0,141	0,034	2,43
Telha, zinco ou outro material	-	-	-	-	-	-
Total de dormitórios						
Três ou mais	β_8	0,224	0,028	0,215	0,038	1,96
Dois	β_9	0,128	0,023	0,119	0,029	1,58
Um ou nenhum	-	-	-	-	-	-
Total de cômodos menos quartos, banheiros e cozinha						
Três ou mais	β_{10}	0,161	0,046	**0,120**	**0,067**	2,39
Dois	β_{11}	0,084	0,034	**0,056**	**0,046**	2,12
Um ou nenhum	-	-	-	-	-	-
Total de banheiros						
Três ou mais	β_{12}	0,550	0,046	0,502	0,058	1,77
Dois	β_{13}	0,224	0,031	0,253	0,039	1,59
Um ou nenhum	-	-	-	-	-	-
Localização do banheiro						
Dentro do domicílio	β_{14}	0,209	0,029	0,209	0,046	2,50
Fora do domicílio	-	-	-	-	-	-
Tipo de água usada para beber						
Mineral	β_{15}	0,152	0,042	0,182	0,055	1,06
Filtrada	β_{16}	0,061	0,021	**0,050**	**0,032**	2,55
Fervida ou natural	-	-	-	-	-	-

Fonte/Fator	Parâmetro	Sem Ponderação		Com Ponderação		
		Estimativa	Erro Padrão	Estimativa	Erro Padrão	EPA
Existência de calçada na frente do domicilio						
Sim	β_{17}	0,073	0,023	0,075	0,038	2,80
Não	-	-	-	-	-	-
Tipo de rua onde o domicílio se localiza						
Asfaltada	β_{18}	0,167	0,022	0,184	0,040	2,91
Paralelepípedo, terra/barro ou outro tipo	-	-	-	-	-	-
Estado de conservação do domicílio (entrevistador)						
Excelente	β_{19}	0,379	0,048	0,445	0,069	2,44
Bom	β_{20}	0,185	0,033	0,205	0,058	3,36
Regular	β_{21}	0,104	0,029	0,115	0,052	3,81
Ruim	-	-	-	-	-	-
Existência de água canalizada						
Sim	β_{22}	0,073	0,036	**0,098**	**0,064**	3,52
Não	-	-	-	-	-	-
Principal forma de abastecimento de água						
Rede geral ou poço na propriedade	β_{23}	0,182	0,040	0,165	0,068	3,78
Carro pipa	β_{24}	0,523	0,162	0,533	0,263	5,62
Poço fora da propriedade, bica pública ou outra forma	-	-	-	-	-	-
Tipo de escoadouro sanitário						
Fossa séptica	β_{25}	0,303	0,050	0,256	0,083	3,08
Rede coletora de esgoto, fossa rudimentar, vala ou outro tipo	β_{26}	0,165	0,044	**0,134**	**0,074**	3,37
Não tem	-	-	-	-	-	-
Principal combustível para cozinhar						
Eletricidade, gás botijão/canalizado, querosene ou outro	β_{27}	0,307	0,048	0,287	0,079	3,79
Carvão/lenha	-	-	-	-	-	-
Outro tipo de combustível para cozinhar						
Eletricidade, querosene ou outro	β_{28}	0,572	0,117	0,510	0,141	2,69
Gás botijão/canalizado ou não utiliza	β_{29}	0,356	0,045	0,315	0,066	2,34
Carvão/lenha	-	-	-	-	-	-
Existência de telefone						
Sim	β_{30}	0,187	0,027	0,183	0,035	1,75
Não	-	-	-	-	-	-

Fonte/Fator	Parâmetro	Sem Ponderação		Com Ponderação		
		Estimativa	Erro Padrão	Estimativa	Erro Padrão	EPA
Destino do lixo domiciliar						
Coletado	β_{31}	0,371	0,061	0,338	0,143	4,98
Queimado/enterrado, jogado em terreno baldio ou outro destino	β_{32}	0,267	0,059	**0,215**	**0,142**	5,26
Jogado em rio, lagoa, etc.	-	-	-	-	-	-
Situação						
Urbana	β_{33}	0,183	0,033	**0,086**	**0,072**	5,01
Rural	-	-	-	-	-	-
Proporção de moradores brancos no setor censitário	β_{34}	0,317	0,039	0,212	0,088	4,66
Proporção de moradores com mais de 25 anos de idade e mais de 11 anos de estudo no setor censitário	β_{35}	0,797	0,188	0,801	0,384	4,01
Idade mediana dos moradores do setor censitário	β_{36}	0,011	0,003	**0,009**	**0,006**	4,86
Densidade de moradores por dormitórios no setor censitário	β_{37}	0,461	0,043	0,453	0,100	5,90
Mediana do logaritmo da renda domiciliar no setor censitário	β_{38}	0,295	0,020	0,379	0,046	5,20
Desvio-Padrão do erro do modelo	σ	0,599	-	0,577	-	-

Tomando-se como base o modelo ajustado sem considerar a ponderação e o plano amostral, e comparando-o com o modelo ajustado considerando as informações do plano amostral, percebe-se que não há mudança nos sinais dos coeficientes e que para a maioria deles não há uma variação de magnitude muito grande. As maiores mudanças ocorreram para os coeficientes estimados para a categoria de tipo de domicílio apartamento ou quarto/cômodo com uma diferença relativa de aproximadamente 60%, e para imóveis em situação urbana, com uma diferença relativa de 53%.

Ambos os ajustes foram realizados utilizando a PROC SURVEYREG do SAS. As estimativas dos erros-padrão são sempre maiores quando obtidas pelo ajuste que leva em conta o desenho amostral. Para as variáveis em nível de setor censitário (cujos parâmetros

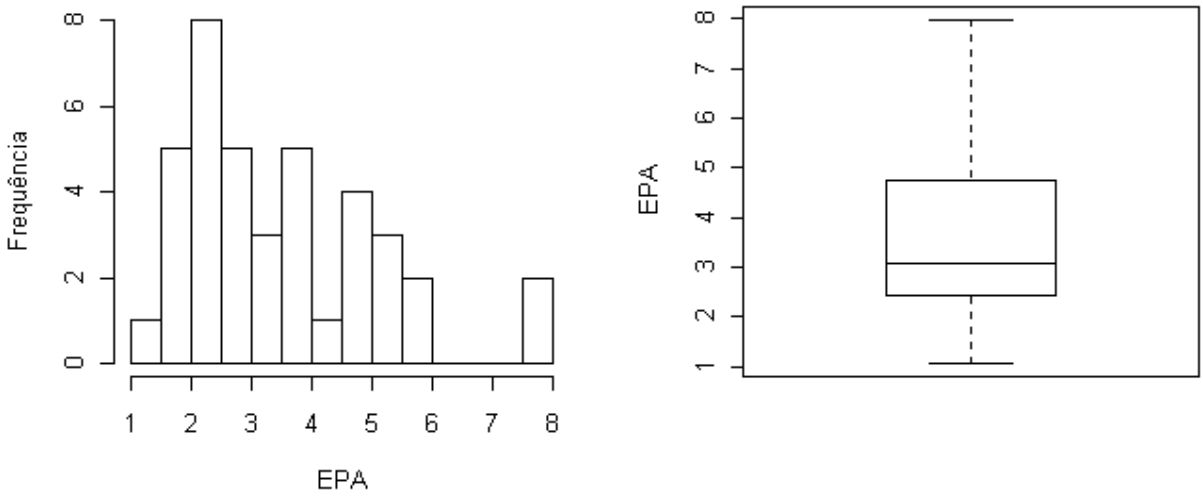
associados vão de β_{33} a β_{38}), além das de destino do lixo em nível de domicílio (parâmetros β_{31} e β_{32}), a diferença dos erros-padrão estimados chega a ser maior que 100%.

São também apresentados na Tabela 5.2 os efeitos do plano amostral – EPA. Estes valores são todos maiores que um, indicando que ao ignorar o plano amostral complexo as variâncias dos coeficientes estimados são subestimadas. Os casos mais agudos são observados para as categorias da condição de ocupação do domicílio. A Tabela 5.3 apresenta algumas medidas resumo dos EPAs da Tabela 5.2. Na Figura 5.1 são apresentados o histograma e o diagrama de caixas para tal medida.

Tabela 5.3 Medidas resumo dos efeitos do plano amostral ampliado – EPA

Mínimo	1º Quartil	Mediana	3º Quartil	Máximo	Média
1,06	2,41	3,08	4,76	7,97	3,51

Figura 5.1 Histograma e diagrama de caixas dos efeitos do plano amostral ampliado – EPA para os coeficientes do modelo 5.2



5.2 – Ajuste do Modelo Linear Hierárquico Usual

É possível que a relação entre o valor do aluguel e atributos físicos e de localização do imóvel possa ser melhor representada utilizando um modelo multinível ao invés de um modelo de regressão linear. A justificativa se baseia na hipótese de que o valor do aluguel do imóvel é diferenciado segundo sua localização geográfica, representada na pesquisa em menor nível de agregação pelo setor censitário. Neste sentido, é proposto um modelo linear hierárquico de dois níveis com interceptos aleatórios, conforme as equações (3.5) e (3.6) da seção 3.2 do capítulo 3. Neste modelo os domicílios são as unidades de nível 1 e os setores censitários as unidades de nível 2.

Um primeiro ajuste foi realizado utilizando um modelo multinível sem nenhuma variável explicativa, também conhecido como modelo nulo. Este ajuste permite estimar o coeficiente de correlação intraclasse conforme definido na equação (3.4), cujo objetivo é estimar a proporção da variância da variável resposta explicada pela variabilidade dos setores censitários. Na Tabela 5.4 são apresentadas as estimativas das variâncias dos erros aleatórios para este modelo.

Tabela 5.4 Estimativas das variâncias dos erros aleatórios e seus respectivos erros padrões para o modelo nulo

Fonte/Fator	Parâmetro	Estimativa	Erro Padrão
$Var(u_i)$	τ_0	0,864	0,055
$Var(\varepsilon_{ij})$	σ^2	0,487	0,010

A estimativa da variância dos erros aleatórios devida aos setores censitários $(\hat{\tau}_0)$, apresentada na Tabela 5.5, foi significativa a 5%, podendo-se concluir que o ajuste de um modelo hierárquico se faz necessário. A estimativa do coeficiente de correlação intraclasse é:

$$\hat{\rho}_I = \frac{\hat{\tau}_0}{\hat{\tau}_0 + \hat{\sigma}^2} = \frac{0{,}864}{0{,}864 + 0{,}487} = 0{,}6396\,. \qquad (5.3)$$

A equação (5.3) mostra que cerca de 64% da variabilidade total do logaritmo do aluguel estimado é devida à variabilidade dos setores censitários, indicando que é necessário utilizar a abordagem hierárquica na estimação dos parâmetros e que o uso de modelos de regressão linear não hierárquicos pode levar a conclusões equivocadas.

Uma vez confirmada a presença do efeito de grupo, foi ajustado um modelo hierárquico de dois níveis utilizando as mesmas variáveis explicativas do modelo de regressão linear múltipla sem ponderação, conforme apresentado na Tabela 5.2. Neste ajuste, as variáveis associadas aos parâmetros β_{33} à β_{38} foram consideradas como variáveis do nível 2 e seus parâmetros passaram a ser referidos como γ_{01} à γ_{06}. Esse modelo foi ajustado usando a PROC MIXED do SAS e o programa utilizado nesta aplicação encontra-se no Anexo 2. Exemplos de modelos multiníveis e da maneira de implementá-los na PROC MIXED podem ser encontrados de forma detalhada em Singer (1998).

Os resultados deste ajuste não são apresentados nesta seção, pois optou-se por apresentá-los na próxima seção comparando-os com os resultados obtidos pelo método de MQGIPP. Antes disto, verificou-se novamente a significância das variáveis com a finalidade de excluir aquelas não significativas. Desta forma, a proporção de moradores com mais de 25 anos de idade e mais de 11 anos de estudo no setor censitário foi excluída do ajuste por ter apresentando uma probabilidade de significância de 0,0856. Foi também realizado um teste de Wald para verificar a existência de diferenças não significativas entre as estimativas dos níveis das variáveis explicativas categóricas, conforme apresentado na Tabela 5.5.

Tabela 5.5 Comparação entre categorias das variáveis com alguma categoria não significativamente diferente da outra.

Variável	Categoria a		Categoria b	Probabilidade de significância
Material predominante no piso	Carpete	x	Madeira aparelhada, cerâmica, lajota, ardósia ou madeira aproveitada	0,9607
	Carpete	x	Cimento, terra ou outro material	0,0008
	Madeira aparelhada, cerâmica, lajota, ardósia ou madeira aproveitada	x	Cimento, terra ou outro material	0,0000
Principal forma de abastecimento de água	Rede geral ou poço na propriedade	x	Carro pipa	0,0751
	Rede geral ou poço na propriedade	x	Poço fora da propriedade, bica pública ou outra forma	0,0004
	Carro pipa	x	Poço fora da propriedade, bica pública ou outra forma	0,0068
Destino do lixo domiciliar	Coletado	x	Queimado/enterrado, jogado em terreno baldio ou outro destino	0,0009
	Coletado	x	Jogado em rio, lagoa, etc.	0,0028
	Queimado/enterrado, jogado em terreno baldio ou outro destino	x	Jogado em rio, lagoa, etc.	0,1448

As variáveis material predominante no piso, principal forma de abastecimento de água e destino do lixo domiciliar apresentaram níveis não significativamente diferentes entre suas categorias ao nível de significância de 5%. Agrupando-se as categorias em que os parâmetros não foram significativamente diferentes chega-se ao modelo explicitado nas equações (3.5) e (3.6) da seção 3.2 do capítulo 3, cujos resultados são apresentados na seção 5.3.

5.3 – Ajuste do Modelo Linear Hierárquico pelo Método MQGIPP

Para a operacionalização do ajuste de modelos lineares hierárquicos ponderado pelas probabilidades, Corrêa (2001) desenvolveu um programa em SAS utilizando a PROC IML (*Iterative Matrix Language*). Uma adaptação deste programa que foi usada nesta dissertação encontra-se no Apêndice 3. Neste programa são utilizados, além das variáveis da modelagem, os pesos dos setores censitários (w_i) e os pesos condicionais dos domicílios dentro do setor

censitário ($w_{j|i}$). Quando $w_i = w_{j|i} = 1$ em qualquer caso, os resultados da execução deste programa nada mais são que estimativas obtidas por Mínimos Quadrados Generalizados Iterativo (MQGI).

Para o ajuste por MQGIPP foram utilizadas como variáveis explicativas somente aquelas que foram significativas no ajuste da seção 5.2. Os resultados do primeiro ajuste, já ponderando pelas probabilidades de inclusão, indicaram que dez variáveis não eram significativas a 5%. Resolveu-se então utilizar o critério de seleção *backward* já utilizado na seção 5.1. A Tabela 5.6 apresenta a relação de variáveis que não foram significativas a 5% e que, portanto foram excluídas do modelo final.

Tabela 5.6 Variáveis não significativas ao nível de significância de 5%

Passo	Variável	Probabilidade de significância
1	Utilização de água filtrada para beber.	0,8113
2	Lixo domiciliar coletado.	0,3862
3	Existência de água canalizada.	0,2495
4	Escoadouro sanitário do tipo rede coletora de esgoto, fossa rudimentar, vala ou outro.	0,1997
5	Idade mediana dos moradores do setor censitário.	0,1090
6	Outro combustível para cozinhar do tipo gás botijão/canalizado ou não utiliza.	0,1011
7	Outro combustível para cozinhar do tipo eletricidade, querosene ou outro.	0,1723
8	Domicílio do tipo apartamento ou quarto/cômodo.	0,1032
9	Rua onde o domicílio se localiza ser do tipo asfaltada.	0,0732
10	Escoadouro sanitário do tipo fossa séptica.	0,0578

As demais variáveis foram significativas a 5% e os resultados das estimativas estão apresentados na Tabela 5.7, juntamente com as estimativas pelo método de MQGI, onde pode-se perceber que o modelo ajustado por MQGI possui um número bem maior de parâmetros estimados que o ajustado por MQGIPP. Os resultados indicam que, embora não

existam diferenças nos sinais, há discrepâncias nas estimativas dos parâmetros quando se compara o método de estimação de mínimos quadrados generalizados iterativo ponderado pelas probabilidades (MQGIPP) com o método de mínimos quadrados generalizados iterativo (MQGI). Adotando-se como base o método MQGI, pode-se verificar que a maior diferença para os coeficientes estimados ocorre existência de calçada em frente ao domicílio, com um aumento de cerca de 105% do coeficiente estimado considerando o método MQGIPP em comparação com o MQGI.

Quando são comparadas as estimativas dos erros padrões dos coeficientes estimados para os dois métodos, as diferenças são ainda maiores. Com exceção da variável água mineral utilizada para beber, existe um viés no método MQGI, no sentido de que os valores dos erros padrões dos coeficientes estão subestimados em relação ao método de ajuste MQGIPP. As variações são particularmente grandes para as condições de ocupação do domicílio, que têm diferenças relativas superiores a 200%.

Outro resultado importante é que o efeito de grupo se torna ainda mais forte para o ajuste realizado pelo método MQGIPP do que no ajuste por MQGI. O coeficiente de correlação intraclasse residual estimado fica em torno de 40%, enquanto que para o método MQGI este valor era de aproximadamente 30%.

Tabela 5.7 Estimativas dos parâmetros e de seus respectivos erros padrões para o modelo hierárquico obtidas pelos métodos MQGI e MQGIPP

Fonte/Fator	Parâmetro	Ajuste por MQGI		Ajuste por MQGIPP	
		Estimativa	Erro Padrão	Estimativa	Erro Padrão
Nível 1 – Domicílio					
Condição de ocupação do domicílio					
Próprio já pago, cedido por empregador ou invasão	β_1	1,497	0,038	1,190	0,121
Alugado ou cedido por outra forma	β_2	1,305	0,040	1,057	0,123
Próprio em aquisição	-	-	-	-	-
Tipo do domicílio					

Fonte/Fator	Parâmetro	Ajuste por MQGI		Ajuste por MQGIPP	
		Estimativa	Erro Padrão	Estimativa	Erro Padrão
Apartamento ou quarto/cômodo	β_3	-0,078	0,031	-	-
Casa	-	-	-	-	-
Material predominante nas paredes externas					
Alvenaria, madeira aparelhada, madeira aproveitada ou tijolo sem revestimento	β_4	0,258	0,050	0,229	0,057
Taipa não-revestida ou outro material	-	-	-	-	-
Material predominante no piso					
Carpete, madeira aparelhada, cerâmica, lajota, ardósia ou madeira aproveitada	β_5	0,199	0,022	0,227	0,044
Cimento, terra ou outro material	-	-	-	-	-
Material predominante no teto					
Laje de concreto, madeira aparelhada ou madeira aproveitada	β_6	0,108	0,023	0,128	0,057
Telha, zinco ou outro material	-	-	-	-	-
Total de dormitórios					
Três ou mais	β_7	0,274	0,026	0,275	0,043
Dois	β_8	0,154	0,022	0,156	0,031
Um ou nenhum	-	-	-	-	-
Total de cômodos menos quartos, banheiros e cozinha					
Três ou mais	β_9	0,274	0,043	0,304	0,074
Um ou dois	β_{10}	0,158	0,033	0,187	0,042
Nenhum	-	-	-	-	-
Total de banheiros					
Três ou mais	β_{11}	0,499	0,042	0,467	0,058
Dois	β_{12}	0,203	0,029	0,226	0,033
Um ou nenhum	-	-	-	-	-
Localização do banheiro					
Dentro do domicílio	β_{13}	0,183	0,028	0,224	0,041
Fora do domicílio	-	-	-	-	-
Tipo de água usada para beber					
Mineral	β_{14}	0,164	0,046	0,194	0,039
Filtrada	β_{15}	0,074	0,020	-	-
Fervida ou natural	-	-	-	-	-
Existência de calçada na frente do domicilio					
Sim	β_{16}	0,079	0,024	0,162	0,045
Não	-	-	-	-	-
Tipo de rua onde o domicílio se localiza					
Asfaltada	β_{17}	0,117	0,026	-	-

Fonte/Fator	Parâmetro	Ajuste por MQGI		Ajuste por MQGIPP	
		Estimativa	Erro Padrão	Estimativa	Erro Padrão
Paralelepípedo, terra/barro ou outro tipo	-	-	-	-	-
Estado de conservação do domicílio (entrevistador)					
Excelente	β_{18}	0,462	0,043	0,505	0,074
Bom	β_{19}	0,270	0,031	0,330	0,056
Regular	β_{20}	0,175	0,025	0,227	0,051
Ruim	-	-	-	-	-
Existência de água canalizada					
Sim	β_{21}	0,070	0,033	-	-
Não	-	-	-	-	-
Principal forma de abastecimento de água					
Rede geral, poço na propriedade ou carro pipa	β_{22}	0,129	0,033	0,181	0,060
Poço fora da propriedade, bica pública ou outra forma	-	-	-	-	-
Tipo de escoadouro sanitário					
Fossa séptica	β_{23}	0,188	0,048	-	-
Rede coletora de esgoto, fossa rudimentar, vala ou outro	β_{24}	0,085	0,039	-	-
Não tem	-	-	-	-	-
Principal combustível para cozinhar					
Eletricidade, gás botijão/canalizado, querosene ou outro	β_{25}	0,131	0,042	0,122	0,054
Carvão/lenha	-	-	-	-	-
Outro tipo de combustível para cozinhar					
Eletricidade, querosene ou outro	β_{26}	0,331	0,091	-	-
Gás botijão/canalizado ou não utiliza	β_{27}	0,128	0,042	-	-
Carvão/lenha	-	-	-	-	-
Existência de telefone					
Sim	β_{28}	0,171	0,025	0,200	0,036
Não	-	-	-	-	-
Destino do lixo domiciliar					
Coletado	β_{29}	0,112	0,031	-	-
Queimado/enterrado, jogado em terreno baldio, rio, lagoa ou outro destino	-	-	-	-	-
Nível 2 – Setor Censitário					
Intercepto	γ_{00}	-3,922	0,345	-4,603	0,477
Situação					
Urbana	γ_{01}	0,294	0,055	0,391	0,099
Rural	-	-	-	-	-

Fonte/Fator	Parâmetro	Ajuste por MQGI		Ajuste por MQGIPP	
		Estimativa	Erro Padrão	Estimativa	Erro Padrão
Proporção de moradores brancos no setor censitário	γ_{02}	0,339	0,075	0,425	0,105
Idade mediana dos moradores do setor censitário	γ_{03}	0,018	0,005	-	-
Densidade de moradores por dormitórios no setor censitário	γ_{04}	0,514	0,077	0,479	0,103
Mediana do logaritmo da renda domiciliar no setor censitário	γ_{05}	0,374	0,030	0,504	0,040
Variâncias dos erros aleatórios					
$Var(u_i)$	τ_0	0,108	0,009	0,135	0,015
$Var(\varepsilon_{ij})$	σ^2	0,256	0,006	0,198	0,011

5.4 – Ajuste do Modelo Linear Hierárquico pelo Método da Distribuição Amostral via MCMC

As estimativas para esta abordagem foram obtidas por simulação via MCMC utilizando a versão 1.4 do *software* WinBUGS. O algoritmo utilizado na simulação foi o amostrador de Gibbs. Como discutido na seção 4.7 este algoritmo consiste em amostrar alternadamente da distribuição a posteriori de cada parâmetro desconhecido, dados os valores atuais dos parâmetros remanescentes e os dados. Para este ajuste foram consideradas apenas as variáveis que foram significativas no ajuste por MQGIPP.

Foram geradas simultaneamente duas cadeias, cada uma com 20.000 amostras da distribuição a posteriori dos parâmetros. Para verificar se as simulações convergiram foi calculada a estatística $\sqrt{\hat{R}}$ definida em (4.30) para cada parâmetro de interesse. Os resultados apresentados na Tabela 5.8 mostram que a convergência pode ser considerada satisfatória, uma vez que os valores da estatística estão muito próximos de um para todos os parâmetros.

As estimativas dos parâmetros foram calculadas como médias empíricas das 10.000 amostras resultantes após o descarte das 10.000 primeiras amostras, que foram consideradas

como amostra de aquecimento. Como a convergência foi verificada, optou-se por estimar as quantidades apenas com base nos valores gerados da primeira cadeia. Os resultados são apresentados na Tabela 5.8.

Tabela 5.8 Estimativas dos parâmetros, dos seus respectivos erros padrões e da estatística $\sqrt{\hat{R}}$ para o modelo hierárquico obtidas pelo método do modelo de distribuição amostral por MCMC

Fonte/Fator	Parâmetro	Estimativa	Erro Padrão	$\sqrt{\hat{R}}$
Nível 1 – Domicílio				
Condição de ocupação do domicílio				
Próprio já pago, cedido por empregador ou invasão	β_1	1,508	0,037	1,004
Alugado ou cedido por outra forma	β_2	1,323	0,039	1,003
Próprio em aquisição	-	-	-	-
Material predominante nas paredes externas				
Alvenaria, madeira aparelhada, madeira aproveitada ou tijolo sem revestimento	β_4	0,278	0,049	1,005
Taipa não-revestida ou outro material	-	-	-	-
Material predominante no piso				
Carpete, madeira aparelhada, cerâmica, lajota, ardósia ou madeira aproveitada	β_5	0,211	0,022	1,000
Cimento, terra ou outro material	-	-	-	-
Material predominante no teto				
Laje de concreto, madeira aparelhada ou madeira aproveitada	β_6	0,110	0,022	1,000
Telha, zinco ou outro material	-	-	-	-
Total de dormitórios				
Três ou mais	β_7	0,277	0,026	1,000
Dois	β_8	0,156	0,022	1,000
Um ou nenhum	-	-	-	-
Total de cômodos menos quartos, banheiros e cozinha				
Três ou mais	β_9	0,275	0,043	1,001
Um ou dois	β_{10}	0,162	0,032	1,003
Nenhum	-	-	-	-
Total de banheiros				
Três ou mais	β_{11}	0,499	0,043	1,000

Fonte/Fator	Parâmetro	Estimativa	Erro Padrão	$\sqrt{\hat{R}}$
Dois	β_{12}	0,200	0,029	1,000
Um ou nenhum	-	-	-	-
Localização do banheiro				
Dentro do domicílio	β_{13}	0,255	0,025	1,000
Fora do domicílio	-	-	-	-
Tipo de água usada para beber				
Mineral	β_{14}	0,104	0,043	1,000
Filtrada, fervida ou natural	-	-	-	-
Existência de calçada na frente do domicilio				
Sim	β_{16}	0,136	0,023	1,000
Não	-	-	-	-
Estado de conservação do domicílio (entrevistador)				
Excelente	β_{18}	0,503	0,043	1,000
Bom	β_{19}	0,307	0,030	1,000
Regular	β_{20}	0,200	0,025	1,000
Ruim	-	-	-	-
Principal forma de abastecimento de água				
Rede geral, poço na propriedade ou carro pipa	β_{22}	0,159	0,032	1,000
Poço fora da propriedade, bica pública ou outra forma	-	-	-	-
Principal combustível para cozinhar				
Eletricidade, gás botijão/canalizado, querosene ou outro	β_{25}	0,073	0,036	1,001
Carvão/lenha	-	-	-	-
Existência de telefone				
Sim	β_{28}	0,186	0,025	1,000
Não	-	-	-	-
Nível 2 – Setor Censitário				
Intercepto	γ_{00}	-4,019	0,352	1,003
Situação				
Urbana	γ_{01}	0,468	0,055	1,000
Rural	-	-	-	-
Proporção de moradores brancos no setor censitário	γ_{02}	0,448	0,075	1,000
Densidade de moradores por dormitórios no setor censitário	γ_{04}	0,421	0,072	1,000
Mediana do logaritmo da renda domiciliar no setor censitário	γ_{05}	0,434	0,030	1,000
Variâncias dos erros aleatórios				

Fonte/Fator	Parâmetro	Estimativa	Erro Padrão	$\sqrt{\hat{R}}$
$Var(u_i)$	τ_0	0,124	0,010	1,000
$Var(\varepsilon_{ij})$	σ^2	0,259	0,006	1,001

Pode-se perceber na Tabela 5.8 que todos os parâmetros são significativos a 5%. Entretanto, o fato que mais chama atenção é que as estimativas são bem próximas àquelas encontradas pelo método MQGI, ou seja, o ajuste efetuado ignorando o plano amostral. Na Tabela 5.9 são apresentadas as diferenças absolutas e relativas das estimativas dos parâmetros dos ajustes pelos métodos de MQGIPP e MCMC em relação ao método de MQGI. Enquanto que os resultados para o método MQGIPP diferem consideravelmente do método MQGI, a única diferença maior que 10% entre os resultados dos métodos MCMC e MQGI é para o parâmetro β_{25}, entretanto a magnitude deste parâmetro é relativamente pequena.

Tabela 5.9 Diferenças absolutas e relativas das estimativas dos parâmetros obtidas com os métodos de estimação MQGIPP e MCMC em relação às obtidas com o método de MQGI

Fonte/Fator	Parâmetro	Diferença Absoluta		Diferença Relativa	
		MQGIPP	MCMC	MQGIPP	MCMC
Nível 1 – Domicílio					
Condição de ocupação do domicílio					
Próprio já pago, cedido por empregador ou invasão	β_1	-0,317	0,001	-21,04%	0,07%
Alugado ou cedido por outra forma	β_2	-0,264	0,002	-19,98%	0,15%
Próprio em aquisição	-	-	-	-	-
Material predominante nas paredes externas					
Alvenaria, madeira aparelhada, madeira aproveitada ou tijolo sem revestimento	β_4	-0,055	-0,006	-19,37%	-2,11%
Taipa não-revestida ou outro material	-	-	-	-	-
Material predominante no piso					
Carpete, madeira aparelhada, cerâmica, lajota, ardósia ou madeira aproveitada	β_5	0,017	0,001	8,10%	0,43%
Cimento, terra ou outro material	-	-	-	-	-
Material predominante no teto					

Fonte/Fator	Parâmetro	Diferença Absoluta		Diferença Relativa	
		MQGIPP	MCMC	MQGIPP	MCMC
Laje de concreto, madeira aparelhada ou madeira aproveitada	β_6	0,018	0,000	16,36%	0,27%
Telha, zinco ou outro material	-	-	-	-	-
Total de dormitórios					
Três ou mais	β_7	0,002	0,004	0,73%	1,61%
Dois	β_8	0,002	0,002	1,30%	1,10%
Um ou nenhum	-	-	-	-	-
Total de cômodos menos quartos, banheiros e cozinha					
Três ou mais	β_9	0,029	0,000	10,55%	0,00%
Um ou dois	β_{10}	0,023	-0,002	14,02%	-0,98%
Nenhum	-	-	-	-	-
Total de banheiros					
Três ou mais	β_{11}	-0,026	0,006	-5,27%	1,16%
Dois	β_{12}	0,027	0,001	13,57%	0,60%
Um ou nenhum	-	-	-	-	-
Localização do banheiro					
Dentro do domicílio	β_{13}	-0,032	-0,001	-12,50%	-0,51%
Fora do domicílio	-	-	-	-	-
Tipo de água usada para beber					
Mineral	β_{14}	0,095	0,005	95,96%	5,25%
Filtrada, fervida ou natural	-	-	-	-	-
Existência de calçada na frente do domicilio					
Sim	β_{16}	0,022	-0,004	15,71%	-2,71%
Não	-	-	-	-	-
Estado de conservação do domicílio (entrevistador)					
Excelente	β_{18}	0,006	0,004	1,20%	0,72%
Bom	β_{19}	0,024	0,001	7,84%	0,33%
Regular	β_{20}	0,028	0,001	14,07%	0,40%
Ruim	-	-	-	-	-
Principal forma de abastecimento de água					
Rede geral, poço na propriedade ou carro pipa	β_{22}	0,01	-0,012	5,85%	-7,02%
Poço fora da propriedade, bica pública ou outra forma	-	-	-	-	-
Principal combustível para cozinhar					
Eletricidade, gás botijão/canalizado, querosene ou outro	β_{25}	0,028	-0,021	29,79%	-22,00%

Fonte/Fator	Parâmetro	Diferença Absoluta		Diferença Relativa	
		MQGIPP	MCMC	MQGIPP	MCMC
Carvão/lenha	-	-	-	-	-
Existência de telefone					
Sim	β_{28}	0,017	0,003	9,29%	1,75%
Não	-	-	-	-	-
Nível 2 – Setor Censitário					
Intercepto	γ_{00}	-0,578	0,006	14,36%	-0,15%
Situação					
Urbana	γ_{01}	-0,047	0,030	-10,73%	6,76%
Rural	-	-	-	-	-
Proporção de moradores brancos no setor censitário	γ_{02}	-0,019	0,004	-4,28%	0,86%
Densidade de moradores por dormitórios no setor censitário	γ_{04}	0,073	0,015	17,98%	3,74%
Mediana do logaritmo da renda domiciliar no setor censitário	γ_{05}	0,067	-0,003	15,33%	-0,76%
Variâncias dos erros aleatórios					
$Var(u_i)$	τ_0	0,013	0,002	10,66%	1,48%
$Var(\varepsilon_{ij})$	σ^2	-0,060	0,001	-23,26%	0,47%

Estes resultados não estão tão distantes dos encontrados por Pfeffermann, Moura e Silva (2001) em seu estudo. Os autores simularam esquemas de seleção em dois estágios com quatro possibilidades, quais sejam: i) seleção não informativa nos dois estágios; ii) seleção não informativa no estágio 1 e informativa no estágio 2; iii) seleção informativa no estágio 1 e não informativa no estágio 2; e iv) seleção informativa nos dois estágios. Os maiores vieses em relação às estimativas ignorando o plano amostral foram obtidos quando o processo de seleção do estágio 2 era informativo. No caso (iii) que é análogo ao esquema de seleção da PPV, os vieses não foram significativos para a maioria dos parâmetros.

A Tabela 5.10 apresenta os resultados da comparação dos erros padrões obtidos pelos métodos de MQGIPP e MCMC em relação ao método de MQGI. Observa-se que, em média, as estimativas dos erros padrões obtidas pelo método MCMC são mais próximas às obtidas pelo método MQGI do que às obtidas por MQGIPP.

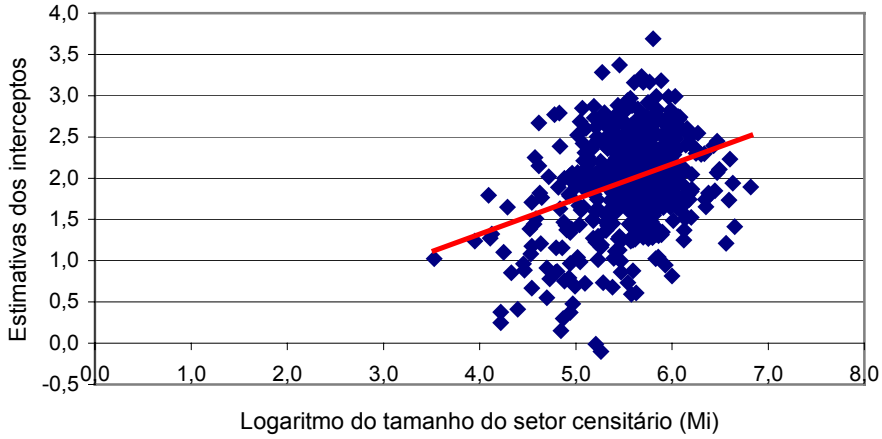
Tabela 5.10 Diferenças absolutas e relativas das estimativas dos erros padrões das estimativas dos parâmetros obtidas com os métodos de estimação MQGIPP e MCMC em relação às obtidas com o método de MQGI

Fonte/Fator	Parâmetro	Diferença Absoluta		Diferença Relativa	
		MQGIPP	MCMC	MQGIPP	MCMC
Nível 1 – Domicílio					
Condição de ocupação do domicílio					
Próprio já pago, cedido por empregador ou invasão	β_1	0,024	-0,060	24,74%	-61,86%
Alugado ou cedido por outra forma	β_2	0,024	-0,060	24,24%	-60,61%
Próprio em aquisição	-	-	-	-	-
Material predominante nas paredes externas					
Alvenaria, madeira aparelhada, madeira aproveitada ou tijolo sem revestimento	β_4	0,004	-0,004	7,55%	-7,55%
Taipa não-revestida ou outro material	-	-	-	-	-
Material predominante no piso					
Carpete, madeira aparelhada, cerâmica, lajota, ardósia ou madeira aproveitada	β_5	0,021	-0,001	91,30%	-4,35%
Cimento, terra ou outro material	-	-	-	-	-
Material predominante no teto					
Laje de concreto, madeira aparelhada ou madeira aproveitada	β_6	0,035	0,000	159,09%	0,00%
Telha, zinco ou outro material	-	-	-	-	-
Total de dormitórios					
Três ou mais	β_7	0,016	-0,001	59,26%	-3,70%
Dois	β_8	0,009	0,000	40,91%	0,00%
Um ou nenhum	-	-	-	-	-
Total de cômodos menos quartos, banheiros e cozinha					
Três ou mais	β_9	0,029	-0,002	64,44%	-4,44%
Um ou dois	β_{10}	0,010	0,000	31,25%	0,00%
Nenhum	-	-	-	-	-
Total de banheiros					
Três ou mais	β_{11}	0,004	-0,011	7,41%	-20,37%
Dois	β_{12}	0,003	-0,001	10,00%	-3,33%
Um ou nenhum	-	-	-	-	-
Localização do banheiro					
Dentro do domicílio	β_{13}	0,016	0,000	64,00%	0,00%
Fora do domicílio	-	-	-	-	-

Fonte/Fator	Parâmetro	Diferença Absoluta		Diferença Relativa	
		MQGIPP	MCMC	MQGIPP	MCMC
Tipo de água usada para beber					
Mineral	β_{14}	-0,003	0,001	-7,14%	2,38%
Filtrada, fervida ou natural	-	-	-	-	-
Existência de calçada na frente do domicilio					
Sim	β_{16}	0,022	0,000	95,65%	0,00%
Não	-	-	-	-	-
Estado de conservação do domicílio (entrevistador)					
Excelente	β_{18}	0,026	-0,005	54,17%	-10,42%
Bom	β_{19}	0,022	-0,004	64,71%	-11,76%
Regular	β_{20}	0,023	-0,003	82,14%	-10,71%
Ruim	-	-	-	-	-
Principal forma de abastecimento de água					
Rede geral, poço na propriedade ou carro pipa	β_{22}	0,023	-0,005	62,16%	-13,51%
Poço fora da propriedade, bica pública ou outra forma	-	-	-	-	-
Principal combustível para cozinhar					
Eletricidade, gás botijão/canalizado, querosene ou outro	β_{25}	0,011	-0,007	25,58%	-16,28%
Carvão/lenha	-	-	-	-	-
Existência de telefone					
Sim	β_{28}	0,009	-0,002	33,33%	-7,41%
Não	-	-	-	-	-
Nível 2 – Setor Censitário					
Intercepto	γ_{00}	0,068	-0,057	16,63%	-13,94%
Situação					
Urbana	γ_{01}	0,029	-0,015	41,43%	-21,43%
Rural	-	-	-	-	-
Proporção de moradores brancos no setor censitário	γ_{02}	0,035	0,005	50,00%	7,14%
Densidade de moradores por dormitórios no setor censitário	γ_{04}	0,027	-0,004	35,53%	-5,26%
Mediana do logaritmo da renda domiciliar no setor censitário	γ_{05}	0,007	-0,003	21,21%	-9,09%
Variâncias dos erros aleatórios					
$Var(u_i)$	τ_0	0,004	-0,001	36,36%	-9,09%
$Var(\varepsilon_{ij})$	σ^2	-0,003	-0,008	-21,43%	-57,14%

Outro resultado que pode ser observado é que as estimativas dos erros padrões obtidas pelo método MQGIPP tendem a ser maiores que as obtidas por MQGI, o que implica em modelos com menos parâmetros a serem estimados. Para as estimativas do método MCMC ocorre o inverso, com as estimativas dos erros padrões tendendo a ser menores que as obtidas por MQGI. Uma explicação possível para este comportamento é que ao considerar as probabilidades de inclusão na amostra como dados adicionais no ajuste, o método de MCMC pode levar à estimativas mais precisas. Entretanto, é possível que a aproximação da distribuição amostral no método MCMC não tenha sido a mais adequada. A Figura 5.2 ilustra a relação entre o logaritmo do tamanho do setor e as estimativas dos interceptos.

Figura 5.2 Logaritmo do tamanho do setor censitário (Mi) versus as estimativas dos interceptos



A equação da reta estimada por Mínimos Quadrados Ordinários na Figura 5.2 é a seguinte: $\text{Intercepto} = -0{,}37 + 0{,}42 Log(\text{M}_\text{i})$, com um R^2 de 11% e o coeficiente para $Log(\text{M}_\text{i})$ significativo. Embora os interceptos não sejam quantidades observáveis, esta figura fornece indícios de que a relação entre as duas quantidades não tem uma tendência linear muito forte. Este resultado implica que outras formas de relação devem ser investigadas para tentar descrever melhor o modelo de distribuição amostral.

6 – CONCLUSÕES E TRABALHOS FUTUROS

Nesta dissertação procurou-se discutir um conjunto de métodos utilizados para o ajuste de modelos lineares hierárquicos com dados de pesquisas amostrais complexas. Foram estudados apenas modelos hierárquicos lineares de dois níveis com interceptos aleatórios para evitar complicações computacionais. Entretanto, a teoria apresentada pode ser estendida para casos mais gerais.

O ajuste "ingênuo" do modelo, ou seja, ignorando o plano amostral, já é amplamente disseminado e utilizado por analistas. Contudo, a maioria das pesquisas amostrais é realizada com algum procedimento amostral complexo, que pode ser informativo para a análise de interesse. Desta forma, faz-se necessário o uso de procedimentos que considerem as informações do plano amostral na modelagem dos dados.

O método de Mínimos Quadrados Generalizados Iterativo Ponderado pelas Probabilidades proposto por Pfeffermann *et al.* (1998a), embora corrija as estimativas utilizando as probabilidades de seleção, não é muito acessível para os analistas já que não está disponível em pacotes estatísticos usuais. Sua implementação depende de programação complexa, o que pode ser uma tarefa bastante custosa. Além disto, para este método de ajuste do modelo não existem métodos simples de diagnóstico do ajuste efetuado.

A abordagem do modelo de distribuição amostral tem como principal vantagem o uso de procedimentos padrões de inferência. Entretanto, existe a dificuldade de identificar a forma correta das esperanças condicionais necessárias para aproximar a distribuição amostral. Na estimação dos parâmetros do modelo da distribuição amostral através de simulações estocásticas por MCMC o principal limitador é a capacidade computacional. À medida que os modelos vão se tornando mais complexos, o tempo de processamento cresce muito, além de ser necessária uma grande capacidade de armazenamento. Na aplicação, foram

experimentados problemas de convergência quando se tentou melhorar/modificar o modelo para a relação dos tamanhos dos setores com as variáveis de interesse do modelo.

O modelo ajustado para a relação entre o logaritmo do valor estimado do aluguel residencial com as características físicas e de localização do imóvel mostrou que existe um forte efeito de grupo, com domicílios localizados num mesmo setor tendendo a ter valores de aluguel mais próximos. Isto confirmou a necessidade de considerar a estrutura hierárquica no processo de modelagem. No entanto, há uma dificuldade considerável para obter medidas representativas do aluguel no nível de setor censitário, uma vez que as pesquisas não são desenhadas para ter estimativas precisas neste nível de agregação. Adicionalmente, evidenciou-se a existência de uma tendência geral de subestimação de variâncias para modelos ajustados de forma ingênua, isto é, ignorando o efeito do plano amostral. Como conseqüência, os modelos eleitos nestes tipos de ajustes contêm mais parâmetros que o necessário.

Existem diversas aplicações para a teoria apresentada nesta dissertação além da que foi aqui discutida. Um estudo importante para o qual a estrutura de hierarquia é claramente presente é o realizado pelo Programa Internacional de Avaliação de Estudantes – PISA. Esta avaliação é desenvolvida conjuntamente pelos países-membros da Organização para Cooperação e Desenvolvimento Econômico – OCDE – e tem como objetivo principal aferir os conhecimentos essenciais de alunos com idade entre 15 anos e 3 meses e 16 anos e 3 meses. No Brasil, o estudo é realizado através de um levantamento amostral probabilístico complexo envolvendo estratificação e conglomeração, com alunos sorteados dentro de escolas que foram sorteadas num primeiro estágio. Neste caso, caracteriza-se a possível necessidade do ajuste de modelos hierárquicos e existe a hipótese do plano ser informativo, uma vez que os conhecimentos dos alunos podem ser relacionados aos mecanismos de estratificação da amostra, tais como estrutura física e tipo de rede (pública ou privada) das escolas.

Como continuação desta dissertação, seria interessante estudar outras funções possíveis para a forma das esperanças condicionais no modelo de distribuição amostral e fazer comparações com os resultados aqui obtidos. Outra contribuição importante seria trabalhar com casos mais gerais de modelos hierárquicos, como modelos com mais níveis de variação ou com coeficientes aleatórios além de interceptos aleatórios. Outro aspecto importante é a necessidade de desenvolvimento e implementação dos procedimentos apresentados nesta dissertação em pacotes que facilitem sua utilização para os analistas em geral.

BIBLIOGRAFIA

ALBIERI, Sonia; BIANCHINI, Zélia M. **Aspectos de amostragem relativos à pesquisa sobre padrões de vida.** Rio de Janeiro: IBGE, Departamento de Metodologia, 1997. mimeo.

BRYK, A.S.; RAUDENBUSH, S.W. **Hierarquical Linear Models: applications and data analysis methods.** Newbury Park: Sage Publications, 1992.

CAILLAUX, Elisa L. **Pesquisa sobre padrões de vida 1996-1997.** Rio de Janeiro: IBGE, Departamento de População e Indicadores Sociais, 1996.

CASELLA, G.; GEORGE, E.I. Explaining the Gibbs Sampler. **The American Statistician,** v 46, p. 167-174. 1992.

COCHRAN, W.G. **Sampling techniques.** Nova York: John Wiley and Sons, 1977.

CORRÊA, S. T. **Modelos lineares hierárquicos em pesquisas por amostragem - relacionando o índice de massa corporal às variáveis da pesquisa sobre padrões de vida/IBGE.** Rio de Janeiro, 2001. 85 p. Dissertação (Mestrado) - Escola Nacional de Ciências Estatísticas.

CHROMY, J.R.; ABEYASEKERA, S. Statistical analysis of survey data. **Household surveys in developing and transition countries: Design, implementation and analysis**, 2003.

Disponível em (http://unstats.un.org/unsd/HHsurveys/pdf/Chapter_19.pdf) Acesso em: 01/07/2005.

GAMERMAN, Dani. **Simulação Estocástica Via Cadeias de Markov.** 12° Simpósio Nacional de Probabilidade e Estatística, MG. 1996.

GELFAND, A.E.; GHOSH, S.K. Model choice: A minimum posterior predictive loss approach. **Biometrika,** v. 85, p. 1-11. 1998.

GELMAN, Andrew; et. al. **Bayesian Data Analysis.** Londres: Chapman & Hall, 1995.

GOLDSTEIN, H.I. **Multilevel statistical models.** Londres: Edward Arnold, 1995.

IBGE, **Contas Nacionais Trimestrais.** Série Relatórios Metodológicos número 28, Rio de Janeiro 2004.

IBGE, **Sistema de Contas Nacionais.** Série Relatórios Metodológicos número 24, Rio de Janeiro 2004.

INEP, **Pisa 2000 – Relatório Nacional.** Brasília 2001.

LITTELL, R.C.; MILLIKEN, G.A.; STROUP, W.W. *et al*. **SAS System for Mixed Models.** Cary: SAS Istitute, 1996.

LYBERG, Lars; et. al. **Survey Measurement and Process Quality.** Nova Iorque: Wiley, 1997.

MORAIS, M.P.; CRUZ, B.O. Demand for housing and urban services in Brazil: a hedonic approach. **Texto Para Discussão 946, IPEA**, 2003.

NATIS, Lílian. **Modelos lineares hierárquicos.** São Paulo, 2000. 65 p. Dissertação (Mestrado) - Universidade São Paulo.

NETER, J.; et. al. **Applied Linear Statistical Models.** Chicago: Irwin, 1996.

PESQUISA sobre padrões de vida 1996-1997 [CD-ROM]. Microdados. Rio de Janeiro: IBGE, 1998. 1 disco a laser; 4 ¾ pol.

PESQUISA sobre padrões de vida 1996-1997 [CD-ROM]. Departamento de População e Indicadores Sociais, 2. ed. Rio de Janeiro: IBGE, 1999. 149 p.

PESSOA, D.G.C.; SILVA, P.L. do N. **Análise de dados amostrais complexos.** 13° Simpósio Nacional de Probabilidade e Estatística, MG. 1998.

PFEFFERMANN, D. Small Area Estimation – New Developments and Directions. **International Statistical Review,** v. 70, p. 125-143. 2002.

PFEFFERMANN, D.; MOURA, F.; SILVA, P.L. **Multi-level modelling under informative probability sampling.** Proceedings. International Statistical Institute – International Association of Survey Statisticians – Invited Papers – Seoul – p. 505-524. 2001.

PFEFFERMANN, D.; SKINNER, C.J.; HOLMES, D.J.; GOLDSTEIN, H.; et. al. Weighting for unequal selection probabilities in multilevel models. **Journal of the Royal Statistical Society B,** v. 60, parte 1, p. 23-40. 1998a.

PFEFFERMANN, D.; KRIEGER, A.M., RINOTT, Y. Parametric distributions of complex survey data under informative probability sampling. **Statistica Sinica,** v. 8, p. 1087-1114. 1998b.

REIS, E.J.; TAFNER, P.; REIFF, L.O. Distribuição de riqueza imobiliária e de renda no Brasil: 1992-1999. **Texto Para Discussão, IPEA** v. 75, 2001.

SANTOS, C.H.M.; CRUZ, B.O. A dinâmica dos mercados habitacionais metropolitanos: aspectos teóricos e uma aplicação para a Grande São Paulo. **Texto Para Discussão 713, IPEA**, 2000.

SIDDHARTHA, C.; GREENBERG, E. Understanding the Metropolis-Hastings Algorithm. **The American Statistician,** v. 49, p. 327-335. 1995.

SINGER, J.D. Using SAS PROC MIXED to Fit Multilevel Models, Hierarchical Models and Individual Growth Models. **Journal of Educational and Behavioral Statistics,** p. 323-355. 1998.

SKINNER, C.J.; HOLT, D.; SMITH, T.M.F. (eds). **Analysis of complex surveys.** Chichester: Wiley, 1989.

SNIJDERS, T.A.B; BOSKER, R.J. **Multilevel analysis: an introduction to basic and advanced multilevel modeling.** London: Sage Publications, 1999.

SPIEGELHALTER, D., THOMAS, A., BEST, N.G., and LUNN, D. **Bayesian Inference using Gibbs Sampling.** WinBUGS version 1.4, User manual. MRC Biostatistics Unit, Institute of Public Health, Robinson Way, Cambridge, U.K. 2003.

APÊNDICE

Apêndice 1 - Relação de variáveis explicativas utilizadas no modelo de regressão linear múltipla

01. Condição de ocupação do domicílio
    1. Alugado
    2. Próprio em aquisição
    3. Próprio já pago
    4. Cedido por empregador
    5. Cedido por outra forma
    6. Invasão
02. Tipo de domicílio
    1. Casa
    2. Apartamento
    3. Quarto/cômodo
03. Material predominante nas paredes externas
    1. Alvenaria
    2. Madeira aparelhada
    3. Tijolo sem revestimento
    4. Taipa não-revestida
    5. Madeira aproveitada
    6. Outro
04. Material predominante no piso
    1. Madeira aparelhada
    2. Carpete
    3. Cerâmica, lajota, ardósia
    4. Cimento
    5. Madeira aproveitada
    6. Terra
    7. Outro
05. Material predominante no teto
    1. Telha
    2. Laje de concreto
    3. Madeira aparelhada
    4. Zinco
    5. Madeira aproveitada
    6. Outro
06. Total de dormitórios
    1. Um ou nenhum

2. Dois
3. Três
4. Quatro ou mais

07. Total de cômodos menos dormitórios, cozinha e banheiros
1. Nenhum
2. Um
3. Dois
4. Três ou mais

08. Total de cômodos usados permanentemente como dormitório
1. Um
2. Dois
3. Três
4. Quatro ou mais

09. Existência de cômodos usados exclusivamente para trabalho, estudo, etc.
1. Sim
2. Não

10. Existência de cozinha
1. Sim
2. Não

11. Total de banheiros
1. Nenhum
2. Um
3. Dois
4. Três ou mais

12. Banheiro de uso exclusivo dos moradores
1. Sim
2. Não

13. Localização do banheiro
1. Dentro do domicílio
2. Fora do domicílio

14. Tipo de água usada para beber
1. Filtrada
2. Fervida
3. Mineral
4. Natural

15. Existência de calçada na frente do domicilio
1. Sim
2. Não

16. Tipo de rua onde se localiza o domicilio
1. Asfaltada

2. Paralelepípedo
3. Terra/barro
4. Outro

17. Localização do domicilio
    1. Condomínio regularizado de casas e/ou apartamentos
    2. Favelas ou conjuntos não-regularizados
    3. Casa-de-cômodos ou cortiço
    4. Construção isolada
18. Estado de conservação do domicílio (entrevistador)
    1. Excelente
    2. Bom
    3. Regular
    4. Ruim
19. Existência de água canalizada
    1. Sim
    2. Não
20. Principal forma de abastecimento de água
    1. Rede geral
    2. Poço na propriedade
    3. Poço fora da propriedade
    4. Bica pública
    5. Carro pipa
    6. Outra forma
21. Outra forma de abastecimento de água
    1. Rede geral
    2. Poço na propriedade
    3. Poço fora da propriedade
    4. Bica pública
    5. Carro pipa
    6. Outra forma
    7. Não utiliza
22. Tipo de escoadouro sanitário
    1. Rede coletora de esgoto
    2. Fossa séptica
    3. Fossa rudimentar
    4. Vala
    5. Outro
    6. Não tem
23. Principal tipo de iluminação
    1. Elétrica

2. Gerador
3. Lampião
4. Vela

24. Outro tipo de iluminação
    1. Elétrica
    2. Gerador
    3. Lampião
    4. Vela
    5. Não utiliza
25. Principal combustível para cozinhar
    1. Eletricidade
    2. Gás botijão/canalizado
    3. Querosene
    4. Carvão/lenha
    5. Outro
26. Outro tipo de combustível para cozinhar
    1. Eletricidade
    2. Gás botijão/canalizado
    3. Querosene
    4. Carvão/lenha
    5. Outro
    6. Não utiliza
27. Existência de telefone
    1. Sim
    2. Não
28. Destino do lixo domiciliar
    1. Coletado
    2. Queimado/Enterrado
    3. Jogado em terreno baldio
    4. Jogado em rio, lagoa, etc.
    5. Outro
29. Situação
    1. Urbana
    2. Rural
30. Densidade de moradores por dormitório
31. Mediana do logaritmo do rendimento do setor censitário
32. Proporção de moradores brancos no setor censitário
33. Idade mediana dos moradores do setor censitário
34. Proporção da população com mais de 25 anos e com mais de 11 anos de estudo no setor censitário

Apêndice 2 – Programa com a PROC MIXED do SAS para ajustar modelo linear hierárquico sem considerar que os dados provêm de amostra complexa

```
proc mixed data=base method=reml cl alpha=0.05 covtest;
      format v01b01 v01b01a. v01a01 v01a01a. v01a02 v01a02a.
              v01a03 v01a03a. v01a04 v01a04a. dormit dormit.
              outrosc outrosc. banheiro banheiro. v01a13 v01a13a.
              v01a15 v01a15a. v01b06 v01b06a. v01b10 v01b10a.
              v01b15 v01b15a. v01b17 v01b17a. v01b21 v01b21a.;
      class v01b01 v01a01 v01a02 v01a03 v01a04 dormit outrosc
             banheiro ban_loc v01a13 v01a14 v01a15 v01a17 v01b05
             v01b06 v01b10 v01b15 v01b17 v01b19 v01b21 area;
      model log_alug = v01b01 v01a01 v01a02 v01a03 v01a04 dormit
                       outrosc banheiro ban_loc v01a13 v01a14
                       v01a15 v01a17 v01b05 v01b06 v01b10 v01b15 v01b17
                       v01b19 v01b21 area p_bran med_idad densmor lmedren
                       / solution ddfm=bw;
      random intercept / subject=nsetor;
run;
```

Apêndice 3 – Programa em SAS para ajustar modelo linear hierárquico ponderado pelas probabilidades de inclusão na amostra

```
/* ---------------------------------------------------------------------

Esta macro utiliza o método de mínimos quadrados generalizados iterativo
considerando os pesos amostrais (MQGIPP) para estimar os parâmetros fixos e
aleatórios (e seus respectivos devios padrões) de um modelo multinível (2
níveis: setor e domicílio) com apenas um efeito aleatório de nível 2 no
intercepto.

      Descrição dos parâmetros da macro:
      &bdnivel1 - arquivo de dados de domicílios.
      &bdnivel2 - arquivo de dados de setores.
      &idniv2   - nome da variável que identifica os setores nos arquivos
                  de dados.
      &pesoniv1 - variável do arquivo de domicílios que contém o peso
                  condicional de cada domicílio, dentro do setor.
      &pesoniv2 - variável do arquivo de setores que identifica o peso do
                  setor.
      &coly     - nome da variável resposta.
      &colx     - nomes das variáveis explicativas.
      &iteracao - é o número máximo de iterações a executar para ajustar o
                  modelo.
      &prec     - é a precisão desejada para a convergência das estimativas
                  dos parâmetros.
 --------------------------------------------------------------------- */

%macro
mqgipp(bdnivel1,bdnivel2,idniv2,pesoniv1,pesoniv2,coly,colx,iteracao,prec);
proc iml;
        /* Abre arquivo de dados das unidades de nivel 1 da amostra */
        use &bdnivel1;
         /* Lê valores da variável resposta */
         read all var { &coly } into dadosy;
         /* Lê valores das variáveis explicativas */
         read all var { &colx } into dadosx [colname=nomex];
         /* Lê coluna de valores da variável identificadora de unidades de
          nível 2 a que pertence cada unidade de nível 1 */
         read all var { &idniv2 } into idset;
         /* Lê valores dos pesos condicionais de domicílios */
         read all var { &pesoniv1 } into widi;
         /* Lê vetor com pesos de setor repetidos para os domicílios do
            mesmo setor */
         read all var { &pesoniv2 } into wi_rep;
        close &bdnivel1;

        /* Abre arquivo de dados das unidades de nivel 2 da amostra */
        use &bdnivel2;
         /* Lê pesos dos setores */
         read all var { &pesoniv2 } into pesoset;
        close &bdnivel2;

        /* Inicializa matrizes diversas */
        nset = nrow(pesoset);
        nvar = ncol(dadosx) + 1;
       /* Cria uma matriz de nset linhas e nvar*nvar colunas preenchida por
          zeros */
        mat_t1i = j(nset,nvar##2,0);
```

```
 mat_t3i = j(nset,nvar,0);
/* Cria um vetor de nset linhas preenchido por zeros */
 vec_t6  = j(nset,1,0);
 vec_aux = j(nset,1,0);

/* Cria vetores com os nomes que aparecerao no resultado */
 nome1 = {const}||{&colx};
 nome2 = {Estimativa DP Z_set};
 nome3 = {var_u var_e};

 /* Calcula vetor lambdai que será usado como escala para os pesos
    parciais de domicílio */
/* Processa um setor por vez */
 do linha = 1 to nset by 1;
  ndom  = loc(idset=linha);
  np     = ncol(ndom);
  k      = widi[ndom,][:,];
  parte  = j(np,1,k);
  if linha = 1 then lambidai = parte;
  else lambidai = lambidai // parte;
 end;
 /* Calcula os pesos parciais de domicílio escalados pelo método 2*/
 widi_set = widi / lambidai;
 wi_set   = pesoset / (pesoset[:,]);
 wiset_r  = wi_rep / (pesoset[:,]);
 /* Calcula valor inicial de beta */
 do linha = 1 to nset by 1;
  /* Processa um setor por vez */
  ndom   = loc(idset=linha);
  yi     = dadosy[ndom,];
  x0i    = dadosx[ndom,];
  np     = nrow(x0i);
  xi     = j(np,1,1) || x0i;
  diag   = diag(widi_set[ndom,]);
  wi     = wi_set[linha,];
  /* Calcula as somas parciais para cada setor */
  mat_t1i[linha,]  = rowvec( ( xi` * diag * xi ) # wi );
  mat_t3i[linha,]  = rowvec( ( xi` * diag * yi ) # wi );
 end;
 somat1   = shape(mat_t1i[+,],nvar);
 somat3   = mat_t3i[+,];
 /* Calcula o vetor inicial de betas  */
 herm=hermite(somat1);
 if herm=i(nvar) then beta0 = inv(somat1) * (somat3`);
 else beta0 = ginv(somat1) * (somat3`);
 /* Calcula valor inicial de teta - v(u0j) e v(eij) */
 do linha = 1 to nset by 1;
  /* Processa um setor por vez */
  ndom  = loc(idset=linha);
  yi     = dadosy[ndom,];
  x0i    = dadosx[ndom,];
  np    = nrow(x0i);
  xi    = j(np,1,1)  || x0i;
  wi    = wi_set[linha,];
  /* Calcula os resíduos de nível 1 para cada domicílio de cada
     setor */
  resid = yi - xi * beta0;
  /* Calcula o resíduo de nível 2 - setor */
  ui0 = (widi_set[ndom,]` * resid)/widi_set[ndom,][+,];
  /* Calcula as matrizes parciais para o cálculo das variâncias*/
  vec_t6[linha,] = (widi_set[ndom,]` * ((resid - ui0) ## 2)) # wi;
```

```
   vec_aux[linha,]= wi # ( widi_set[ndom,][+,]- 1 );
  end;
  /* Calcula as variâncias iniciais dos efeitos aleatórios */
  teta0 = 0.000001 // ( vec_t6[+,] / vec_aux[+,] );


/*  Módulo que executa o procedimento iterativo MQGIPP */
 /* início do procedimento iterativo */
 matp = j(nset,nvar##2,0);
 matq = j(nset,nvar,0);
 r11  = j(nset,1,0);
 r12  = j(nset,1,0);
 r22  = j(nset,1,0);
 s11  = j(nset,1,0);
 s21  = j(nset,1,0);
 beta_ant = beta0;
 beta     = beta_ant # 2;
 teta_ant = teta0;
 teta     = teta_ant # 2;
 itera=1;
 aux1 = max( abs( (beta-beta_ant)/beta_ant ) );
 aux2 = max( abs( (teta-teta_ant)/teta_ant ) );
 do while ( itera <= &iteracao & (aux1 + aux2 > &prec) );
  /* print itera; */
  /* Processo iterativo para o cálculo de beta */
  do linha = 1 to nset by 1;
   /* Processa um setor por vez */
   ndom  = loc(idset=linha);
   yi     = dadosy[ndom,];
   x0i    = dadosx[ndom,];
   np    = nrow(x0i);
   xi    = j(np,1,1) || x0i;
   v     = widi_set[ndom,];
   diag  = diag(v);
   wi    = wi_set[linha,];
   t5i   = v[+,];
   if itera = 1 then ai=(teta0[1,] / (teta0[1,] # t5i + teta0[2,]));
   else  ai = ( teta[1,] / ( teta[1,] # t5i + teta[2,] ) );
   t1i      =  ( xi` * diag * xi );
   t2i      =  ( xi` # v` )[,+];
   t3i      =  ( xi` * diag * yi );
   t4i      =   v` *  yi;
   matp[linha,] = rowvec( wi # (t1i - ai # t2i * t2i`) );
   matq[linha,] = rowvec( wi # (t3i - ai # t2i # t4i ) );
  end;
  /* Calcula beta estimado */
  s_matp  = shape(matp[+,],nvar);
  s_matq  = matq[+,];
  if itera ^= 1 then beta_ant = beta;
  /* Verifica se a matriz é inversível */
  herm = hermite(s_matp);
  /* Calcula vetor de beta estimado */
  if herm = i(nvar) then beta = inv(s_matp) * (s_matq`);
  else  beta = ginv(s_matp) * (s_matq`);
  /* Processo iterativo para o cálculo de teta = inv(R) x S -------*/
  do linha = 1 to nset by 1;
   /* processa um setor por vez */
   ndom  = loc(idset=linha);
   yi    = dadosy[ndom,];
   x0i   = dadosx[ndom,];
   np    = nrow(x0i);
```

```
   xi    = j(np,1,1)  || x0i;
   wi    = wi_set[linha,];
   v     = widi_set[ndom,];
   t5i   = v[+,];
   eij = yi - xi * beta;
   ui  = (v` * eij) / t5i;
   vij = eij - ui;
   t6i = v` * (vij ## 2);
   if itera = 1 then do;
    bi  = t5i /  ( teta0[1,] # t5i + teta0[2,] );
    r22[linha,] = wi #(teta0[2,]##(-2)#(v[+,]-1)+(bi##2)/(t5i##2));
    s21[linha,] = wi #(teta0[2,]##(-2)#t6i+(bi##2)#(ui##2)/t5i);
   end;
   else do;
    bi  = t5i /  ( teta[1,] # t5i + teta[2,] );
    r22[linha,] = wi#(teta[2,]##(-2)#(v[+,]-1)+(bi##2)/(t5i##2));
    s21[linha,] = wi#(teta[2,]##(-2)#t6i+(bi##2)#(ui##2)/t5i);
   end;
   r11[linha,] = wi # (bi ## 2);
   r12[linha,] = wi # (bi ## 2) / t5i;
   s11[linha,] = wi # (bi ## 2) # (ui ## 2);
  end;
  /* Soma para todos os setores e retorna à forma original da matriz
     r 2x2 */
  matr = r11[+,] || r12[+,] || r12[+,] || r22[+,];
  mats = s11[+,] // s21[+,];
  r = shape(matr,2);
  s = mats;
  /* Calcula o vetor de estimativas das variâncias de u0j e eij */
  if itera ^= 1 then teta_ant = teta;
  /* Verifica se a matriz é inversível */
  herm = hermite(r);
  /* Calcula o vetor teta estimado */
  if herm = i(2) then teta = inv(r) * s;
  else teta = ginv(r) * s;
  itera = itera + 1;
  aux1 = max( abs( (beta-beta_ant)/beta_ant ) );
  aux2 = max( abs( (teta-teta_ant)/teta_ant ) );
  /* Fim do procedimento iterativo */
 end;
 /* Calcula o número de iterações */
 n_it = itera - 1;
 /* print n_it; */
 /* Estima as variâncias de beta e teta */
 /* Inicializa matrizes*/
 mat_c = j(nset,nvar##2,0);
 mat_d = j(nset,4,0);
 do linha = 1 to nset by 1;
  /* Processa um setor por vez */
ndom  = loc(idset=linha);
  yi     = dadosy[ndom,];
  x0i    = dadosx[ndom,];
  np    = nrow(x0i);
  xi    = j(np,1,1)  || x0i;
  wi    = wi_set[linha,];
  v     = widi_set[ndom,];
  di    = diag(v);
  /* Cálculos parciais para a variância de beta */
  ei = yi - xi * beta;
  t5i= v[+,];
  ai = teta[1,]/(teta[2,] + teta[1,] # t5i);
```

```
         t2i= (xi` # v`)[,+];
         ci = (xi` * di * ei) - ( ai # t2i # (v`* ei) );
         mat_c[linha,] = rowvec( (wi##2) # (ci * ci`) );
         /* Cálculos parciais para a variância de teta */
         bi   = t5i / (teta[1,]# t5i + teta[2,]);
         ui   = (v`* ei) / t5i;
         vij = ei - ui;
         t6i = v` * (vij ## 2);
         d11  = (bi##2) # ( ui ## 2 - teta[1,] - teta[2,]/t5i );
         auxd = (teta[2,]##(-2)) # (t6i - (v[+,] - 1) # teta[2,] );
         d2i  = d11  //  ( auxd + d11/t5i );
         mat_d[linha,] = rowvec( (wi##2) # (d2i * d2i`) );
        end;
        /* Calcula as variâncias de beta e teta */
        /* Testa se matriz é inversível */
        herm = hermite(s_matp);
        if herm = i(nvar) then
           var_beta = inv(s_matp)*((nset/(nset-1))#shape(mat_c[+,],nvar))
                      * inv( s_matp );
        else var_beta = ginv(s_matp)*((nset/(nset-1)) #
                        shape(mat_c[+,],nvar) ) * ginv( s_matp );
        dp_beta  = sqrt(vecdiag(var_beta));
        /* Testa se matriz é inversível */
        herm = hermite(r);
        if herm = i(2) then
             var_teta = inv(r)*((nset/(nset-1))#shape(mat_d[+,],2))*inv(r);
        else var_teta = ginv(r)*((nset/(nset-1))#shape(mat_d[+,],2))*
                        ginv(r);
        /* Calcula o desvio padrao de teta */
        dp_teta  = sqrt(vecdiag(var_teta));

        /* Calcula os Zs */
        z_set = beta/dp_beta;
        z_set2 = teta/dp_teta;

        /* Impressao dos resultados */
        result1 = beta || dp_beta || z_set;
        result2 = teta || dp_teta || z_set2;

        print "N. Iteraçoes:" n_it;
        print result1 [rowname=nome1 colname=nome2];
        print result2 [rowname=nome3 colname=nome2];

quit;
%mend mqgipp;

title 'Ajuste por MQGI';
%mqgipp(tese.base2,tese.nivel2,nsetor,teste2,teste1,log_alug,
x1 x2 x3 x4 x5 x6 x7 x8 x9 x10 x11 x12 x13 x14 x15 /*x16*/ x17 x18
x19 x20 x21 /*x22 x23 x24*/ x25 x26 x27 x28 /*x29*/ x30 p_bran med_idad
densmor lmedren,
50,0.000001);

title 'Ajuste por MQGIPP';
%mqgipp(tese.base2,tese.nivel2,nsetor,peso2,peso1,log_alug,
x1 x2 x3 x4 x5 x6 x7 x8 x9 x10 x11 x12 x13 x14 x15 /*x16*/ x17 x18
x19 x20 x21 /*x22 x23 x24*/ x25 x26 x27 x28 /*x29*/ x30 p_bran med_idad
densmor lmedren,
50,0.000001);
```