Universidade Federal de Santa Catarina
Departamento de Informática e Estatística
Curso de Sistemas de Informação

APLICAÇÃO DE MODELOS LINEARES HIERÁRQUICOS PARA PREDIÇÃO DO PAGAMENTO DE IPTU-DA DO MUNICÍPIO DE ITAJAÍ.

Trabalho de Conclusão de Curso

Josué Celso Cim

UFSC 2008

JOSUÉ CELSO CIM

APLICAÇÃO DE MODELOS LINEARES HIERÁRQUICOS PARA PREDIÇÃO DO PAGAMENTO DE IPTU-DA DO MUNICÍPIO DE ITAJAÍ.

Trabalho de conclusão de curso apresentado
como parte dos requisitos para obtenção do
grau de Bacharel em Sistemas de Informação

Orientador: Paulo José Ogliari, Dr.
Co-Orientador: Dalton Francisco de Andrade, PhD.

Florianópolis
2008

JOSUÉ CELSO CIM

APLICAÇÃO DE MODELOS LINEARES HIERÁRQUICOS PARA PREDIÇÃO DO PAGAMENTO DE IPTU-DA DO MUNICÍPIO DE ITAJAÍ.

Trabalho de conclusão de curso apresentado
como parte dos requisitos para obtenção do
grau de Bacharel em Sistemas de Informação

Orientador:

_____
Paulo José Ogliari

Banca Examinadora:

_____
Dalton Francisco de Andrade

_____
João Goulart Júnior

Florianópolis
2008

Dedicatória

A minha mãe, Márcia e ao meu pai Celso,

Pelo apoio, carinho, incentivo e lições concedidas

durante toda minha vida.

Agradecimentos

À minha família, pelo apoio e compreensão.

Ao meu orientador, Prof. Paulo José Ogliari, que apoiou o desenvolvimento do trabalho, pela orientação cientifica, pela confiança depositada em mim e no trabalho.

Ao Prof. Dalton Francisco de Andrade, pela atenção, pela valiosa co-orientação e esclarecimentos.

Ao Prof. João Goulart Júnior, membro da banca avaliadora, pelas contribuições e aperfeiçoamento deste trabalho.

A todos que direta ou indiretamente contribuíram para a conclusão deste trabalho, entre eles professores, colegas de trabalho e amigos.

Resumo

CIM, Josué Celso. Aplicação de Modelos Hierárquicos para a Predição do Pagamento de IPTU-DA do Município de Itajaí. 2008. 79f. Trabalho de Conclusão de Curso – Sistemas de Informação, UFSC, Florianópolis.

Este trabalho propõe um método de classificação dos lançamentos de IPTU-DA, fundamentado nas características dos terrenos ou dos imóveis, por meio de Modelos Lineares Hierárquicos Generalizados, que efetivamente considera e avalia, de forma sistematizada, a correlação de medidas existentes entre os bairros. Os dados de trabalho são referentes aos tributos de IPTU-DA do exercício de 2006 do município de Itajaí, SC. A partir da estimação dos modelos propostos e da utilização das variáveis preditoras individuais dos terrenos, imóveis e bairros; foram devidamente quantificadas e explicadas às variabilidades nas respostas de interesse aqui enfocadas. Por fim, os modelos foram aplicados em uma base de dados para validação, e então avaliados, em termos de capacidade preditiva. Os principais resultados alcançados pela condução deste trabalho foram: a agregação de conhecimento do comportamento tributário e dos Modelos Lineares Hierárquicos Generalizados; A incapacidade das variáveis selecionadas, construírem modelos hierárquicos com um bom índice de predição da situação de pagamento do tributo de IPTU-DA.

Palavras Chave: modelos lineares hierárquicos generalizados, IPTU-DA, predição, situação do pagamento.

Abstract

CIM, Josué Celso. Aplicação de Modelos Hierárquicos para a Predição do Pagamento de IPTU-DA do Município de Itajaí. 2008. 80f. Trabalho de Conclusão de Curso – Sistemas de Informação, UFSC, Florianópolis.

This work proposes a method of classifying the releases of ITPU-DA. It is based on characteristics of the land or construction. The method uses a Hierarchical Generalized Linear Models which can effectively considers and evaluates, in a systematic way, the correlation between measures of the different districts. The data used in this work refers to taxes from the 2006 - IPTU-DA of the city Itajaí, SC. After the model has been created, it was tested to verify its validity. The main results of this study were: the aggregation of knowledge about the tributary behavior and the generalized hierarchical linear models; The inability of selected variables to build hierarchical models with a good índex which could predict the situation for the payment of tribute IPTU-DA.

Key-Words: generalized hierarchical linear models, IPTU-DA, situation of the payment, prediction.

Sumário

Lista de Figuras

Lista de Quadros

Lista de Tabelas

Lista de Reduções

ANCOVA – Análise de Covariância.

ANOVA – Análise de Variância.

CCIC – Coeficiente de Correlação Intraclasse.

DM – *Data Mining.*

EUA – Estados Unidos da América.

IPTU – Imposto Predial e Territorial Urbano.

IPTU-DA – Imposto Predial e Territorial Urbano – Dívida Ativa.

ISS – Imposto sobre Serviço.

KDD – *Knowledge Discovery in Database*.

LRF – Lei de Responsabilidade Fiscal.

MD – Mineração de Dados.

MHL – Modelo Linear Hierárquico.

MLGH – Modelo Linear Generalizado Hierárquico.

MQG - Mínimos Quadrados Generalizados.

MQO – Mínimos Quadrados Ordinais.

MV – Máxima Verossimilhança.

MVR – Máxima Verossimilhança Restrita.

SC – Santa Catarina.

SGBD – Sistemas de Gerência de Banco de Dados.

SQL – *Structured Query Language.*

TI – Tecnologia da Informação.

1- INTRODUÇÃO

A atual contextualização da administração tributária requer um controle a curto, médio e longo prazo, dos níveis de arrecadação de tributos e contribuições administrados.

Um planejamento pró-ativo pode ser um dos mediadores para o alcance do controle dos fluxos de caixa públicos, e nesse sentido os sistemas de tributação, fiscalização e arrecadação devem ser contemplados pelos gestores dos órgãos tributários (Barreto, 2005).

A prefeitura municipal de Itajaí, **SC**, responsável pela administração, fiscalização e arrecadação de impostos, procura acompanhar os recentes avanços em **TI**, pois hoje é um fator determinante para suas possibilidades futuras de sucesso econômico e social e, sendo assim, fazendo parte do atual panorama econômico e tecnológico brasileiro o qual tem se mostrado eficiente tanto quanto países de primeiro mundo.

Com a informatização de todo seu sistema de tributação, a prefeitura municipal de Itajaí contém uma grande base de dados, que favorece o desenvolvimento e a aplicação de metodologias para a solução de problemas, procurando identificar informações relevantes e transformando-as em conhecimento útil principalmente no suporte e apoio à decisão.

Recentemente os modelos lineares hierárquicos ou multiníveis começaram a ser utilizados em diversos campos do conhecimento científico. Estes modelos possuem uma estrutura que permitem uma interpretação mais detalhada dos efeitos relacionados com os diferentes níveis da hierarquia natural dos dados. Sendo que no presente trabalho, será constatado se o nível bairro (localidade)

influência na quitação dos débitos de determinado imóvel ou lote localizado no município de Itajaí.

Importantes propriedades dos modelos lineares hierárquicos permitem que a variabilidade da variável resposta seja explicada através de variáveis preditoras incluídas em diferentes níveis hierárquicos e sendo possível quantificar em cada nível, a proporção da variabilidade explicada. Assim, as análises multiníveis apresentam estimativas mais fiéis, já que não pressupõe a independência entre as observações. Estas propriedades serão vistas com mais detalhes nos próximos capítulos deste trabalho.

Segundo BRYK & RAUDENBUSH (1992), esses modelos não são uma solução para todos os problemas, contudo, eles representam um grande passo para auxiliar as análises por serem estatisticamente corretos e não desperdiçarem informação.

Assim o presente trabalho procura contribuir, com um método de classificação dos contribuintes, a partir das características da situação de pagamento dos seus tributos, fundamentado na previsão de seus comportamentos tributários por meio de modelos lineares generalizados hierárquicos, para dar subsídios para a melhoria da classificação é a principal justificativa da realização deste trabalho.

1.1 - Objetivos

1.1.1 - Geral

O objetivo geral é construir dois **MLGHs** capazes de estimar se um imóvel ou um lote, do município de Itajaí, terá a tendência de não pagar ou pagar suas dívidas relativas ao tributo **IPTU-DA**. Serão consideradas as características do terreno, do imóvel (quando o terreno possui construção), e do bairro em que o mesmo está situado como variáveis preditoras do modelo.

1.1.2 - Específicos

I – Levantar, pré-processar e analisar exploratoriamente os dados relevantes à pesquisa.

II – Identificar as variáveis preditoras (independentes) significativas para comporem o modelo.

III – Verificar se o modelo proposto atende todas as suposições teóricas consideradas inicialmente para a sua existência.

IV – Implementar os modelos e verificar a sua utilidade.

1.2 – Justificativa

Como em uma empresa privada, o fluxo de recursos essencial para o seguimento funcional da organização é a receita. Nos órgãos públicos essa receita

é denominada receita pública, a qual consiste no recolhimento de bens aos cofres públicos. O complexo de problemas que se concentram em torno do processo de receitas-despesas do governo é denominado de finanças públicas.

Atualmente com a implantação da Lei de Responsabilidade Fiscal (**LRF**), nos âmbitos municipal, estadual e federal, exige que os órgãos públicos controlem com mais rigor suas finanças. A **LRF** consiste em um mecanismo legal de delimitação de conduta dos administradores públicos, fixando-lhes normas e limites no gerenciamento financeiro dos recursos públicos. Ela tem como objetivo básico à melhoria da administração das contas públicas, onde os administradores deverão ter compromisso com o orçamento e com as metas aprovadas pelo Legislativo;

Para tanto, a **LRF** no seu art. 1º, § 1º explicita que:

> *(...) a responsabilidade na gestão fiscal pressupõe a ação planejada e transparente, em que se previnem riscos e corrigem desvios capazes de afetar o equilíbrio das contas públicas, mediante o cumprimento de metas de resultados entre receitas e despesas e a obediência a limites e condições no que tange a renúncia à receita, geração de despesas de pessoal, da seguridade social e outras, dívidas consolidada e mobiliária, operações de crédito, inclusive por antecipação de receita, concessão de garantia e inscrição em Restos a Pagar.*

Como a **LRF** prega o planejamento orçamentário, este trabalho surge como uma alternativa para auxiliar na previsão das receitas referentes aos tributos de **IPTU** em dívida ativa, sempre levando em consideração as características dos lotes, imóveis e dos bairros. Sendo assim, tenciona ser uma opção decisória nos momentos que exijam pronta reação por parte da prefeitura. Políticas para o aumento na arrecadação em bairros, ou determinados conjuntos de edificações poderão ser mais eficientes. Indicará novos critérios para obras de infra-estrutura dos bairros, etc.

1.3 - Metodologia

Com a finalidade de alcançar o objetivo geral do trabalho, que é a obtenção dos modelos preditivos, a metodologia de desenvolvimento desse trabalho pode ser dividida em três partes. A fundamentação teórica, a aplicação do processo de descoberta de conhecimento, e a construção do modelo e sua aceitação.

Após a conclusão da fundamentação teórica é dado inicio as atividades do processo de descoberta de conhecimento por meio de análise da base de dados e da seleção, preparação e transformação das variáveis a serem consideradas na etapa da construção do modelo.

Sobre as variáveis preditoras, leva-se em conta que o modelo linear hierárquico considera as variáveis referentes ao comportamento individual dos lotes ou imóveis, e também justificam a variabilidade da resposta entre os bairros.

Os dados disponíveis (objeto de estudo) para esta pesquisa são: os relacionados às informações dos lotes e dos imóveis, os com os bairros do município, e os referentes aos lançamentos de **IPTU-DA** do exercício de 2006.

E, finalmente quanto aos aspectos de amostragem, com o objetivo de viabilizar o processo tradicional em modelagem estatística, os dados foram previamente divididos em duas amostras: estimação e validação.

Para finalizar as atividades aqui propostas, faz-se necessário a implantação do modelo e conclusões.

1.4 – Limitações da Pesquisa

Dentre os vários modelos estatísticos que podem ser utilizados, restringiu-se nesta pesquisa, ao uso de modelos lineares hierárquicos.

Além disso, o método de classificação proposto e aplicado por este trabalho se limita apenas as dívidas ativas de **IPTU-DA**, referente ao exercício de 2006, dos imóveis e lotes do município de Itajaí. A opção de limitar a pesquisa, a esse tipo de contribuinte e tributo, deve-se principalmente à sua relevância fiscal.

1.5 - Estrutura do trabalho

O relatório do **TCC** está estruturado em cinco capítulos, mais as referências bibliográficas e os apêndices, construídos de forma a facilitar o entendimento e compreensão do leitor desde os objetivos até a conclusão.

O primeiro capitulo, denominado de introdução, faz uma contextualização do assunto, cita o problema, os objetivos, justificativas, os métodos para o desenvolvimento do trabalho e sua estrutura.

O segundo capítulo trata de uma revisão da literatura sobre pré-processamento de dados e modelos lineares hierárquicos. Apresenta os principais conceitos e definições relacionadas aos mesmos.

No terceiro capítulo, estão expostas as atividades efetuadas na base de dados, tais como transformações de variáveis, análises estatísticas, seleção entre outros.

O capítulo quatro apresenta a metodologia estatística necessária para a construção e o diagnóstico do modelo. Aplica o modelo e faz a sua validação.

O quinto capítulo faz uma conclusão com a base nos estudos realizados nos capítulos anteriores e apresenta sugestões para outras possíveis analises que podem ser realizadas.

2 – FUNDAMENTAÇÃO TEÓRICA

Esse capítulo expõe as considerações provenientes da pesquisa bibliográfica do trabalho distribuída em duas partes. A primeira parte diz-se respeito ao processo de descoberta de conhecimento (**KDD**). A segunda parte aborda alguns aspectos relacionados aos Modelos Lineares Hierárquicos, onde os mesmos serão apresentados e formalizados, seguindo principalmente as noções descritas em Raudenbush e Bryk (2002) e Bryk e Raudenbush (1992). Estas seções são fundamentais para o entendimento e acompanhamento do processo de execução do presente trabalho.

2.1 – Processo de Descoberta de Conhecimento

Ao longo das ultimas três décadas, muitas organizações produziram grande quantidade de dados, sendo armazenados e processados utilizando a tecnologia de banco de dados através da linguagem **SQL**, que é a principal linguagem para acesso e manipulação dos dados na maioria dos bancos de dados transacionais (Kamber, 2001). Como o acesso à informação é um fator de sucesso competitivo para a grande parte das organizações, cada vez mais gerentes e executivos precisam obter com rapidez e facilidade as informações sobre o seu negócio para auxiliarem em suas decisões. É impraticável a obtenção destas informações pelos profissionais do nível tático e estratégico da organização através da utilização da linguagem **SQL**, pois ela é altamente estruturada e pressupõe que seu usuário esteja ciente da estrutura do banco de dado. A linguagem **SQL** é mais utilizada por

profissionais predominantemente do nível operacional da empresa (Machado, 2005).

Baseado nesse contexto, e além do desafio de implementar técnicas que consigam mensurar e descobrir padrões relevantes na base de dados, resultante, sobretudo, do aumento da complexidade das tarefas operacionais e decisórias. A partir disto, uma nova perspectiva é apresentada, em que a análise de dados é vista com um caráter exploratório. De maneira mais abrangente, encontra-se *Knowledge Discovery in Database* (**KDD**), sendo esta a designação para o processo que envolve a seleção, o pré-processamento e a transformação dos dados, bem como a aplicação de algoritmos, a interpretação dos resultados e a geração de conhecimento (Figura 1).

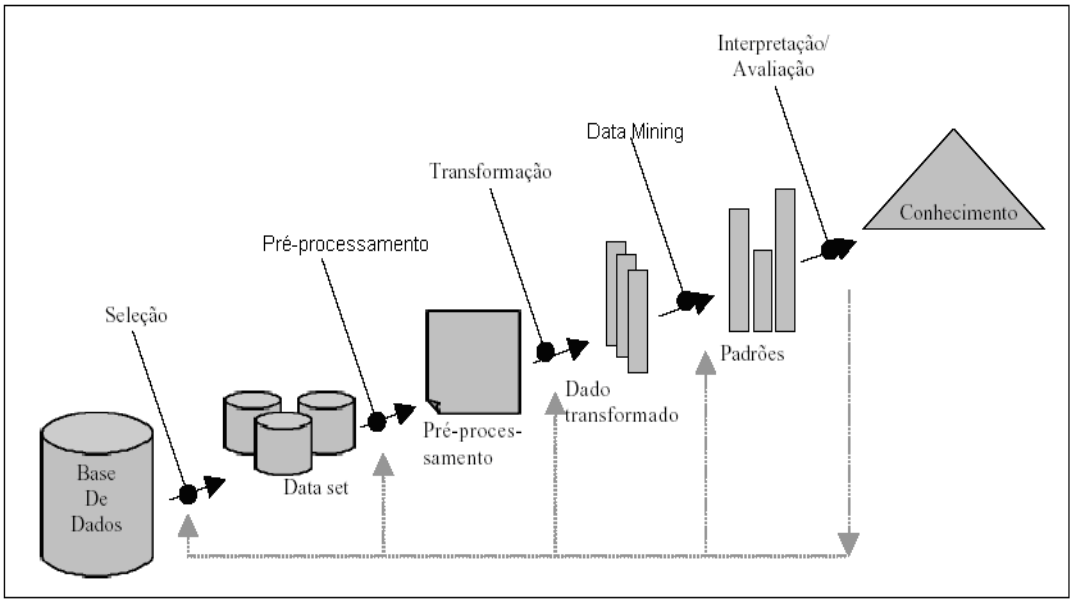
O processo de descoberta de conhecimento é um composto por diversas etapas, envolvendo metodologias e técnicas de *data mining*. O seu objetivo é o de “aperfeiçoar e automatizar o processo de descrição das tendências e dos padrões contidos nesse processo, potencialmente úteis e interpretáveis” (Ogliari, 2004).

Dentro desse processo encontra-se *Data Mining,* ou Mineração de Dados (**MD**), que é uma etapa do **KDD**, responsável pela aplicação dos algoritmos com a finalidade de identificar padrões sobre uma base de dados (Fayyad, 1996b), ou gerar um conjunto de regras que descrevam o comportamento de uma base de dados.

Segundo PRASS (2004), o processo de descoberta de conhecimento compreende todo o ciclo que os dados percorrem até virar informação.

Nas seções a seguir, resumidamente, são apresentados os conceitos do processo de descoberta de conhecimento com suas etapas e metodologias.

Figura 2.1 - Processo de KDD. Adaptado de Fayyad et al. (1996a)



2.1.1 – Seleção de Dados

Nessa primeira fase do processo de descoberta de conhecimento devem ser definidos os objetivos da análise. A definição dos objetivos é realizada com o auxilio de um especialista no assunto, pois é difícil fazer a seleção adequada sem ter um bom domínio do problema em estudo e, muitas vezes, é dessa escolha de variáveis que depende o resultado bem sucedido de todo o **KDD** (Ogliari, 2004).

Em seguida deve-se buscar conhecimento sobre as fontes dos dados, conhecendo sua estrutura e verificando como estes dados podem ser utilizados na mineração. A partir da seleção dos dados, estes deverão ser organizados e armazenados em uma nova base de dados para análise, que pode ser mantida por um **SGBD** ou ser apenas um único arquivo.

2.1.2 – Pré-Processamento

A fase de pré-processamento é importante para o *data wharehouse* e para *data mining,* pois nessa fase são identificados e corrigidos os problemas presentes nos dados selecionados. Quando existem distorções, como dados inconsistentes, ou valores discrepantes gerados devido a erro na entrada dos dados, ou ainda falta de valores para alguns campos importantes para a mineração e que não eram tão importantes na entrada de dados, torna-se necessário tratar os dados.

Para se obter bons resultados no **KDD**, além da correta seleção das variáveis a serem utilizadas nas analises, devem-se ter os dados limpos e corretos, pois as ferramentas de mineração de dados são altamente sensíveis a ruídos nos dados. A identificação desses problemas pode ser obtida através da analise exploratória dos dados (Barreto, 2005).

A análise exploratória dos dados “emprega técnicas estatísticas descritivas e gráficas para estudar um conjunto de dados, detectando *outliers* e anomalias, e testando as suposições do modelo” (Ogliari, 2004).

Resumidamente, o pré-processamento dos dados tem por objetivo assegurar a qualidade dos dados selecionados.

2.1.3 – Redução, Transformação e Integração dos Dados.

Após os dados desejados terem sido selecionados e os dados a serem minerados tenha sido identificado, normalmente é necessário realizar algumas transformações nestes dados. As transformações variam de conversão de tipos de dados, conversão de valores nominais em valores numéricos, definição de novos atributos aplicando-se operadores matemáticos ou lógicos, etc. Resultando em

dados integrados e transformados, ou seja, estavam em um formato padrão para a utilização.

Podemos empregar algumas técnicas para transformar os dados, entre elas a discretização, agregação, normalização, alisamento entre outras.

A integração dos dados segundo (Barreto, 2005), consiste em reunir num único arquivo (semelhante a uma planilha eletrônica) ou tabela de banco de dados, todos os dados de diferentes fontes que foram selecionados para utilização nas análises. Na integração deve-se ter cuidado para não incluir variáveis duplicadas (que podem ter diferentes nomes em cada fonte de dados selecionada) e variáveis redundantes (que podem ser verificadas através da análise de correlação). Tomando esses cuidados se contribui para aumentar a precisão e a velocidade nos processos de mineração de dados.

2.1.4 – Data Mining

Mineração de dados é a fase do **KDD**, onde são aplicadas as técnicas de busca de conhecimento. Esta fase é caracterizada pela busca de padrões de interesse em uma forma particularmente representativa ou em um conjunto dessas representações. Nesse momento que se descobrem novas relações, não identificáveis a olho nu, mas que podem ser visualizadas com Técnicas de Inteligência Artificial e Técnicas Estatísticas, por meio de uma análise sistemática e exaustiva sobre a base de dados.

Um conceito muito difundido e errado sobre mineração de dados é o que define os sistemas de mineração de dados como sistemas que podem automaticamente minerar todos os conceitos valiosos que estão escondidos em um

grande banco de dado sem intervenção ou direcionamento humano (Kamber, 2001).

Para nós, mineração de dados é um processo altamente cooperativo entre homens e maquinas que visa à exploração de grandes bancos de dados, com o objetivo de extrair conhecimentos através do reconhecimento de padrões e relacionamento entre variáveis, conhecimento esses que possam ser obtidos por técnicas comprovadamente confiáveis e validados pela sua expressividade estatística.

Pode-se citar como exemplos de metodologias de mineração de dados o *teste de hipótese*, *descoberta supervisionada de conhecimento, descoberta não supervisionada de conhecimento.* Para maiores detalhes sobre metodologias e técnicas de mineração de dados pode-se consultar sobre o assunto em Kamber (2001).

Uma grande variedade de técnicas analíticas tem sido utilizada em mineração de dados, técnicas que vão desde as tradicionais da estatística multivariada, como análise de agrupamentos e regressões, até modelos mais atuais de aprendizagem, como redes neurais, lógica difusa e algoritmos genéticos, tudo dependendo da metodologia adotada.

No presente trabalho utilizaremos a metodologia de *teste de hipótese* junto com a tarefa de *previsão e predição,* que consiste em determinar o valor que uma variável (situação de pagamento) ira assumir no futuro tendo como base seus valores anteriores, ou seja, os valores que essa variável assumiu em períodos de tempo passados são utilizados para predizer seu valor futuro.

2.1.6 – Análise e Interpretação

A etapa de pós-processamento ocorre quando é concluída a execução do algoritmo de mineração de dados sobre a base de dados pré-processada, utilizando os parâmetros definidos. O resultado será um conjunto de padrões, ou modelo, que descreve os dados. Nessa etapa pode-se também chegar à conclusão de que o modelo obtido não atende às expectativas, ou seja, ao objetivo definido inicialmente. Neste caso, é necessário analisar todo o processo de **KDD** e identificar qual passo deve ser revisto e refeito. Dessa forma, os passos subseqüentes ao passo refeito também devem ser refeitos para que um novo modelo seja obtido e também avaliado.

Esta avaliação deve ser feita em conjunto com um especialista no assunto, devem-se interpretar os resultados e analisá-los quanto a sua relevância e qualidade.

2.2 – Modelos Lineares Hierárquicos Generalizados

2.2.1 – Introdução

Esta seção apresentará à lógica e as formulações referentes aos Modelos Lineares Hierárquicos, que se caracterizam por possuir uma estrutura de hierarquia aos modelos lineares e que são também denominados como Modelos Multiníveis. Estas idéias estão amplamente desenvolvidas em Bryk e Raudenbush (1992). Porém, a denominação de “Modelos Lineares Hierárquicos é bem mais antiga e, de acordo com Natis (2000, p.3), ela surgiu originalmente como fruto dos trabalhos de Lindley e Smith (1972) e Smith (1973) sobre a estimação Bayesiana de modelos lineares”.

Entretanto os estudos anteriores apresentavam muitas vezes problemas de cálculo e imprecisão nas estimativas, acarretando em um desestímulo na exploração desses modelos. Contudo avanços estatísticos isoladas foram reunidos de forma a aperfeiçoar as estimativas hierárquicas. Em Natis (2000) pode ser consultada uma breve cronologia das pesquisas em estatísticas ao longo das ultimas três décadas.

Este capítulo, então se propõe a abordar tópicos referentes à **MLGH**, inicialmente fazendo um breve estudo sobre modelos lineares, posteriormente modelos lineares hierárquicos e concluindo com os modelos lineares generalizados hierárquicos.

2.2.2 - Modelos Lineares

Os modelos lineares têm por objetivo analisar a influência que uma determinada variável *Y* (variável dependente) sofre ao ser afetado por outras variáveis (variáveis independentes ou explicativas) por intermédio de uma relação linear. Em todos os casos, temos a presença de variáveis que ajudam a explicar a variação da variável de interesse. Denotamos por *Y* a variável dependente e *X*1, *X*2,... *Xp* as variáveis explicativas, todas com *n* observações. Assim, temos que:

$$\gamma_i = \beta_1\chi_{1\,i} + \beta_2\chi_{2\,i} + \beta_3\chi_{3\,i} + ... + \beta_p\chi_{p\,i} + \varepsilon_i, \text{ onde } i = 1,..,n,$$

em que γ_i denota a *i*-ésima observação da variável *Y* ; *xij* denota a *i*-ésima observação da variável *Xj*, *j* = 1, ..., *p*, ε_i denota o *i*-ésimo erro, com (i = 1, ...,n) ~ $N(0;\Sigma)$; e ($\beta_{1,\,....}\,\beta_p$) são parâmetros desconhecidos.

Este modelo é chamado de *modelo linear* ou *modelo de regressão linear*. Dizemos que o modelo é “simples” quando existe apenas uma variável explicativa, e “múltiplo” quando existem mais de uma variável explicativa.

Modelos lineares são aplicados quando os termos ε_i são considerados como não correlacionados, ou seja, com média zero e variância constante. Isto combinado com a suposição de que os erros são normalmente distribuídos, resulta na suposição adicional tradicional em regressão, que é: os efeitos aleatórios são independentes entre si.

Contudo, existem duas situações típicas em que essa suposição de independência deve ser relaxada, sob pena de obtenção de resultados não consistentes (DOBSON, 2002).

A primeira é o caso de dados longitudinais, onde as repostas são medidas repetidamente ao longo do tempo da mesma fonte. Nesse caso, as medidas tomadas a partir do mesmo indivíduo tendem a ser mais parecida entre si do que as medidas tomadas em indivíduos distintos.

A outra situação é quando as repostas de interesse são medidas a partir de indivíduos agrupados em unidades distintas, que é freqüentemente denominada na literatura por estrutura aninhada ou hierárquica de dados.

Um estudo comparativo dessas medidas (nas situações supracitadas) pode levar a resultados enganosos. Concluindo então que a correlação entre os dados tem de ser incorporada à modelagem de alguma maneira, de forma a produzir inferências estatísticas válida, mas isso é contrário a algumas suposições iniciais que sustentam as estimativas dos modelos lineares anteriormente apresentados, particularmente a independência entre efeitos aleatórios do modelo. A possibilidade de se ajustar uma equação para cada grupo seria operacionalmente custosa e fortemente condicionada à quantidade de dados existente a cada grupo. Uma solução bem conhecida em Modelos Lineares é a utilização de matriz bloco-diagonal de covariâncias no processo regular de estimação dos parâmetros, porem sem possibilitar a explicação da variabilidade das medidas intergrupos.

Um modelo que incorpore em si a existência de correlação entre as medidas internas e intergrupos, é o Modelo Linear Hierárquico.

2.2.3 - Modelos Lineares Hierárquicos

Está seção discutirá a lógica e a formulação inerente aos Modelos Lineares Hierárquicos, que se caracterizam por conferir uma estrutura de hierarquia aos modelos lineares e que são também conhecidos como Modelos Multiníveis. As idéias aqui expostas de forma introdutória estão amplamente desenvolvidas em Bryk e Raudenbush (1992), que juntamente com Goldstein (1995), são considerados como referencia na área.

Conforme na seção anterior, é de conhecimento, que existe certa dificuldade de utilizar modelos de regressão considerando uma variável de reposta que termos de erros correlacionados, já que isso, além de derrubar algumas suposições típicas que os sustentam, pode causar imprecisão nas estimativas, dificuldade de operacionalização dos modelos e até a infactibilidade de produção de inferências.

Recentemente, estudos utilizando Modelos Lineares Hierárquicos estão sendo bastante utilizados, pois se desenvolveu bastante nas ultimas décadas os estudos nessa área, principalmente pelas técnicas de estimação de parâmetros em amostras de dados mais aprimoradas, os estudos estatísticos isolados da década de 70 e 80, foram reunidos e aperfeiçoados. Acarretando em diferentes pacotes computacionais feitos especificamente para **MHL**, sendo uns dos mais populares o HLM (Scientific Software International, Inc) e o MLwiN (Centre for Multilevel Modelling).

Naturalmente, com o passar dos anos e aperfeiçoamento das técnicas, os pesquisadores puderam propor hipóteses mais ricas, proporcionando a aferição das variabilidades devidas a cada nível e a provisão de uma modelagem única fornecendo estimativas individuais e também para os grupos, por meio da

ponderação conjunta dos resultados em função da precisão fornecida por cada uma das unidades.

Em relação aos resultados possíveis de ser obtido por esta modelagem, Osborne (2000) comparou as estimativas obtidas para amostras de dados do *National Education Longitudinal Survey of 1988* dos **EUA**, por três estratégias diferentes: desagregada, totalmente agregada e hierárquica. Assumiu-se que o modelo hierárquico fornece a melhor estimativa para o real relacionamento entre as preditoras e a variável resposta e, portanto, constituiu-se na referência de análise. As conclusões apontaram para uma propensão à subestimação significativa dos erros padrões dos estimadores dos efeitos de nível dois por parte da estratégia desagregada. Por outro lado, a estratégia agregada tendeu a superestimação das magnitudes dos coeficientes dos efeitos de nível um e a substimação das magnitudes dos coeficientes dos efeitos de nível dois. Sua conclusão final foi de que estratégias agregadas e desagregadas estimadas por **MQO** não produzem resultados correspondentes ao real relacionamento entre a resposta e as preditoras. Essas conclusões corroboram os estudos de Kreft e de Leeuw (apud Santos et al., 2000, p. 74) que indicam que os modelos hierárquicos fornecem estimativas geralmente mais conservadoras em relação à estratégia desagregada de trabalho.

Uma das características proporcionada pelos **MLH** é a estimação de médias individuais da resposta para cada um dos níveis do modelo. Isso confere ao método um grande atrativo interpretativo, já que com um único modelo obtêm-se estimativas ótimas para cada grupo, e que, além disso, segundo Lindley e Smith (apud Bryk e Raundenbush, 1992, p.40), são estimadores de menor valor de erro quadrático médio esperado.

Em resumo, os **MLH** representam o fenômeno estudado de forma mais fidedigna, já que não pressupõem indevidamente a independência dos termos de erros aleatórios nas equações de regressão, possibilitam o ajuste de um único modelo levando em consideração os efeitos aleatórios intragrupos, e produzem inferências mais apropriadas e resultados mais interessantes, ao darem ocasião à explicação da variabilidade da resposta segundo os diferentes níveis hierárquicos.

Essa modelagem tem sido utilizada nos mais diversos ramos do conhecimento, porém, a prevalência ainda é na área de pesquisas sociais, tradicionalmente educacionais e socioeconômicos, estudos de organizações (instituições), controle epidemiológico entre outras diferentes áreas do conhecimento: geográfica, demográfica, econômica etc.

2.2.3.1 - O Modelo Hierárquico Nulo

Este modelo é a estrutura mais simples possível do **MLH** em dois níveis, não possuindo variáveis preditoras em nenhum dos seus níveis (totalmente não condicional) e, assim o coeficiente β_{ij} no nível *i* equivale à zero para todos *j*. Suas equações são:

Nível 1:

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

onde,

β_{0j}: valor da resposta esperada para o nível j,

r_{ij}: erro aleatório associado ao i-ésimo registro do nível j,

suposições do modelo: $r_{ij} \sim N(0, \sigma^2)$ e r_{ij}'s independentes entre si.

Nível 2:

$$\beta_{0j} = Y_{00} + u_{0j}$$

onde,

$Y_{00:}$ valor da resposta esperada para a toda população,

u_{0j}: efeito aleatório associado ao setor j,

suposições do modelo: $u_{0j} \sim N(0, \tau_{00})$ e u_{0j}'s independentes.

Substituindo a equação do nível 1 na equação do nível 2, obtém o modelo combinado:

$$Yij = Y_{00} + u_{0j} + r_{ij}$$

O modelo nulo pode ser considerado o primeiro passo em modelagens hierárquicas, pois permite a avaliação da variabilidade da resposta em cada um dos níveis. A partir deste modelo, pode-se estruturar a matriz de variâncias/covariâncias para um nível. Então podendo calcular a correlação entre indivíduos do mesmo grupo, o qual denominou de Coeficiente de Correlação Intraclasse (**CCIC**) e mede a proporção da variabilidade da resposta devida exclusivamente ao segundo nível. Sua estimação é importante, na medida em que quanto maior for o **CCIC**, mais se está auferindo ganhos de precisão nas estimativas por meio da utilização do **MLH**.

Na literatura de referencia podem ser consultados diversos submodelos de **MLH**, mas neste trabalho é apresentado apenas o chamado “modelo nulo ou incondicional”, já que na aplicação prática de modelagem hierárquica sua estimação inicial é sempre recomendada para avaliação da variabilidade da resposta devida a cada um dos níveis hierárquicos.

2.2.3.2 - Alguns Aspectos de locação de variáveis

Uma vez estimado um modelo nulo, um pesquisador provavelmente desejara incluir variáveis preditoras em seu modelo. Nesta seção é contemplado um breve esclarecimento sobre locação de variáveis. Entende-se, por locação de variáveis, a questão da escolha da métrica da variável a ser utilizada na modelagem.

Segundo Barreto (2005), um aspecto importante a se reconhecer é que, em modelos com coeficientes aleatórios, como o **MHL**, a alteração da métrica de uma variável preditora produz efeitos distintos em relação ao modelo com coeficientes fixos (regressão tradicional). Neste, o fato de se acrescentar uma constante às medidas de uma variável afeta apenas a magnitude do intercepto, sendo mantidos os demais resultados (coeficientes e estimativas de variância). Já nos modelos com coeficientes aleatórios, os aspectos de locação afetam os procedimentos de inferência e seus resultados, e, na prática, a depender da locação escolhida, são obtidas diferentes respostas.

Existem três hipóteses básicas de eleição para possíveis locações, quais sejam: a métrica natural, o centro na grande media e o centro na media do grupo.

Em **MHL**, a métrica natural de uma variável X deve ser alterada se ela não fizer sentido na prática, pois pode levar a resultados incorretos e com viés. Já em

relação às demais alternativas de locação, o efeito mais imediato se da em relação à interpretação dos interceptos estimados. Essas duas últimas são as locações mais utilizadas em **MHL**. Entretanto, se for conhecida a media populacional de uma variável, pode-se centrá-la em torno dela. Há ainda outras opções, como as que envolvem a locação de variáveis categóricas e seus possíveis efeitos, mas não serão discutidos aqui. Porém, são minuciosamente discutidos e exemplificados em Bryk e Raudenbush (1992).

Natis (2000), diz que não há uma regra fixa para a escolha da locação dos preditores em modelagens hierárquicas, já que isso vai depender de aspectos interpretativos e de outros até, como a presença de multicoliniariedade entre as preditoras, e ainda questões envolvendo estabilidade computacional.

2.2.4 - Modelos lineares generalizados hierárquicos

Na seção anterior viu-se que os modelos lineares hierárquicos prestam-se bem aos casos em que se supõem variáveis respostas normalmente distribuídas em cada um dos níveis. Em modelos lineares, mesmo quando uma variável continua é assimétrica, uma transformação pode prontamente aproximá-la da Normal tornando mais adequada à modelagem. Porém em muitas aplicações de regressão, a variável de resposta é do tipo qualitativa, com dois ou mais resultados possíveis, ou uma variável de contagem, da mesma forma pretende-se estimar essa resposta não só em termos de características individuais, mas também de grupos. Essa generalização em termos de aplicação pode ser alcançada plenamente pelos Modelos Lineares Generalizados Hierárquicos (**MLGH**). Na verdade o **MHL** é um caso particular do **MLGH**, portanto nesta seção terá feito um breve apanhado sobre a generalização hierárquica segundo Raudenbush e Bryk (2002).

Primeiramente, observe-se que os **MLGH** apresentam a mesma estruturação em níveis já conhecida, adequando-se, portanto, à hierarquia que se deseja modelar, O nível 1 de um **MLGH** consiste em três partes distintas: um modelo amostral, uma função de ligação e um modelo estrutural.

À semelhança entrem dos **MHL**, em **MLGH** as funções de reposta podem seguir as diferentes distribuições da família exponencial, a depender do interesse do estudo, mas o modelo estrutural linear de nível 1 mantém-se por meio de uma função de ligação linear nos parâmetros como na equação a seguir:

$$g(\mu_i) = \mathbf{x}_i^T \boldsymbol{\beta} = \eta_i$$

onde se tem que:

$$E(Y_i) = \eta_i$$

Utilizando-se dessa composição genérica, a escolha da função de ligação adequada ao modelo depende da natureza da resposta a ser estudada, se ela é contínua, binária, politômica, de contagem, ordinal etc. Com isso é possível identificar e utilizar a distribuição de probabilidade da família exponencial e a ela aplicável.

É relevante também que em **MLGH**, o **CCIC** não é informativo, como em **MHL**, isto é ocasionado pela própria característica de se modelar uma função de resposta não linear.

2.2.5 - Conclusão

Viu-se, por todo o exposto neste capitulo, que os modelos lineares hierárquicos incorporam aspectos de modelagens estatísticas, especialmente desenvolvidas para aplicações envolvendo estruturas de dados correlacionados (Barreto 2005).

Alem disso, os modelos lineares hierárquicos possuem importantes propriedades. A primeira delas permite que a variabilidade da variável resposta nos diferentes níveis hierárquicos possa ser explicada através de variáveis preditoras incluídas no modelo em diferentes níveis. A outra característica é que esses modelos permitem quantificar quanto à variabilidade da resposta se deve a cada nível, tal que a proporção da variabilidade explicada possa ser comparada diretamente. Ao mesmo tempo, as análises hierárquicas são mais fidedignas ao

não assumirem erroneamente o pressuposto de independência entre as observações das unidades pertencentes a uma estrutura maior, como na analise contextual (KREFT & LEEU, 1998), e nem desperdiçam informação ao tomarem as medias das medidas, como na análise com dados agregados.

2.3 – Inferência em Modelos Lineares Generalizados Hierárquicos

2.3.1 – Introdução

Neste capitulo são abordados alguns métodos de estimação, bastante utilizados em modelos lineares e modelos lineares generalizados, importantes na medida em que são intensamente aplicados em conjunto com métodos adicionais, para a produção de estimativas em **MHL** e **MLGH**. Assim, os aspectos que envolvem as estimativas de mínimos quadrados, detalhados em Neter et al. (1996) e Charnet et al. (1999), são enfocados nas seções 2.3.2 e 2.3.3, e o método de máxima verossimilhança, sob os enfoques de Davidson e Mackinnon (1993), Neter et al. (1996) e Dobson (2002), na seção 2.3.4. A seção 2.3.5 faz menção aos intervalos de confiança. Ao final do capitulo, seção 2.3.6, alguns aspectos de inferência e testes de hipótese em **MHL** e **MLGH** serão sinteticamente apresentados.

2.3.2 - Estimação por mínimos quadrados ordinários

A estimação por **MQO** vem a ser a de utilização mais imediata em estatística, já que é aplicável para os modelos de regressão simples, linear nos parâmetros e na preditora. Além disso, é um modelo com erros normais.

As estimativas de **MQO** possuem certas propriedades, garantidas pela prova do Teorema de Gauss-Markov, a saber, (DAVIDSON e MACKINNON, 1993; NETER et al. 1996):

1 - As estimativas dos parâmetros são não viesadas, por exemplo, para o modelo:

$\mathbf{Y = X\beta + \varepsilon}$ onde se supõe,

$E(\varepsilon) = 0$ e $Var(\varepsilon) = \sigma^2 \mathbf{I}$

Indica que: $E(\boldsymbol{b})=\boldsymbol{\beta}$

2 - Sob as suposições do modelo acima citado os estimadores de **MQO** são os mais eficientes (de menor variância) dentre os estimadores não viesados que sejam uma função linear do vetor de observações Y.

2.3.3 – Mínimos quadrados generalizados

O método de Mínimos Quadrados Generalizados (**MQG**), como já informa sua denominação, é uma generalização do **MQO** aplicável no caso em que os termos de erro são não só heterocedásticos, mas também correlacionados entre si (DAVIDSON e MACKINNON, 1993). Com isso, a propriedade 2 acima passa a não ser mais válida para os estimadores de **MQO**, devendo ser considerada no processo de estimação uma matriz **V** estruturando as variâncias/covariâncias das observações da variável resposta, de forma a que sejam encontrados os estimadores de máxima eficiência.

Maddala (1997, p. 450) prova que “o estimador **MQG** é mais eficiente do que o de **MQO**, na medida em que $Var(\mathbf{b}_{MQO}) - Var(\mathbf{b}_{MQG})$ fornece uma matriz positiva semidefinida”, ou seja, os estimadores de **MQG** possuem menor variância.

Quando a matriz **V** não é conhecida, o que é muito comum, pode-se partir diretamente para as estimativas de máxima verossimilhança, ou então utilizar o **MQG** iterativo. (DOBSON, 2002).

Finalmente, vale ressaltar que os métodos de mínimos quadrados não requisitam que seja especificada a forma funcional da distribuição de probabilidade conjunta dos Y*i*, para obtenção de suas estimativas.

2.3.4 - Estimação por máxima verossimilhança

Este outro método de estimação de parâmetros bastante utilizado quando a forma funcional da distribuição de probabilidade dos termos de erro é especificada, é o de Máxima Verossimilhança (**MV**). A idéia básica do método, como diz seu nome, é encontrar um conjunto de estimadores dos parâmetros, tal que a probabilidade de se ter obtido a amostra de dados em mãos seja a máxima possível. Para isso, a função de probabilidade conjunta da amostra sob o modelo especificado é avaliada para cada uma das observações da variável resposta, sendo tratada como uma função dos parâmetros do modelo. O método busca, portanto, a maior consistência possível com a amostra de dados.

De maneira geral, a verossimilhança é dada pela função de probabilidade conjunta da amostra proporcionada pelas *n* observações, em função do vetor de parâmetros do modelo, $\boldsymbol{\beta}$**,** sendo denotada por $L(\boldsymbol{\beta})$. O método escolhe como estimador de **MV** um vetor **b** que forneça o maior valor possível para a função $L(\boldsymbol{\beta})$ (Barreto, 2005).

Os estimadores obtidos por **MV** podem ser de forma analítica ou numérica. Existe também um processo chamado de máxima verossimilhança restrita (**MVR**),

que corrige o viés na estimativa de **MV** de σ^2. Podem ser encontrados detalhes a respeito dessa estimação em Ogliari (1998).

2.3.5 – Intervalos de confiança

O intervalo de confiança fornece informação sobre a precisão das estimativas. É o intervalo do qual se pode afirmar, com certa confiança, que o verdadeiro valor de um parâmetro populacional está contido nele, ou seja, o intervalo de confiança estabelece limites para o valor objeto de estudo. Os detalhes sobre o estabelecimento de intervalos de confiança em **MHL** e em **MLGH**, cujos procedimentos são os mesmo, podem ser acessados em Raudenbush e Bryk (2002) e Goldstein (2003).

2.3.6 – Testes de hipóteses em MLGH

O teste de hipótese é uma regra usada para decidir se uma hipótese estatística deve ser rejeitada ou não. O objetivo do teste de hipótese é decidir se uma hipótese sobre determinada característica da população é ou não apoiada pela evidência obtida de dados amostrais. Os testes de hipóteses são os primeiros estudos realizados para a verificação da validade do modelo (Gazola, 2002). São apresentados os princípios que norteiam os testes de hipóteses, seguindo, principalmente, o proposto por Bryk e Raundenbush (1992).

2.3.6.1 – Testes relacionados a efeitos fixos

Testar um único parâmetro envolve a seguinte hipótese nula:

H_0: $\gamma_{qs} = 0$.

Isso implica em testar se o efeito de uma determinada preditora de nível 2, W_{sj}, equivale à zero. A estatística de interesse é dada por:

$$Z = \frac{\hat{\gamma}_{qs}}{(\hat{V}\hat{\gamma}_{qs})^{1/2}}$$

onde

$\hat{\gamma}_{qs}$: estimativa para o qs-ésimo efeito fixo,

$\hat{V}\hat{\gamma}_{qs}$: variância estimada para $\hat{\gamma}_{qs}$.

Esta estatística que exprime a razão entre o coeficiente estimado e seu erro padrão, segue assintoticamente, uma distribuição Normal padronizada. Contudo, considerando a formula acima, seguindo uma distribuição *t* com numero de graus de liberdade dado por (J – S_q – 1) permite inferências mais acuradas (RAUDENBUSH e BRYK, 2002). Um teste envolvendo diversos efeitos fixos pode também ser executado. Seja um vetor de efeitos fixos γ^T = (γ_{00}, γ_{01}, γ_{10}, γ_{11}). O interesse pode ser testar se γ_{01} e γ_{11} são iguais à zero.

2.3.6.2 – Testes relacionados a coeficientes aleatórios de nível 1

Esse teste segue basicamente a mesma filosofia em relação aos efeitos fixos. Seja a hipótese nula:

H_0: $\beta_{qj} = 0$.

Aqui se têm a opção de utilizar no teste as estimativas de **MQO** ou então as empíricas de Bayes. Empregando-se aqui a empírica de Bayes tem-se:

$$z = \beta^*_{qj} / V^{*1/2}_{qqj}$$

onde,

β^*_{qj} : q-ésima estimativa empírica de Bayes do grupo j,

V^*_{qqj}: q-ésimo elemento diagonal da matriz de dispersão posterior dos coeficientes β^*_{qj} .

Esta estatística z segue assintoticamente, uma distribuição Normal padronizada.

É preciso ressaltar que existem ainda outras possíveis alternativas para a realização desse teste disponível ao pesquisador, não apresentados aqui, porém pode ser consultado em Raudenbush e Bryk (2002).

2.3.6.3 – Testes relacionados a componentes de variância/covariância

Uma das preocupações do pesquisador em modelos hierárquicos é avaliar se os coeficientes aleatórios de nível 1 efetivamente possuem efeito aleatório ou devem ser especificados como fixos em relação aos grupos. Isso pode ser aferido por um teste de variância/covariância. Quando ele envolve um único parâmetro, a hipótese nula é:

H_0: $\tau_{qj} = 0$.

No caso de todos – ou quase todos – os grupos propiciarem uma estimativa de **MQO** para β_{qj}, uma estatística possível para o teste, já que a estimativa da variância do estimador de:

β_{qj}, sob a hipótese nula, é equivalente a $\mathbf{V}_j = \sigma^2 (\mathbf{X}^T_j \mathbf{X}_j)^{-1}$, é dada por:

$$h = \sum_j \left(\hat{\beta}_{qj} - \hat{\gamma}_{q0} - \sum_{s=1}^{S_q} \hat{\gamma}_{qs} W_{sj} \right)^2 / \hat{V}_{qqj}$$

onde,

$\hat{V}_{qqj}$: q-ésimo elemento diagonal de $\hat{V}_j$.

Essa estatística segue aproximadamente uma distribuição χ^2 com $(J - S_q - 1)$ graus de liberdade. Os testes multiparâmetros para componentes de variância/covariância são fundamentados no teste de razão de verossimilhança e a filosofia do teste é bem definida por Natis:

Temos na hipótese nula todos os componentes que se deseja testar considerados nulos. Dessa forma, a hipótese nula corresponde a uma forma

reduzida da hipótese alternativa. Vale ressaltar que tanto na hipótese nula quanto na hipótese alternativa, os modelos devem ser idênticos com relação aos efeitos fixos (NATIS, 2000).

A hipótese nula, nesse caso é dada por H_0: $\mathbf{T} = \mathbf{T}_0$, a ser testada contra a alternativa H_1: $\mathbf{T} = \mathbf{T}_1 \mathbf{.} \mathbf{T}_0$ representa uma matriz de variância/covariância associada a um modelo reduzido em relação ao modelo mais geral correspondente à matriz $\mathbf{T}_1$.

$$H_{RV} = D_0 - D_1$$

onde D_0 e D_1 são as *deviances* proporcionadas pelo ajuste, respectivamente, dos modelos reduzidos e gerais. A *deviance* é calculada por meio de $D = -2log L(\theta)$, entendendo θ como vetor de parâmetros do modelo $L(\theta)$ sendo avaliada em seu Maximo. Sabe-se que quanto maior a *deviance* pior o ajuste obtido para o modelo. Essa estatística segue uma distribuição χ^2 com (*m*) graus de liberdade, onde *m* é a diferença entre o numero de parâmetros previstos para os dois modelos. Valores elevados para essa estatística indicam que a hipótese nula é muito simples para explicar os dados observados e a redução na *deviance* ocasionada pelo modelo mais completo se justifica (BARRETO, 2005).

2.3.7 – Conclusão

Neste capitulo foram apresentados os processos que fundamentam as estimativas de parâmetros em **MHL** e **MLGH** aplicados a estrutura de dados correlacionados. Viu-se que os processos utilizam de uma complexa combinação de métodos já tradicionais em estatística com métodos adicionais relativamente

recentes. O resultado é que as estimativas de efeitos fixos, coeficientes aleatórios de nível 1 e componentes de variância/covariância provêem ao pesquisador uma maior riqueza interpretativa para os fenômenos em estudo, isso em uma modelagem única e robusta.

Este modelo estatístico é aplicado no capitulo 5, a partir da análise exploratória de dados executada no próximo capitulo.

3 – ANÁLISE EXPLORATÓRIA DOS DADOS

O presente estudo utiliza-se de uma amostra da base de dados constituída de 8244 registros que representam lançamentos em dívida ativa, de cobrança do imposto territorial e predial urbano do município de Itajaí. Estes lançamentos são referentes ao exercício de 2006. Na base de dados de origem existiam 16513 registros. Utilizando então cerca de 50 por cento dos dados para amostra, considerado uma quantidade bem expressiva de dados, para este tipo de estudo. É importante deixar claro, que a amostra utilizada foi estratificada. Os bairros foram considerados os estratos, mantendo a representatividade de cada bairro na amostra também.

Existem variáveis do tipo quantitativas e qualitativas, representando características do terreno e/ou imóvel, assim como características do débito e do bairro que este lote/imóvel está localizado. O software utilizado para a análise estatística foi o *STATISTICA Version 6.0*, enquanto que para trabalhar com as transformações de variáveis e criações de variáveis *dummy,* utilizou o software *SQL Manager 2007 for PostgresSQL* e *Microsoft Excel*.

3.1 – Identificação e apresentação das variáveis

A variável dependente é a situação do pagamento, que representa as seguintes situações em que um débito pode se encontrar: Não pago, quando o débito meramente não está quitado no momento; Pago, quando este débito foi quitado. Para este trabalho, foram separadas as informações referentes aos lotes, em uma base de dados, das informações dos imóveis. Pois muitas informações

dos imóveis não existem para os lotes, e principalmente, porque lotes e imóveis possuem comportamento diferentes. Por esses motivos estudaremos os mesmos em separado e, portanto, será implementado um modelo, o **MLGH1**, para os imóveis e outro, **MLGH2**, para os lotes.

As variáveis independentes estão relacionadas e descritas nos quadros 3.1 e 3.2.

Quadro 3.1: Descrição das variáveis independentes quantitativas

Variáveis	Unidade de medida	Descrição	Tabela Lote	Tabela Imóvel
valor_lancamento	Reais	Valor do débito.	Existe	Existe
id_bairro	-	Identificador de qual bairro o terreno/imóvel está localizado.	Existe	Existe
area_total_lote	m²	Área total do terreno.	Existe	Existe
fracao_ideal	m²	Percentual que um imóvel ocupa do terreno em que está situado.	Não Existe	Existe
area_tributavel	m²	Área do imóvel construída.	Não Existe	Existe

Quadro 3.2: Descrição das variáveis independentes qualitativas

Variáveis	Categorias	Descrição	Tabela Lote	Tabela Imóvel
passeio	cimento; terra; outros	Material da calçada do lote ou imóvel.	Existe	Existe
topografia	abaixo do nível; acima do nível; no nível; alagado.	Característica do relevo do terreno.	Existe	Existe
situacao_lote	beco; encravado; esquina; meio de quadra.	Localização do terreno na quadra.	Existe	Existe
benfeitoria	normal; sem muro; sem passeio;	Obras executadas para conservação ou melhoria do	Existe	Existe

	sem muro e sem passeio.	mesmo		
agua_luz_drenagem	agua; agua e luz; agua e luz e drenagem; luz.	Existência ou não de água, luz e drenagem.	Existe	Existe
pavimento	asfalto; lajota; macadame; paralelepípedo; terra.	Pavimento rodoviário; Rua.	Existe	Existe
coleta_lixo	nao tem; 3 vezes por semana; 5 vezes por semana.	Existência ou não de coleta de lixo.	Existe	Existe
especie_imovel	Alvenaria; madeira; mista.	Espécie do material do imóvel.	Não Existe	Existe
tipo_uso_imovel	comercio; industria; residência.	Utilização do imóvel.	Não Existe	Existe
regime_utilizacao_imovel	alugado; fechado; próprio.	Regime de utilização do imóvel.	Não Existe	Existe
tipo_imovel	apartamento; casa; barraco; cinema; comercio; deposito aberto; deposito fechado; escola; galpão aberto; galpão fechado; hospital; hotel c/restaurante; hotel s/ restaurante; industria; oficina; outros; posto de serviço; restaurante ou bar; serviço público.	Tipo do imóvel.	Não Existe	Existe
conservacao_imovel	bom; mau; regular.	Conservação do imóvel.	Não Existe	Existe

acabamento_imovel	bom; comum; luxo; normal; popular.	Acabamento do imóvel.	Não Existe	Existe
situacao_pagamento	0 = Não pagou; 1 = Pago.	Situação do pagamento da dívida de IPTU.	Existe	Existe

3.2 – Pré-processamento

A base de dados origem, utilizada nas análises foi adquirida na Prefeitura Municipal de Itajaí. O **SGBD** utilizado é o *postgres*, onde foi utilizados **SQL's** para retirar as informações das variáveis supracitadas. A base de dados origem possui cerca de 300 tabelas, na terceira forma normal.

Conforme já descrito na introdução deste capitulo, foram pré-selecionados cerca de 16500 registros, estes registros, são a população total do nosso estudo, porém para viabilizar a pesquisa, utilizará uma amostra estratificada de aproximadamente 8244 registros, dentre os quais 2517 são referentes aos lotes e 5727 referentes aos imóveis.

Na identificação dos registros com valores *missing* foi encontrada apenas a presença de duas tuplas com essa característica. Esses registros foram removidos da base de dados da pesquisa.

3.3 – Análise exploratória das variáveis

A existência de variáveis qualitativas e quantitativas fez-se necessária, primeiramente a transformação de algumas variáveis, para a realização do estudo do relacionamento entre elas.

Algumas variáveis qualitativas foram re-categorizadas, pois existia um grande numero de categorias ou algumas categorias que não possuíam números relevantes de registros. Assim, essas variáveis foram categorizadas novamente conforme o Quadro 3.3.

Quadro 3.3: Nova categorização das variáveis qualitativas.

VARIAVEL	CATEGORIAS	NOVAS CATEGORIAS
pavimento	cimento; terra; outros.	cimento; outros.
topografia	abaixo do nível; no nível; acima do nível; alagado.	abaixo do nível; no nível; acima do nível.
situacao_lote	beco; encravado; meio de quadra; esquina	beco ou encravado; meio de quadra; esquina
benfeitoria	normal; sem muro; sem passeio; sem muro e sem passeio.	normal; sem muro ou sem passeio.
tipo_imovel	apartamento; casa; barraco; cinema; comercio; deposito aberto; deposito fechado; escola; galpão aberto; galpão fechado; hospital; hotel c/restaurante; hotel s/ restaurante; industria; oficina; outros; posto de serviço; restaurante ou bar; serviço público.	apto; casa; depositos; comercio; serv. publicos e outros.
regime_utilizacao_imovel	proprio; fechado; alugado.	proprio; alugado.
tipo_uso_imovel	comercio; industria; residencia.	comercio e ind; residencia.
pavimento	asfalto; lajota; macadame; paralelepípedo; terra;	asfalto; lajota; paralelepípedo; terra.
especie_imovel	alvenaria; madeira; mista.	alvenaria; madeira ou mista.
coleta_lixo	nao tem; 3 vezes por semana; 5 vezes por semana.	com coleta; sem coleta.
conservacao_imovel	bom; mal; regular.	bom; ruim; regular.

Para estudo da assimetria das variáveis quantitativas foram construídos diagramas de caixa. Os resultados são mostrados na Figura 3.1 para a base de dados dos lotes e na Figura 3.2 para os imóveis. Observa-se que a variáveis independentes são assimétricas positiva.

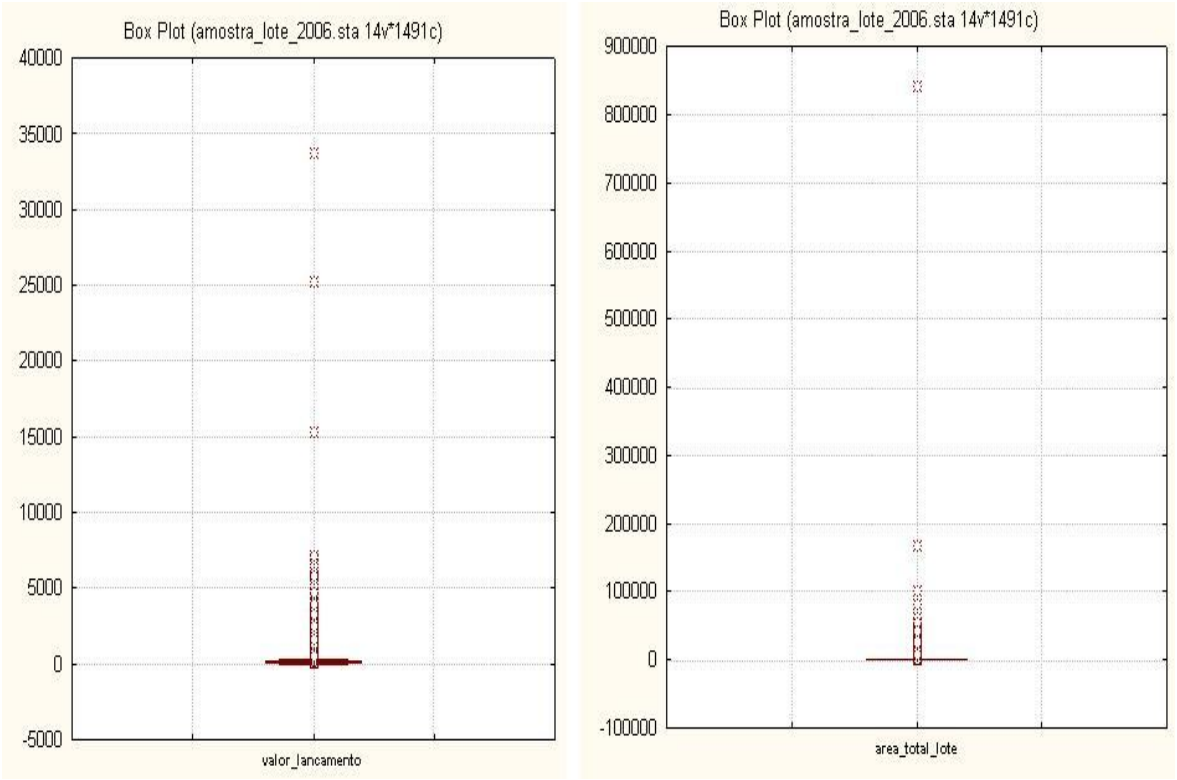


Figura 3.1: Diagramas de caixa das variáveis quantitativas referente aos lotes.

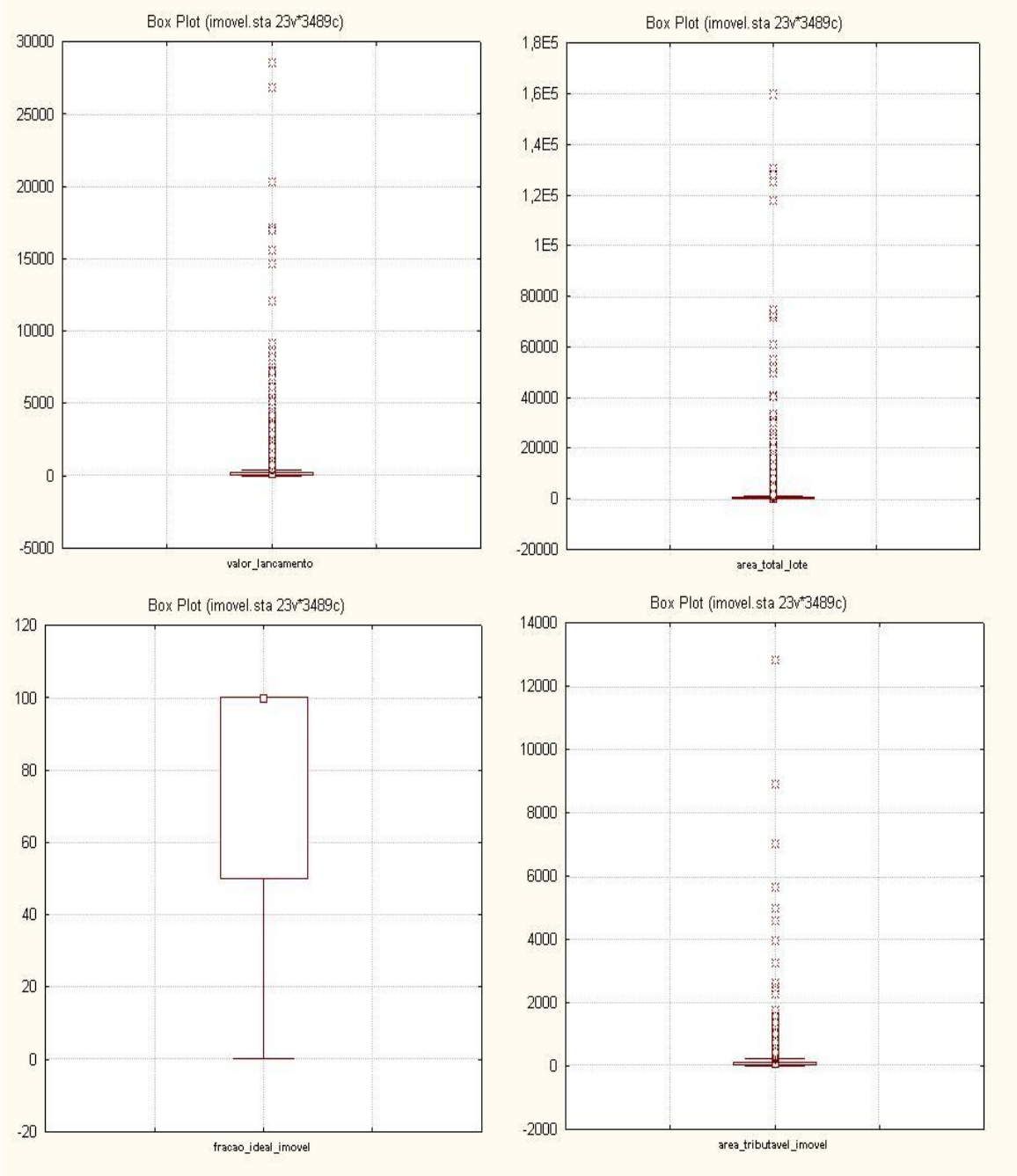


Figura 3.2: Diagramas de caixa das variáveis quantitativas referente aos imóveis.

Com a transformação logarítmica, nas variáveis independentes apresentadas acima, foi amenizado o problema da simetria. Exceto na variável fracao_ideal, pois esta variável tem valores da faixa de zero a cem, sendo assim impossibilitado de executar a transformação logarítmica. Esta melhora, mais o fato de que os valores

das variáveis diferem em relação uma das outras, levam a concluir que se deve trabalhar com estas variáveis transformadas logaritmicamente. A Figura 3.3 e Figura 3.4 mostram os diagramas de caixas e os histogramas dessas variáveis transformadas.

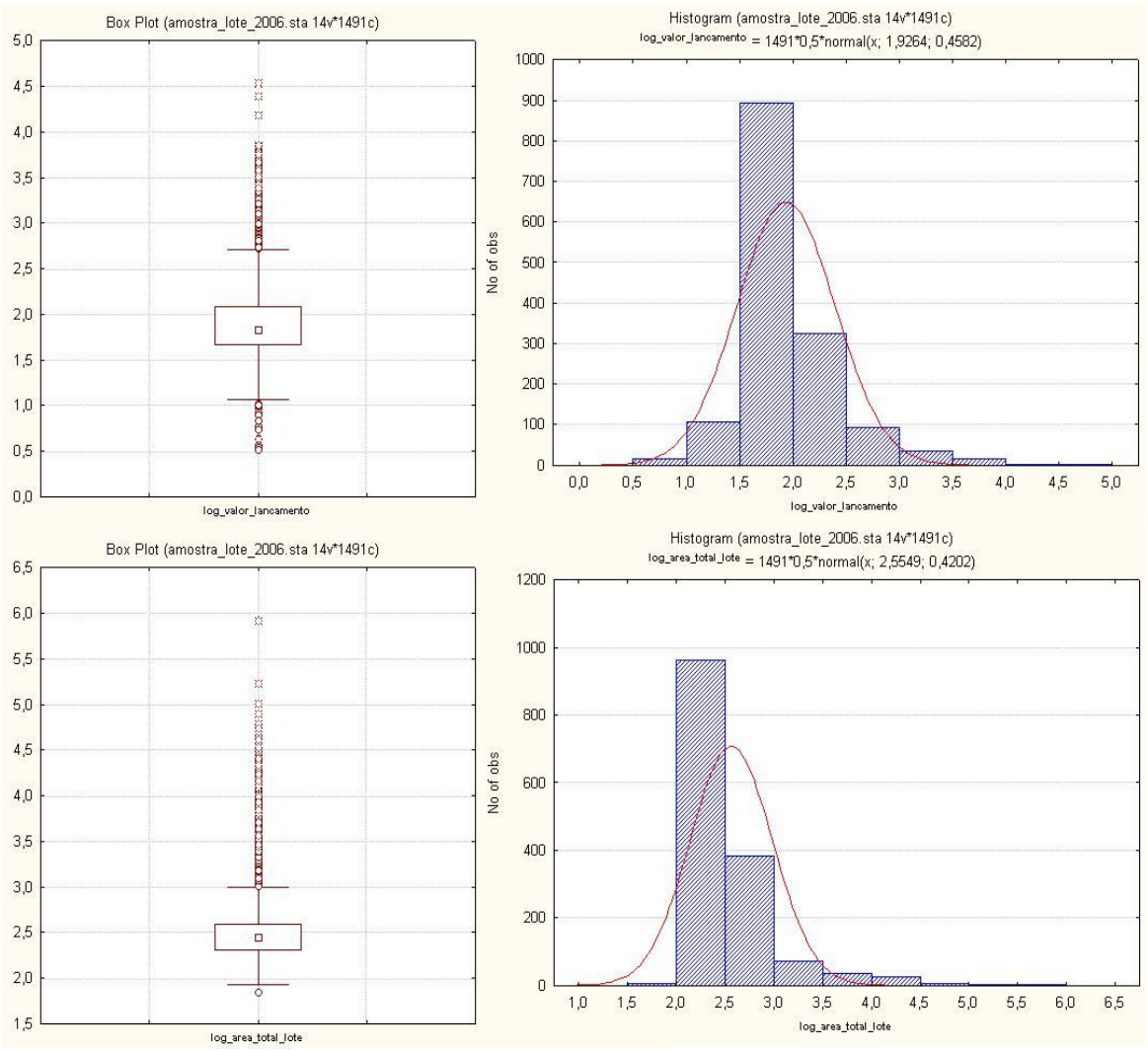


Figura 3.3: Diagramas de caixa e histogramas do logaritmo das variáveis quantitativas do lote.

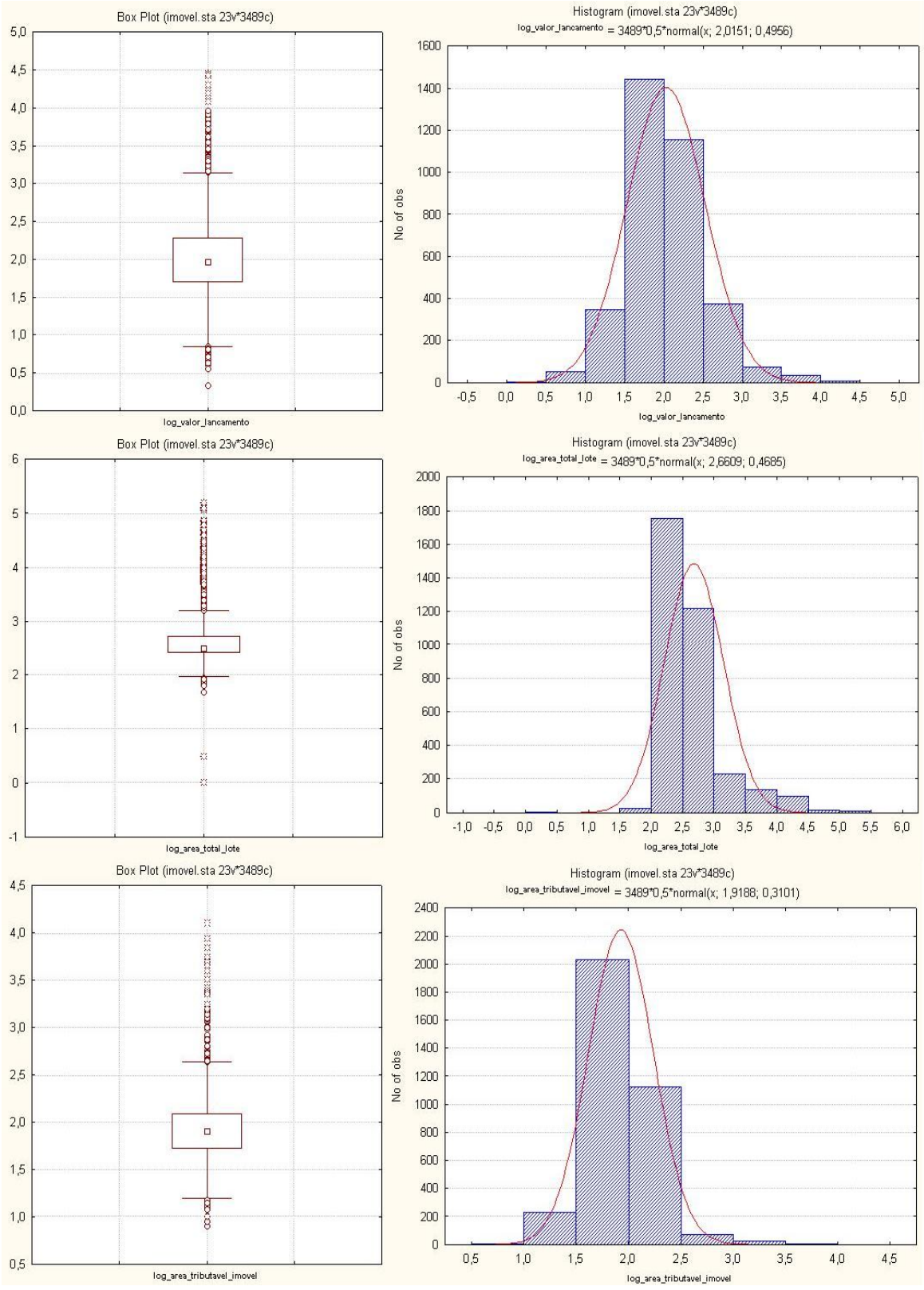


Figura 3.4: Diagramas de caixa e histogramas do logaritmo das variáveis quantitativas do imóvel.

Desta forma, passou a se ter dois conjuntos de dados: o primeiro com as variáveis quantitativas e as qualitativas categorizadas, ambas na forma original; e o segundo com as variáveis quantitativas transformadas, exceto a variável fracao_ideal, junto com as qualitativas que foram re-categorizadas.

3.4 – Relação da variável dependente com as variáveis independentes.

A Tabela 3.1 apresenta algumas características das variáveis independentes quantitativas com relação à variável dependente situação do pagamento.

As análises de variância, como esperado, diferem quando os registros representam lote, e quando representam os imóveis. O resultado é marginalmente significante para os lotes e altamente significante para os imóveis. Para os lotes nota-se que apenas a variável log_valor_lancamento é significante. Não se nota significância da variável log_area_total_lote. Para a análise dos imóveis, log_valor_lancamento, log_area_tributavel_imovel e fração_ideal são altamente significantes e log_area_total_lote é significante.

Tabela 3.1: Variável dependente (situacao_pagamento) e as variáveis independentes quantitativas.

Variável	Lote / Imóvel	Valor p	Valor F
log_valor_lancamento	Lote	0,2656	1,3265
log_area_total_lote	Lote	0,2275	1,4814
log_valor_lancamento	Imóvel	0,000000	50,8972
log_area_total_lote	Imóvel	0,000529	1,49450
log_area_tributavel_imovel	Imóvel	0,000000	10,3511
fracao_ideal	Imóvel	0,001334	5,22858

Para investigar possíveis relações entre a variável dependente e as variáveis qualitativas, fez-se o uso do teste F. A Tabela 3.2, abaixo, mostra os resultados. Como o numero de registro da amostra utilizado é grande, se o nível de significância for acima de 0,05 podemos excluir essa variável.

Tabela 3.2: Teste F da **ANOVA** para a variável dependente (situacao_pagamento) versus variáveis independentes qualitativas.

Variável	Lote / Imóvel	Pearson(chi-Square)	Pearson(p)
passeio	Lote	2,401762	0,30093
topografia	Lote	33,73501	0,00000
situação_lote	Lote	15,89767	0,00316
benfeitoria	Lote	0,329663	0,84804
agua_luz_drenagem	Lote	30,32119	0,00003
pavimento	Lote	25,28496	0,00030
coleta_lixo	Lote	29,83866	0,00001
passeio	Imóvel	0,00309	0,99846
topografia	Imóvel	3,805909	0,43291
situação_lote	Imóvel	15,23781	0,00423
Benfeitoria	Imóvel	3,848045	0,14602
agua_luz_drenagem	Imóvel	6,229803	0,39795
pavimento	Imóvel	59,02214	0,00000
coleta_lixo	Imóvel	4,874236	0,30045
especie_imovel	Imóvel	21,49481	0,00002
tipo_uso_imovel	Imóvel	9,411082	0,00905
regime_utilizacao_imovel	Imóvel	2,703152	0,25884
tipo_imovel	Imóvel	27,35658	0,00061
conservacao_imovel	Imóvel	37,67626	0,00000
acabamento_imovel	Imóvel	71,06654	0,00000

Com base nos resultados obtidos nessa seção, pode-se decidir pela exclusão das variáveis quantitativa referentes aos lotes por não ser importante na explicação da situação do pagamento. Já para as variáveis qualitativas, da base de dados dos lotes, foram excluídas passeio e benfeitoria. E da base de dados dos imóveis, foram excluídas as seguintes variáveis: passeio, topografia, benfeitoria, água_luz_drenagem, coleta_lixo e regime_utilizacao_imovel.

Assim passamos a trabalhar com um conjunto menor de variáveis independentes. As variáveis qualitativas para serem utilizadas no modelo devem ser transformadas em variáveis do tipo *Dummy.* Com estas novas categorias, estas variáveis puderam coerentemente, serem transformadas, conforme o Quadro 3.4 para os lotes e Quadro 3.5 para os imóveis.

Quadro 3.4: Variáveis Dummy para os registros referentes aos lotes.

VARIAVEL	NOVAS VARIAVEIS	SIGNIFICADO
situacao_lote	sl_beco_enc sl_meio sl_esquina	Lote situado num beco ou encravado. Lote situado no meio de quadra. Lote situado na esquina.
topografia	topo_abaixo; topo_acima; topo_nivel.	Topografia abaixo do nível. Topografia acima do nível. Topografia no nível.
agua_luz_drenagem	ald_agua ald_luz ald_agua_luz ald_agua_luz_drena	Possui somente água. Possui somente luz. Possui água e luz. Possui água, luz e drenagem.
pavimento	pav_asfalto pav_lajota pav_paralelepipedo pav_terra	Rua asfaltada. Rua pavimentada com lajotas. Rua calcada com paralelepípedo. Rua sem pavimentação.
Coleta_lixo	coleta_nao coleta_3 coleta_5	Não tem coleta de lixo. 3 vezes por semana. 5 vezes por semana.

Quadro 3.5: Variáveis Dummy para os registros referentes aos imóveis.

VARIAVEL	NOVAS VARIAVEIS	SIGNIFICADO
situacao_lote	sl_beco_enc sl_meio sl_esquina	Imóvel está em um beco ou encravado. Imóvel situado no meio de quadra. Imóvel situado na esquina.
pavimento	pav_asfalto pav_lajota pav_paralelepipedo pav_terra	Rua asfaltada. Rua pavimentada com lajotas. Rua calcada com paralelepípedo. Rua sem pavimentação.
especie_imovel	esp_alvenaria esp_mista_madeira	Imóvel de alvenaria. Imóvel de madeira ou misto.
tipo_uso_imovel	tp_comercio tp_residencia	Uso para comercio ou industria. Uso residencial.

tipo_imovel	ti_apto ti_casa ti_comercio ti_deposito ti_servico	Apartamento Casa ou barracos. Comercio em geral. Depósito ou galpão. Serviços públicos ou outros.
conservacao_imovel	cons_bom cons_regular cons_mau	Imóvel bem conservado. Imóvel com conservação regular. Imóvel mal conservado.
acabamento_imovel	acab_bom acab_comum acab_normal acab_popular acab_luxo	Acabamento bom. Acabamento comum. Acabamento normal. Acabamento popular. Acabamento luxuoso.

Em face dos resultados alcançados nesta seção, obtivemos um conjunto menor de variáveis para aplicar no modelo. Para a base de dados referente aos lotes dezessete variáveis *dummy* e uma quantitativa. Para a base dos imóveis, utilizaremos quatro variáveis quantitativas e vinte e quatro variáveis *dummy*.

4- ESTIMAÇÃO E APLICAÇÃO DO MODELO

Segundo a literatura de referência em **MLGH** apresentado, o primeiro passo de uma abordagem hierárquica é o ajuste de um modelo nulo, de forma de se avaliar e quantificar a existência de variabilidade do nível em relação à resposta esperada. Para isso, é preciso que se considere a estrutura hierárquica da unidade de interesse.

4.1 – A estrutura hierárquica em unidades

Já foi mencionado, que a hierarquia do **MLGH** irá ser estruturada de acordo com os bairros em que os lotes ou imóvel estão localizados. O nível bairro foi selecionado pelo motivo de agregar um conjunto de imóveis ou bairros, relativamente parecidos. Essa divisão de bairro na cidade de Itajaí, não é meramente geográfica, existem fatores culturais, históricos e políticos. Teoricamente, os indivíduos intra-bairros agrupam características comuns e características distintas dos indivíduos dos outros bairros.

4.2 – O MLGH1 nulo

O primeiro modelo nulo tem por objetivo avaliar a existência de variabilidade dos imóveis, em relação às suas possíveis situações de pagamento de **IPTU** em dívida ativa: “não pagou”, “pago em dia”.

Para o ajuste do modelo nulo, foi utilizado o seguinte procedimento, no aplicativo *The SAS System for Windows V8*.

```
PROC NLMIXED data=mhl.Imovel technique=newrap;
PARMS b0=0 sd=1.5 thres1=1;
bounds sd > 0;
  z = b0 + u;
  p1 = 1 / (1 + exp(-(0+z)));
  p2 = (1/(1 + exp(-(thres1+z)))) - (1/(1 + exp(-(0+z))));
  p3 = 1 - (1 / (1 + exp(-(thres1+z))));
  ll = sit1*log(p1) + sit2*log(p2) + sit3*log(p3);
model sit1 ~ general(ll);
RANDOM u~NORMAL(0,sd*sd) SUBJECT=id_bairro;
RUN;
```

Este modelo nulo fornece as estimativas da Tabela 4.1:

Tabela 4.1: Estatísticas do modelo nulo 1.

Parâmetro	Estimativa	Erro Padrão	Graus de liberdade	Valor t	Nível descritivo
B0	0.3436	0.06904	26	4.93	<. 0001
Sd	0.3079	0.06039	26	5.10	<. 0001
tres1	0.3021	0.01447	26	20.88	<. 0001

Os níveis descritivos da Tabela 4.1, com base em testes de hipótese, demonstram que a hipótese nula é altamente implausível, indicando que existe variabilidade significativa, e, portanto poderemos continuar utilizado o **MLGH** para alcançarmos nosso objetivo de estudo.

4.2 – O MLGH2 nulo

O segundo modelo nulo, o qual tem por objetivo avaliar a existência de variabilidade dos lotes, em relação às suas possíveis situações de pagamento de **IPTU** em dívida ativa: “não pagou”, “pago em dia”.

Para o ajuste deste modelo nulo, foi utilizado o seguinte procedimento, no aplicativo *SAS*, como resultado as seguintes estatísticas demonstradas na Tabela 4.2.

```
PROC NLMIXED data=mhl.Lote technique=newrap;
PARMS b0=0 sd=1.5 thres1=1;
bounds sd > 0;
  z = b0 + u;
  p1 = 1 / (1 + exp(-(0+z)));
  p2 = (1/(1 + exp(-(thres1+z)))) - (1/(1 + exp(-(0+z))));
  p3 = 1 - (1 / (1 + exp(-(thres1+z))));
  ll = sit1*log(p1) + sit2*log(p2) + sit3*log(p3);
model sit1 ~ general(ll);
RANDOM u~NORMAL(0,sd*sd) SUBJECT=id_bairro;
RUN;
```

Tabela 4.2: Estatísticas do modelo nulo 2.

Parâmetro	Estimativa	Erro Padrão	Graus de liberdade	Valor t	Nível descritivo
B0	0.7578	0.1229	26	6.16	<. 0001
Sd	0.5332	0.1007	26	5.30	<. 0001
tres1	0.2063	0.01955	26	10.55	<. 0001

Este modelo nulo, também demonstra que a hipótese nula é altamente implausível, indicando que existe variabilidade significativa, e, portanto poderemos continuar utilizado o **MLGH** para alcançarmos nosso objetivo de estudo.

4.2 – Variáveis candidatas aos modelos

Conforme os resultados obtidos no capítulo três, resultou em dois conjuntos de dados. O primeiro referentes aos imóveis, utilizado no **MLGH1** e o outro conjunto referente aos lotes utilizado no **MLGH2**. Na Tabela 4.3 e Tabela 4.4, estão listadas todas as variáveis, após as transformações.

Tabela 4.3: Variáveis pré-selecionadas para os modelo 1.

Variável	Tipo transformação
log_valor_lancamento	Logarítmica
log_area_tributavel	Logarítmica
pav_asfalto	Dummy
pav_lajota	Dummy
pav_terra	Dummy
acab_popular	Dummy
acab_comum	Dummy
acab_normal	Dummy
acab_bom	Dummy
cons_bom	Dummy
cons_regular	Dummy
esp_alvenaria	Dummy
tp_comercio	Dummy
sl_meio	Dummy
sl_esquina	Dummy
log_area_total_lote	Logarítmica
ti_apto	Dummy
ti_casa	Dummy
ti_deposito	Dummy
ti_comercio	Dummy
fracao_ideal_imovel	Sem transformação

Tabela 4.4: Variáveis pré-selecionadas para os modelo 2.

Variável	Tipo transformação
log_valor_lancamento	Logarítmica
topo_acima	Dummy
topo_nivel	Dummy
coleta_nao	Dummy
coleta_3	Dummy
ald_agua	Dummy
ald_luz	Dummy
ald_agua_luz	Dummy
pav_asfalto	Dummy
pav_lajota	Dummy
pav_terra	Dummy
sl_meio	Dummy
sl_esquina	Dummy

4.3 – Processo de seleção de variáveis dos níveis 1 e 2

Para as variáveis de nível 1, o procedimento adotado foi a inclusão uma a uma das variáveis no modelo (método stepwise forward), em diversas rodadas.

As variáveis foram avaliadas tomando-se por base o quanto a sua inclusão ao modelo é significante para a variável resposta. Nesse processo, se incluí uma a uma variável e analisa-se o modelo. Caso ele continue significante mantém a mesma no modelo, senão descarta-se a variável e faz-se o mesmo com a próxima variável. As rodadas se sucedem até que se convirja para um modelo considerado final em que não há mais variáveis candidatas, mas sim incluídas ou descartadas do modelo.

É importante registrar que nem sempre as variáveis que eventualmente mais contribuam para a explicação da resposta em termos isolados formam bons modelos em conjunto, provavelmente em virtude de correlações entre as variáveis e também da questão da estabilidade de estimativas. Na Tabela 4.5 segue os resultados das variáveis do nível imóvel.

Tabela 4.5: Variáveis do nível imóvel candidatas do MHL1.

Variável	**Variáveis *Dummy***	**Incluída/Descartada**
log_valor_lancamento	Quantitativa	Incluída
log_area_tributavel	Quantitativa	Descartada
pavimento	pav_asfalto; pav_lajota; pav_terra	Incluída
acabamento_imovel	acab_popular; acab_comum; acab_normal; acab_bom	Incluída
conservacao_imovel	cons_bom; cons_regular	Incluída
especie_imovel	esp_alvenaria	Incluída
tipo_imovel	tp_comercio	Incluída
situacao_imovel	sl_meio; sl_esquina	Descartada
log_area_total_imovel	Quantitativa	Descartada
tipo_uso_imovel	ti_apto; ti_casa;ti_deposito; ti_comercio	Descartada
fracao_ideal_imovel	Quantitativa	Descartada

Para as variáveis de nível do bairro, o procedimento adotado foi à inclusão uma a uma das variáveis no modelo (método forward), em diversas rodadas.

As variáveis foram avaliadas tomando-se por base o quanto a sua inclusão ao modelo é significante para a variável resposta. Esse processo busca avaliar as variáveis candidatas, considerando-as isoladamente no modelo. Ao final da primeira rodada forma-se um modelo provisório inicial incluindo a de melhor desempenho. Caso ela seja significante, ela é mantida no modelo, e inclui-se a próxima variável de melhor desempenho, formando um segundo modelo provisório. As rodadas se sucedem como no caso do nível dos imóveis. O parâmetro de significância utilizado foi o log verossimilhança.

Após a primeira rodada completa chegou-se ao seguinte resultado mostrado na Tabela 4.6 que segue abaixo.

Tabela 4.6: Variáveis do nível bairro candidatas do MHL1.

Variável	Nível de significância	Incluída/Descartada
carencia_pavimentacao	7165.2	Descartada
qtd_apartamento	7160.4	Incluída
qtd_hospital	7162.1	Descartada
tem_clube	7161.7	Incluída
qtd_serv_publico	7163.1	Descartada
qtd_comercio	7163.6	Descartada
qtd_barraco	7160.7	Incluída
qtd_industria	7166.4	Descartada
qtd_escola	7166.1	Descartada
qtd_casa	7166.1	Descartada

Concluindo as próximas rodadas chegamos ao resultado de que a variável tem_clube é insignificante para o modelo optando em descartá-la junto com carência_pavimentacao, qtd_hospital, qtd_serv_publico, qtd_comercio, qtd_industria, qtd_escola e qtd_casa. Optando então em incluir as variáveis qt_apartamento e qtd_barracos ao modelo.

Já para o **MLGH2** encontramos como resultados para o nível 1, características do lote as variáveis da Tabela 4.7 e na Tabela 4.8 as variáveis referentes aos bairros.

Tabela 4.7: Variáveis do nível lote candidatas do MHL2.

Variável	Variáveis *Dummy*	Incluída/Descartada
log_valor_lancamento	Quantitativa	Incluída
topografia	topo_acima; topo_nivel	Incluída
pavimento	pav_asfalto; pav_lajota; pav_terra	Incluída
coleta_lixo	coleta_nao; coleta_3	Descartada
agua_luz_drenagem	ald_luz; ald_agua; ald_agua_luz	Descartada
situacao_lote	sl_meio; sl_esquina	Incluída

Tabela 4.8: Variáveis do nível bairro candidatas do MHL2.

Variável	Nível de significância	Incluída/Descartada
carencia_pavimentacao	2902.2	Descartada
qtd_apartamento	2902.6	Descartada
qtd_hospital	2900.9	Descartada
tem_clube	2896.7	Incluída
qtd_serv_publico	2901.4	Descartada
qtd_comercio	2902.5	Descartada
qtd_barraco	2896.7	Incluída
qtd_industria	2899.9	Descartada
qtd_escola	2902.7	Descartada
qtd_casa	2902.0	Descartada

Observando os resultados, para o **MLGH2**, foram incluídas as variáveis log_valor_lancamento, topografia, pavimento e situacao_lote das características dos lotes. E das variáveis que representam as características dos bairros foram incluída apenas tem_clube e qtd_barraco. Para as variáveis do nível do bairro, precisou apenas duas rodadas, pois logo na primeira rodada foram descartadas todas as variáveis, exceto a tem_clube e qtd_barraco, únicas que mostraram significantes no primeiro momento. E que na rodada seguinte foi confirmada essa significância sendo ambas incluídas no modelo.

4.3 – O MLGH1

O modelo final dos imóveis é formalizado, no *SAS*, da seguinte forma:

```
PROC NLMIXED data=binomial.Imovel technique=newrap;
PARMS b0=0 b1=0 b2=0 b3=0 b4=0 b5=0 b6=0 b7=0 b8=0 b9=0 b10=0
b11=0 b12=0 b13=0 b14=0 b15=0 b16=0 b17=0 b18=0 b19=0 b20=0
sd=1.5;
bounds sd>0;
      z = b0 +
      b1 * log_valor_lancamento +
      b2 * pav_asfalto + b3 * pav_lajota + b4 * pav_terra +
      b5 * esp_mista + b6 * esp_alvenaria +
      b7 * tp_casa + b8 * tp_apto + b9 * tp_comercio +
      b10 * tp_deposito +
      b11 * sl_esquina + b12 * sl_meio +
      b13 * acab_normal + b14 * acab_popular +
      b15 * acab_bom + b16 * acab_comum +
      b17 * cons_mau + b18 * cons_regular +
      b19 * qtd_apart +
      b20 * qtd_barraco + u;
p = 1 / (1 + exp(-z));
ll = sit_1*log(p)+(1-sit_1)*log(1-p);
model sit_1 ~ general(ll);
RANDOM u~NORMAL(0,sd*sd) SUBJECT=id_bairro;
RUN;
```

As estatísticas e níveis descritivos para cada uma das variáveis de níveis 1 e 2 do **MLGH1** estão relacionadas na Tabela 4.9 a seguir.

Tabela 4.9: Estimativas do MLGH1.

Parâmetro	Estimativa	Erro Padrão	T valor	Significância
b0	2.9824	0.6525	4.58	0.0001
b1	-1.3002	0.08883	-14.64	<.0001
b2	-0.3978	0.09812	-4.05	0.0004
b3	-0.1975	0.09812	-2.37	0.0255
b4	-0.5965	0.08336	-4.93	<.0001
b5	0.3279	0.1211	3.27	0.0030
b6	0.2512	0.1141	2.58	0.0159
b7	-0.9453	0.09743	-3.09	0.0047
b8	-1.1206	0.3061	-3.51	0.0016
b9	-0.8737	0.3191	-2.78	0.0099
b10	-0.5418	0.3138	-1.60	0.1223
b11	0.7364	0.3392	3.34	0.0025
b12	0.7279	0.2202	3.49	0.0017
b13	-0.3955	0.2084	-0.87	0.3899

b14	-1.0423	0.4523	-2.27	0.0319
b15	-0.1300	0.4597	-0.29	0.7768
b16	-0.6617	0.4538	-1.46	0.1569
b17	-0.3505	0.4539	-1.81	0.0812
b18	-0.1540	0.1932	-2.19	0.0378
b19	0.000245	0.07035	3.24	0.0033
b20	-0.01233	0.000076	-2.91	0.0073
Sd	0.1044	0.05155	2.02	0.0533

4.4 – O MLGH2

Já o modelo final dos lotes é formalizado, no *SAS*, da seguinte forma:

```
PROC NLMIXED data=binomial.Lote_binario technique=newrap;
PARMS b0=0 b1=0 b2=0 b3=0 b4=0 b5=0 b6=0 b7=0 b8=0 b9=0 b10=0
sd=1.5;
bounds sd > 0;
      z = b0 +
      b1 * pav_asfalto + b2 * pav_lajota + b3 * pav_terra +
      b4 * topo_abaixo + b5 * topo_nivel +
      b6 * sl_meio + b7 * sl_esquina +
      b8 * log_valor_lancamento +
      b9 * qtd_barraco +
      b10 * tem_clube + u;
p = 1 / (1 + exp(-z));
ll = sit_1*log(p)+(1-sit_1)*log(1-p);
model sit_1 ~ general(ll);
RANDOM u~NORMAL(0,sd*sd) SUBJECT=id_bairro;
RUN;
```

As estatísticas e níveis descritivos para cada uma das variáveis de níveis 1 e 2 do **MLGH2** estão relacionadas na Tabela 4.10 a seguir.

Tabela 4.10: Estimativas do MLGH2.

Parâmetro	Estimativa	Erro Padrão	T valor	Significância
b0	-0.3351	0.4292	-0.78	0.4423
b1	0.5343	0.2185	2.45	0.0218
b2	0.1042	0.2187	0.48	0.6378
b3	-0.3370	0.4427	-1.60	0.1215
b4	-0.8006	0.2921	-1.81	0.0826
b5	0.2969	0.5117	1.02	0.3192
b6	-0.9953	0.1439	-1.94	0.0631
b7	-0.2413	0.1739	-1.68	0.1060

b8	-0.4587	0.1129	-4.06	0.0004
b9	-0.0947	0.04295	-2.20	0.0369
b10	0.1617	0.05214	3.10	0.0047
sd	0.4030	0.08718	4.62	<.0001

4.5 – Interpretação dos modelos estimados

4.5.1 MLGH1

O efeito das variáveis selecionadas no modelo 1, pode-se dividir em dois grupos. As variáveis que tencionam o valor da variável resposta para a situação de pagamento igual a 0 (não pagou), e o outro grupo que tenciona a resposta para a situação de pagamento igual 1, (paga).

De acordo com a Tabela 4.8, podemos separar as variáveis pav_asfalto, pav_lajota, pav_terra, tp_apartamento, tp_casa, tp_deposito, tp_comercio, acab_normal, acab_comum, acab_popular, acab_bom, cons_mau, cons_regular e por ultimo a variável de nível 2, qtd_barraco. Para essas variáveis, a sua incidência em determinado registro, tencionam o valor resposta para o não pagamento do **IPTU-DA**. E de acordo com a variável log_valor_lancamento, quanto maior o valor da mesma, maior a probabilidade do não pagamento.

Já o segundo grupo de variáveis, as que seu indicio, tendência o valor da resposta para o pagamento do tributo, é formado pelas variáveis esp_mista, esp_alvenaria, sl_meio, sl_esquina e qtd_apart.

É importante registrar, que as variáveis com estimativas mais significaticas, as que possuem peso maior na estimação da variável resposta, foram por ordem de

maior impacto: log_valor_lancamento, tp_apart, acab_popular, tp_casa e tp_comercio.

4.5.2 MLGH2

O efeito das variáveis selecionadas no modelo 1, pode-se dividir em dois grupos. As variáveis que tencionam o valor da variável resposta para a situação de pagamento igual a 0 (não pagou), e o outro grupo que tenciona a resposta para a situação de pagamento igual 1, (paga).

De acordo com a Tabela 4.8, podemos separar as variáveis pav_asfalto, pav_lajota, pav_terra, tp_apartamento, tp_casa, tp_deposito, tp_comercio, acab_normal, acab_comum, acab_popular, acab_bom, cons_mau, cons_regular e por ultimo a variável de nível 2, qtd_barraco. Para essas variáveis, a sua incidência em determinado registro, tencionam o valor resposta para o não pagamento do **IPTU**. E de acordo com a variável log_valor_lancamento, quanto maior o valor da mesma, maior a probabilidade do não pagamento.

4.5.3 Comparação entre o MLGH1 e MLGH2 finais

Inicialmente, deve-se atentar para o fato de que os modelos finais, comparados em conjunto, aproveitaram-se, de algumas variáveis significativas em comum. Em termos de nível 1, foram log_valor_lancamento, pav_asfalto, pav_lajota, pav_terra, sl_esquina e sl_meio. E para o nível 2, a única variável que mostrou significativa em ambos modelos, é a qtd_barraco. Para efeitos de

comparação, a Tabela 4.11 lista as estimativas para os coeficientes das variáveis que se apresentaram significativas para os modelos estimados neste capitulo.

Tabela 4.11: Estimativas para variáveis presentes em ambos modelos finais

Variável	Nível	Coeficientes Estimados	
		MLGH1	MLGH2
log_valor_lancamento	1	-1.3002	-0.4587
pav_asfalto	1	-0.3978	0.5343
pav_lajota	1	-0.1975	0.1042
pav_terra	1	-0.5965	-0.3370
sl_esquina	1	0.7364	-0.2413
sl_meio	1	0.7279	-0.9953
qtd_barraco	2	-0.01233	0.1617

A tabela indica que, considerando apenas as variáveis de nível 1, pav_asfalto, pav_lajota, sl_esquina e sl_meio, possuem efeitos contrários em relação aos modelos finais. No modelo **MLGH1** essas variáveis contribuem para elevar a probabilidade da situação de não pagamento do tributo. Porém, essas mesmas variáveis contribuem para elevar a probabilidade de pagamento no **MLGH2**. Já as variáveis log_valor_lancamento e pav_terra apresentam-se com efeitos de mesma orientação para os dois modelos.

Para a variável de nível 2 a situação é inversa, sua estimativa troca de sinal, indicando inversão no sentido do efeito da variável entre um e outro modelo.

O levantamento das motivações que fundamentam essa inversão nas estimativas de algumas variáveis, não se insere entre os objetivos deste trabalho, pois possivelmente, existem vários fatores externos, e inacessíveis no momento para a sua inclusão no modelo.

4.6 – Aplicação e validação dos modelos estimados modelos estimados

Conforme a introdução do capitulo três, a base de dados foi dividida em duas amostras: uma utilizada para a estimação do modelo e outro para executar a validação do mesmo.

A previsão de uma resposta binária em modelos estatísticos envolve basicamente a utilização de uma determinada regra de classificação operacional, e.g., probabilidade predita maior ou igual a 50% classificada como 1, caso contrario classificada como 0. O que claramente requisita a definição de um ponto de corte probabilístico por parte do analista – no caso exemplificado ele seria de 50%.

A aplicação dos modelos **MLGH1** e **MLGH2** à amostra de validação implicam nos seguintes valores para a sensitividade (S) e especificidade (E):

Tabela 4.12: Sensitividade e especificidade para ambos modelos.

Modelo	Sensitividade (%)	Especificidade (%)
MLGH1	20,01	95,64
MLGH2	8,56	98,06

Os resultados da Tabela 4.12 permitem inferir que o **MLGH1** detecta em média 20% dos casos positivos e 95.6% dos casos negativos. E para o **MLGH2** detecta, em média 8.5% dos casos positivos e 98% dos casos negativos.

Estes resultados, para ambos os modelos, indicam a instabilidade dos coeficientes estimados frente às novas observações. Nota-se que os dois modelos tencionam seus resultados para a situação de não pagamento do tributo. Podendo observar na Tabela 4.13 quantitativamente e percentualmente os resultados da validação dos modelos.

Tabela 4.13: Sensitividade e especificidade para ambos os modelos.

Modelo	Situação Pagamento	Real	Estimado	Acertos
MLGH1	Não Pagou	3647	3488	95,64 %
	Pagou	2078	416	20,01 %
MLGH2	Não Pagou	1757	2417	98,06 %
	Pagou	759	99	8,56 %

Neste momento deve-se ressaltar que esses resultados relacionam-se a apenas dois dos muitos possíveis modelos, podendo, portanto serem aperfeiçoados, por exemplo, procedendo-se a rodadas adicionais para testes de variáveis, ou ainda pesquisando informações ainda inexploradas contidas na base de dados e também outras características a titulo de nova variáveis de nível 1 e 2.

5 – CONSIDERAÇÕES FINAIS

O objetivo geral foi construir e aplicar dois modelos capazes de estimar se um imóvel ou um lote terá a tendência de não pagar ou pagar suas dívidas relativas ao tributo **IPTU-DA**. Após a revisão de literatura, análise dos dados e da estimação e aplicação dos modelos propriamente ditos, neste capítulo de fechamento são apresentadas as principais conclusões e contribuições proporcionadas pela pesquisa, assim como identificadas suas limitações e oportunidades para estudos futuros.

5.1 Principais resultados

Neste trabalho de conclusão de curso foram aplicadas técnicas de modelagem hierárquica, com o objetivo de predizer a situação de pagamento do tributo de **IPTU-DA**, de determinado contribuinte do município de Itajaí. Inicialmente pretendia-se que os resultados auxiliassem na tomada de decisão, por parte da prefeitura, e também que estes resultados fossem utilizados como critérios para novas obras ou políticas de arrecadação.

Nesse estudo foram propostos quatro objetivos específicos que apresentavam os meios pelos quais o objetivo geral seria alcançado. Para atingir o primeiro objetivo específico foi preciso analisar o processo de descoberta de conhecimento (metodologias, etapas e técnicas), aliada a métodos já tradicionais de estatística.

Para o segundo, terceiro e quarto objetivo, que tratava de identificar as variáveis preditoras para o modelo, aplicá-lo e validá-lo. Foi necessário

primeiramente analisar os processos que fundamentam os **MLGH** para que fosse obtido o embasamento teórico à sua aplicação.

No capitulo quatro, foram selecionadas as variáveis significativas ao modelo, em seguida aplicado o modelo nulo para ambas as bases, alcançando assim o segundo e terceiro objetivo específico.

Nas subseções finais do capítulo quatro, da estimação e validação dos modelos, obtivesses os principais resultados prático-interpretativos:

I) Dentre as diversas variáveis incluídas no **MLGH1** significativas. As preditoras que demonstraram maior significância foram log_valor_lancamento, tp_apto, acab_popular. Sendo que essas variáveis tencionam a probabilidade de não pagamento ser maior que a do pagamento do tributo.

II) O mesmo acontecendo com as preditoras do **MLGH2,** topo_abaixo, sl_meio e pav_asfalto, tencionando para a situação de não pagamento do lançamento.

III) Em termos de validação dos modelos estimados:

III.I) Ambos modelos finais (**MLGH1 e MLGH2**) não forneceram previsões superiores à mera aleatoriedade, os resultados ficaram claramente tendenciosos para a situação de não pagamento do tributo.

III.II) A validação mostrou que ambos modelos não equilibraram os níveis de sensitividade e especificidade.

III.III) A partir das variáveis selecionadas incluídas nos modelos, não foi possível construção de modelos que obtivessem um bom índice de precisão na predição da situação de pagamento dos contribuintes.

Finalmente, a partir da execução da pesquisa e dos resultados obtidos, pôde-se constatar como já relatado, a conclusão dos quatros objetivos da pesquisa relacionados na subseção 1.1.2. Porém deixando explícito mais uma vez, que os modelos implementados não conseguiram satisfazer de forma plena a predição da variável resposta, situação de pagamento do tributo de **IPTU-DA**.

5.2 Limitações e oportunidades para estudos futuros

Quanto aos possíveis desdobramentos da realização desta pesquisa, uma vez que, em virtude da base de dados disponível, não foram exploradas todas as possibilidades em relação às informações, e também a inclusão de informações de outras origens.

Nesse sentido de exploração de novas possibilidades, podem ainda serem incorporados nos modelos características socioeconômicas dos contribuintes e também aspectos históricos ou temporais.

Alem disso, o método de predição proposto e aplicado nesta pesquisa se delimita aos terrenos ou imóveis do município de Itajaí, com tributos de **IPTU-DA** do exercício 2006.

Os resultados desse trabalho podem motivar as seguintes atividades:

- Incorporar novas características dos bairros (nível 2), assim como características dos contribuintes, podendo criar mais um nível hierárquico.
- Implementar novos modelos com dados de outros tributos, como o tributo de **IPTU**, **ITBI**, **ISS**, entre outros.
- Realizar um estudo para encontrar as prováveis razões que levaram a baixa probabilidade dos modelos implementados neste trabalho.

6 – REFERÊNCIAS BIBLIOGRÁFICAS

BARRETO, A.S. **Previsão de Comportamento e Classificação de Contribuintes Tributários: Uma Abordagem por Modelos Lineares Generalizados Hierárquicos.** Florianópolis: Universidade Federal de Santa Catarina, 2005.

GAZOLA, S. **Construção de um Modelo de Regressão para Avaliação de Imóveis.** Florianópolis: Universidade Federal de Santa Catarina, 2002.

GOLDSTEIN, H. **Multilevel Statistical Models.** 2. Ed. London: Institute of Education, University of London, 1995.

KREFT, I.G.G.; LEEU, W.J.; DLEEDEN, R.V.D. **Review if Five Multilevel Analysis Programs: BMDP-5V, GENMOD, HLM, MLN3, VARCL.** The American Statistican, v. 48, n.4, nov., 1994.

KAMBER, M.; HAN, J. **Data Mining: Concepts and Techniques.** New York: Editora Morgan Kaufmann Publisher, 2001.

OGLIARI, P. J. **Disciplina INE5644 – Data Mining.** Disponível em <http:www.inf.ufsc.br/~ogliari>. Acessado em 8 de outubro de 2008.

MACHADO, Rodrigo Benincá. **Uso de Data Mining e Sistemas de Informações Geográficas no Apoio a Tomada de Decisões.** Trabalho de Conclusão de Curso (Bacharel em Sistemas de Informação), Universidade Federal de Santa Catarina, Florianópolis, 2005.

CHARNET, R.; FREIRE, C. A. de L.; CHARNET, E. M. R.; BONVINO, H. **Análise de Modelos de Regressão Linear com Aplicações.** Editora da Unicamp, 1999.

NETER, J.; WASSERMAN, W.; KUTNER, M. H.; NACHTSHELM, C. J. **Applied Linear Regresseion Models.** 3. Ed. Boston: Times Mirror Hiher Group, Inc., 1996.

PRASS, Fernando Sarturi. **Estudo Comparativo entre Algoritmos de Análise de Agrupamentos em Data Mining.** Dissertação (Mestrado em Ciências da Computação) Universidade Federal de Santa Catarina, Florianópolis, 2004.

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

DAVIDSON, R.; MACKINNON, J. G. **Estimation and Inference in Econometrics.** Oxford University Press, 1993.

DOBSON, A. J. **An Introduction to Generalized Linear Models.** 2d ed. Chapman & Hall/CRC, 2002.

GOLDSTEIN, H. **Multilevel Statistical Models.** 3rd ed. London: Arnold, 2003.

MACHADO, F. N. R. **Tecnologia e Projeto de Data Warehouse.** Rio de Janeiro: Editora Erica, 2002, 320 p.

MADDALA, G. S. **Econometrics.** New York: McGraw-Hill, 1997.

MCCULLAGH, P. e NELDER, J. A. **Generalized Linear Models.** 2 ed. New York: Chapman & Hall/CRC, 1989.

NATIS, L. **Modelos Lineares Hierárquicos. Estudos em avaliação educacional, Fundacao Carlos Chagas**, n. 23, jan/jun 2001.

OGLIARI, P. J. **Modelos Não Lineares para Dados Longitudinais Provenientes de Experimentos em blocos Casualizados.** 1998 Tese (Doutorado em Agronomia) – Escola Superior de Agricultura “Luiz de Queiroz”, USP, Piracicaba.

RAUDENBUSH, S. W.; BRYK, A. S. **Hierarchical Linear Models: Applications and Data Analysis Methods.** 2.ed. Sage Publications, 2002.

SANTOS, C. A. de S. T.; FERREIRA, D. A.; OLIVEIRA, N. F.; DOURADO, M. I. C.; BARRETO, M. L. **Modelagem Multinível. Sitientibus**, Feira de Santana, n. 22, p. 89-98, jan./jun. 2000.

OSBORNE, J. W. **Advantages of Hierarchical Linear Modeling. Practical Assessment, Research & Evaluation**, v. 7(1), 2000. Disponível em <http://edresearch.org/pare/getvn.asp>. Acessado em 13 de junho de 2008.

FAYYAD, U. M.; PIATETSKY-SHAPIRO, G.; SMYTH, P. **From Data Mining to Knowledge Discovery: an Overview.** Cambridge: MIT Press, 1996.

KIMBALL, R. **Data Warehouse Toolkit: Técnicas para Construção de Data Warehouses Dimensionais.** São Paulo: Berkeley, 1998.