Universidade Federal de Santa Catarina
Programa de Pós-Graduação em Engenharia de Produção

CONSTRUÇÃO DE UM MODELO DE REGRESSÃO PARA AVALIAÇÃO DE IMÓVEIS

Dissertação de Mestrado

Sebastião Gazola

Florianópolis
2002

CONSTRUÇÃO DE UM MODELO DE REGRESSÃO PARA AVALIAÇÃO DE IMÓVEIS

Dissertação apresentada no
Programa de Pós-Graduação em
Engenharia de Produção a
Universidade Federal de Santa Catarina
como requisito parcial para obtenção
do grau de Mestre em
Engenharia de Produção.

Orientador: Pedro Alberto Barbetta, Dr.

Florianópolis
2002

Sebastião Gazola

Construção de um Modelo de Regressão para Avaliação de Imóveis

Esta dissertação foi julgada adequada e aprovada para a obtenção do título de **Mestre em Engenharia de Produção** no **Programa de Pós-Graduação em Engenharia de Produção da Universidade Federal de Santa Catarina**

Florianópolis, 08 de outubro de 2002.

Edson Pacheco Paladini, Dr.
Coordenador do Programa

BANCA EXAMINADORA

_____

Prof. Pedro Alberto Barbetta, Dr.
Orientador

_____

Prof. Norberto Hochheim, Dr.

_____

Prof. Paulo José Ogliari, Dr.

A minha esposa, Vilma
pelo apoio constante.
A meus filhos Vicente e Marina.

AGRADECIMENTOS

À minha família, pelo apoio e compreensão, em especial à minha esposa Vilma;

Ao Departamento de Estatística da UEM;

À Profª. Terezinha Aparecida Guedes, pelo apoio;

À Profª. Isolde Previdelli, pela viabilização do curso;

À todos os alunos do curso pelo apoio e amizade e em especial às colegas Angela, Clara, Clédina e Zeza;

Ao Programa de Pós-graduação em Engenharia de Produção da UFSC pelo empenho;

À Profª. Eunice Passaglia e toda equipe do LED, responsáveis pelo funcionamento do curso;

Ao meu orientador, Prof. Pedro Alberto Barbetta, que apoiou o desenvolvimento do trabalho e ofereceu todas as contribuições necessárias para sua realização;

a todos que direta ou indiretamente
contribuíram para a realização
desta pesquisa.

Resumo

GAZOLA, Sebastião. **Construção de um modelo de regressão para avaliação de imóveis**. 2002. 110f. Dissertação (Mestrado em Engenharia de Produção) – Programa de Pós-Graduação em Engenharia de Produção, UFSC, Florianópolis.

Este trabalho apresenta uma estratégia de construção de um modelo de regressão para determinar o preço de um imóvel em função de suas características. O modelo foi determinado utilizando-se a Regressão Linear Múltipla com a técnica de *Ridge Regression*, para contornar o problema de multicolinearidade. A estratégia de construção foi aplicada a um conjunto de dados referentes a apartamentos da cidade de Criciúma, SC. O modelo determinado apresentou-se de fácil interpretação e utilização, utilizando 11 variáveis independentes e proporcionando um bom ajuste aos dados e uma boa capacidade preditiva. Ele atendeu à todas as suposições teóricas para sua existência e utilização.

Palavras-chave: Regressão Linear Múltipla, Avaliação de Imóveis, Multicolinearidade.

Abstract

GAZOLA, Sebastião. **Construção de um modelo de regressão para avaliação de imóveis**. 2002. 110f. Dissertação (Mestrado em Engenharia de Produção) – Programa de Pós-Graduação em Engenharia de Produção, UFSC, Florianópolis.

This work presents a strategy for the building of a regression model to determine the price of a property as a function of its characteristics. The model was determined using the Multiple Linear Regression with the Ridge Regression technique, to outline the multicollinearity problem. The construction strategy was applied to a data set from flats of the Criciúma city, SC. The model was shown to be of easy interpretation and use, applying 11 independent variables and providing a good adjustment to the data and a good predictive capacity. The theoretical suppositions for its existence and usage were met.

Key-Words: Multiple Linear Regression, Evaluation of flats, Multicolinearity.

SUMÁRIO

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA DE TABELAS

_____ Capítulo 1

INTRODUÇÃO

1.1 - CONTEXTUALIZAÇÃO

Os primeiros estudos sobre avaliação de imóveis no Brasil datam de 1918, e em 1923 foram introduzidos novos métodos de avaliação de terrenos, que a partir de 1929 começaram a ser sistematicamente aplicados. A partir daí a engenharia de avaliação no Brasil vem crescendo e evoluindo nas técnicas de avaliação. Atualmente um grande número de profissionais vem desenvolvendo estudos nesse campo, visando dar à matéria o suporte cientifico necessário como apoio aos métodos técnicos até então utilizados (Fiker, 1997, p.17).

O desenvolvimento da engenharia de avaliação, o crescimento do número de profissionais atuando nesse campo e as necessidade do uso das técnicas de avaliação pelo mercado privado e também pelos órgãos públicos, levaram a Associação Brasileira de Normas Técnicas a elaborar a Norma para Avaliação de Imóveis Urbanos n. 5676/89 (Antiga NB 502).

A Caixa Econômica Federal é responsável por aproximadamente 200.000 laudos de avaliação por ano, envolvendo em torno de 3.500 engenheiros, sendo a maior entidade contratante ou executora de serviços de avaliação de imóveis no país (Dantas, 1998).

A possibilidade de contribuir, através de uma simples, clara e acessível metodologia cientifica de Regressão Linear Múltipla e Inferência Estatística, para essa gama de profissionais avaliadores e empresas, no que se refere a dar subsídios para a melhoria da

qualidade das avaliações, é a principal justificativa para a realização deste trabalho.

Uma justificativa do porquê realizar a avaliação em massa de imóveis é, principalmente, para a elaboração ou atualização do cadastro fiscal dos municípios, pois é a base para cobrança dos tributos. Os métodos usados para manter estes cadastros atualizados são, na maioria das vezes, não satisfatórios do ponto de vista da realidade do mercado imobiliário, e também muito onerosos aos municípios. Por isto, a grande maioria dos municípios tem seu cadastro desatualizado, levando a injustiças na tributação e grande perda na arrecadação dos tributos, os quais são fundamentais para a administração pública (Zancan, 1995, p.12-17). Assim, uma metodologia adequada, ou seja, que tenha bom poder de predição e não seja onerosa, é de fundamental importância aos municípios.

Outra importância para a avaliação de imóveis é a realização de laudos de avaliação, relativos a programas habitacionais, a patrimônios da União, seguros, entre outros (Dantas, 1998).

A regressão linear múltipla foi escolhida como método para ser aplicado à avaliação de imóveis por fornecer um modelo de fácil interpretação e, principalmente, de simples aplicabilidade .

Assim, o presente trabalho procura oferecer uma contribuição na área da Engenharia de Avaliação, mais especificamente, nos problemas de avaliação em massa de imóveis urbanos. Os dois pontos principais dessa contribuição são: primeiro, a metodologia para construção de um modelo de regressão linear múltipla para estimar o preço do imóvel, com um bom ajuste aos dados observados e que forneça uma estimativa calibrada, não distante da realidade; e o segundo ponto consiste na metodologia de análise e diagnóstico do modelo construído.

1.2 - PROBLEMA

É possível construir um modelo de regressão linear múltipla que atenda todas as suposições teóricas e que seja adequado para predizer o valor de um imóvel em função de suas características?

1.3 – OBJETIVOS

1.3.1 – Objetivo Geral

O objetivo geral é descrever como se pode construir um modelo de regressão linear múltipla, que possa predizer o valor de um imóvel em função de suas características.

1.3.2 – Objetivos Específicos

1 - Identificar um conjunto de variáveis independentes significativas para comporem o modelo de regressão;

2 - Verificar se o modelo proposto atende todas as suposições teóricas consideradas inicialmente para sua existência;

3 - Realizar o diagnóstico do modelo para tomada de decisão quanto a sua aceitação;

4 – Descrever uma estratégia, para a construção de modelos de regressão para a avaliação de imóveis;

5 – Aplicar o roteiro e a estratégia proposta em um caso prático.

1.4 – MÉTODOS DE DESENVOLVIMENTO DO TRABALHO

Com a finalidade de alcançar o objetivo principal, que é a obtenção do modelo, desenvolve-se um estudo da teoria de regressão linear múltipla e de todas as técnicas consideradas necessárias para a obtenção de um modelo melhor do que foi encontrado na pesquisa bibliográfica.

As primeiras ferramentas descritas são para atender o objetivo de estudo das variáveis que participam da construção do modelo de regressão linear múltipla. As técnicas para isto são a análise de correlação, o determinante da matriz de correlação e as raízes características.

Para escolha das variáveis independentes realiza-se um estudo da técnica de seleção de variáveis “Stepwise” (“Geral”, “Forward” e “Backward”) e, como ferramenta de apoio, realiza-se também um estudo da Análise Fatorial de Correspondências para investigar a existência de associações entre as variáveis.

O refinamento do modelo empírico é realizado através do estudo da inferência estatísticas aplicada à regressão linear múltipla, testes de hipóteses e intervalos de confiança.

O objetivo terceiro, diagnóstico do modelo determinado, é atingido pelas técnicas gráficas da análise de resíduos.

1.5 – DELIMITAÇÕES DA PESQUISA

Dentre as várias técnicas que podem ser utilizadas para a avaliação de imóveis, restringiu-se nesta pesquisa, ao uso das técnicas clássicas de regressão linear múltipla.

A aplicação da estratégia de construção de um modelo de regressão, proposto nesta dissertação, restringe-se aos imóveis descritos por Zancan (1995). Observa-se que os dados foram obtidos pela própria pesquisadora, usando o banco de dados imobiliário de Criciúma, SC, em confronto com o cadastro urbano do município. A amostragem utilizada não está claramente especificada.

1.6 – ESTRUTURA

A dissertação está estruturada em cinco capítulos, construídos de forma à facilitar o entendimento e compreensão do leitor desde os objetivos até a conclusão.

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

O segundo capítulo trata da revisão de literatura sobre a avaliação de imóveis. Apresenta os principais conceitos e definições relacionadas a definição de imóveis urbanos.

O capítulo três apresenta toda metodologia estatística necessária, de forma detalhada, da construção até o diagnóstico do modelo, para realização da análise de dados com a finalidade de atingir os objetivos do trabalho.

O capítulo quatro descreve um roteiro estratégico para a construção do modelo e faz a análise dos dados, do caso dos apartamentos da

cidade de Criciúma (Zancan, 1995), descrevendo detalhadamente os resultados e realizando a análise para cada um deles.

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

Capítulo 2

AVALIAÇÃO DE IMÓVEIS

2.1 - INTRODUÇÃO

A necessidade de uma avaliação adequada de imóveis cresceu e evoluiu junto com o próprio crescimento e evolução do mercado imobiliário, já que ambos são complementares e interdependentes.

A tentativa de contextualizar o ícone “avaliação” cria uma lista interminável de conceitos, onde cada autor contextualiza de acordo com as suas prioridades pessoais. Na verdade, isto é possível se considerarmos a multidisciplinidade da avaliação, ou seja, sua dependência tanto de técnicas racionais (nas áreas de ciências exatas, naturais e sociais) como de percepção não-racional (bom-senso e bom-julgamento). A introdução do “feeling” é que torna possível a variabilidade contextual (IBAPE, 1974, p.64; Ayres, 1996, p.11), e a multidisciplinidade é que torna a avaliação flexível (Dantas, 1998, p.3).

Existem várias formas diferentes de se desenvolver as avaliações de imóveis, dependendo dos dados disponíveis ou da preferência do avaliador. Outras vezes, o imóvel pode ser avaliado percorrendo-se caminhos diferentes para a confirmação do valor de avaliação (Moreira Filho, *et al*. 1993, p.4).

De forma geral e resumida, pode-se definir avaliação como uma operação técnica realizada na estimativa do valor de um bem; ou como uma determinação técnica do valor de um imóvel e/ou de um direito sobre ele (NBR – 5676/90); ou ainda como uma arte, dependente de conhecimento técnico e de bom-senso, de estimar valores à propriedades específicas (Moreira, 1990).

Sem dúvida, esta última definição é genericamente mais apropriada, pois considera o conhecimento técnico e a capacidade de percepção

não-técnica de um bem. Agora, independentemente do conceito de avaliação que tomemos como adequada, é necessário definir o valor do bem considerado e o próprio mercado imobiliário.

2.2 - MERCADO IMOBILIÁRIO

O mercado imobiliário é a instância de determinação dos preços de imóveis urbanos que, como quaisquer outras mercadorias, passa pelo crivo da oferta e da demanda (Moscovitch, 1997).

A existência do mercado imobiliário depende da presença de três componentes: os bens imóveis disponíveis, os vendedores e os compradores. Assim sendo, o fator determinante na formação dos preços será a relação quantitativa dentre os três, onde a situação ideal será aquela onde haja uma abundância equilibrada dentre os mesmos. Isto determinará, num dado momento, um preço de equilíbrio de mercado que podemos considerar como sendo um preço justo. Este mercado, considerado como sendo de concorrência perfeita, é inatingível. O extremo à esta situação, ou seja, um mercado de concorrência imperfeita, cria um desbalanço que faz os preços se afastarem do ideal ou justo. É o caso do monopólio (raro) e oligopólio (mais comum) que viesam os preços para cima; ou do monopsônio (raro) e oligopsônio (mais comum) que viesam os preços para baixo. Obviamente, somente no mercado de concorrência perfeita, a construção do valor de um bem pode seguir a lei da oferta e procura (Dantas, 1998, p.9).

Os imóveis são bens economicamente únicos, pois são heterogêneos, fixos e duráveis. A durabilidade permite a formação de estoque que domina o mercado, a heterogeneidade o torna insubstituível, e a imobilidade o relaciona com a acessibilidade e com a estrutura vicinal (Smith *et al*., 1988). Se associarmos isto ao fato do mercado ser particulado e, portanto, contar com a ação simultânea de vários agentes não coordenados, poderemos explicar, pelo menos

parcialmente, a enorme variabilidade de preços (González e Formoso, 2000).

A indeslocabilidade torna o imóvel um bem imperfeito por natureza, ou seja, com diferenças inter e intra grupos de bens. Por este motivo, o mercado imobiliário será sempre de concorrência imperfeita podendo apresentar todas as gradações de imperfeição. Dessa forma, cada bem imobiliário acabará por gerar em torno de sí um micro-mercado que guardará caracteres tão intimamente relacionados que dificultará a relação deste com o macro-mercado que o circunda. Isto dificulta a avaliação do bem porque condiciona a avaliação à coleta de dados do micro-mercado considerado. Se os elementos amostrais forem insuficientes dentro do micro-mercado, a coleta de elementos do macro-mercado circundante gerará tendências de mercado que invalidariam a avaliação (Auricchio *apud* Trivelloni, 1998, p.12).

2.3 - VALOR

Atribui-se valor a tudo que é útil ou escasso. Cabe à avaliação traduzir essa utilidade ou escassez numa quantia monetária e associar à uma necessidade e/ou desejo de possuir um bem (Ayres, 1996, p.21). Assim, pode-se definir valor como a relação entre a intensidade das necessidades econômicas humanas, objetivas ou subjetivas, e a quantidade de bens disponíveis para atendê-las (Fiker, 1997, p.21).

Vários tipos de valores podem ser atribuídos a um bem (Venal, Comercial, de Mercado, etc). No entanto, numa avaliação, o valor a ser determinado é o valor de mercado (Dantas, 1998, p.7). Estas atribuições são impostas pelo mercado que determina o valor pela lei da oferta e da procura. Assim, o valor de mercado é o preço consciente determinado por um vendedor e pago por um comprador a um bem, sem coação de ambos os lados (Ayres, 1996, p.21).

O valor de um bem pode ser subjetivado dependendo das circunstâncias que envolvem a avaliação e do modo como é examinado, mas sempre dependerá de sua utilidade. A localização do imóvel é um

componente essencial de seu valor. Este valor estaria correlacionado a aspectos que compõem a qualidade de vida da área urbana onde o imóvel está situado, por exemplo, as áreas urbanas mais bem providas de equipamentos públicos são as que possuem imóveis com maiores valores venais. Por outro lado, independente da sua utilização, o valor de um imóvel é a soma de dois sub-valores aditivos, o valor da edificação que é dada pelo seu custo (incluída a remuneração do construtor) e o valor do terreno que está intimamente relacionado às condições urbanas de sua localização (incorpora as vantagens e desvantagens espaciais) (Moscovitch, 1997).

No entanto, o valor do bem difere e não deve ser confundido com o preço do bem, que representa a quantidade de dinheiro paga pelo mesmo. Assim, a necessidade de venda ou compra imediata e/ou a não existência de um livre comércio podem alterar o preço de um bem, tornando-o superior ou inferior ao valor avaliado (Moreira Filho, 1993). Dessa forma, defini-se o preço hedônico como o preço implícito de atributos e são revelados à agentes econômicos a partir da observação de preços de produtos diferenciados e a quantidade específica de características a eles associados (Rosen, 1974).

O valor de mercado é normatizado pela NBR 5676/90, como um valor único num dado instante, independente da finalidade da avaliação e subjugada a um mercado de concorrência perfeita. Obviamente, o mercado imobiliário não é, pela sua própria natureza, de concorrência perfeita (Dantas, 1998, p.8). Na verdade, o mercado imobiliário é um dos segmentos de mercado que mais se ajusta ao mercado teórico da concorrência imperfeita. Isto faz com que o preço de um bem seja desviado daquele determinado teoricamente pelo mercado de concorrência perfeita (Barbosa Filho, 1988). Portanto, o que realmente se paga numa negociação imobiliária é o preço e não o valor (Dantas, 1998, p.8).

Existe, portanto, a necessidade da busca por técnicas que tornem mais precisas as formas de se estimar o valor de um bem aproximando-o ao máximo do seu valor de mercado.

2.4 - MÉTODOS DE AVALIAÇÃO

Pode-se definir as metodologias avaliatórias como sendo as várias e diferentes vias percorridas com o objetivo de atribuir valor a um imóvel. Cada via utilizada é caracterizada como um método de avaliação diferente. No entanto, independentemente da metodologia aplicada, esta deverá apoiar-se em pesquisa de mercado e considerar os preços comercializados e/ou ofertados, bem como outros elementos e atributos que influenciam o valor (NBR-5676/90). A escolha da metodologia mais apropriada para uma dada avaliação depende das condições atuais do mercado, do tipo de serviço a que se presta e da precisão que se deseja.

Os métodos avaliatórios pertencem a dois grupos, por vezes conjugados, os métodos diretos e os métodos indiretos. Considera-se um método como sendo direto quando o valor resultado da avaliação independe de outros. Por outro lado, o método considerado indireto sempre necessita de resultados de algum método direto (Dantas, 1998, p.15).

Os métodos diretos subdividem-se em método comparativo de dados de mercado e método comparativo de custo de reprodução de benfeitorias. Já os métodos indiretos organizam-se em três grupos, o método de renda, o método involutivo e o método residual (NBR-5676/90).

No método comparativo de dados de mercado, o valor do bem é avaliado por comparação com dados do mercado similares quanto as características intrínsecas e extrínsecas; para isto exige a presença de um conjunto atual de dados que represente estatisticamente o mercado. Portanto, qualquer bem pode ser avaliado por este método, desde que existam dados suficientes e atuais no mercado imobiliário que possam ser utilizados para representá-los estatisticamente (Trivelloni, 1998, p.20; NBR 5676/90).

Pelo método comparativo de custo de reprodução de benfeitorias, o valor das benfeitorias é avaliado pela reprodução dos custos componentes, via composição dos custos baseada em orçamento

simples ou detalhado, podendo incluir o valor do terreno e o custo da comercialização e considerando o grau de desgaste físico e/ou o arcaicaísmo funcional (Zancan, 1995; NBR 5676/90).

O método da Renda avalia o valor do imóvel ou de suas partes componentes em função de um rendimento já existente ou previsto pelo bem no mercado, ou seja, o valor econômico do bem (Ayres, 1996, p.23; NBR 5676/90).

No método involutivo, o valor do terreno é estimado por estudos da viabilidade técnica-econômica do seu aproveitamento, considerando como aproveitamento eficiente a realização de um empreendimento imobiliário hipotético compatível com as características do imóvel e com as condições do mercado (Moreira Filho, 1993, p.5; NBR 5676/90).

Já pelo método residual, obtêm-se o valor do terreno a partir da diferença entre o valor total do imóvel e o valor das benfeitorias, levando-se em conta o fator de comercialização (Fiker, 1997, p.27; NBR 5676/90).

A utilização dos métodos diretos têm preferência e sempre que existirem dados de mercado suficientes para utilização do método comparativo ele deve ser escolhido (Dantas, 1998, p.15)

Quando analisa-se os vários métodos citados anteriormente, pode-se observar que de uma forma, ou de outra, todos são comparativos. No método comparativo comparam-se bens semelhantes; no método de custo, comparam-se os próprios custos no mercado; nos métodos da renda e involutivo compara-se a possibilidade de renda do bem; e no método residual, compara-se o grau de comercialização do mercado (Dantas, 1998, p.44).

No entanto, quando a questão é avaliação de imóveis, o método mais utilizado e recomendado é o método comparativo de dados de mercado, já que este método permite que a estimativa considere as diferentes tendências do mercado imobiliário que, por sua vez, diferenciam-se das tendências de outros ramos da economia. Este método estima valores baseado na comparação com outros semelhantes, partindo-se de um grupo de dados somado às informações sobre transações e ofertas do mercado, e originando com

isto uma amostragem estatística de dados do mercado imobiliário. Na prática, de modo geral, a semelhança entre o imóvel avaliado e os componentes da amostra é imperfeita e incompleta, por faltar algum atributo que tenha influenciado no valor ou por apresentá-lo de forma parcial. Portanto, os atributos dos dados pesquisados que influenciam o valor devem ser ponderados por homogeneização ou inferência estatística, respeitando os níveis de rigor definidos na NBR-5676/89. A utilização da inferência estatística permite uma avaliação isenta de subjetividade e repleta de confiabilidade (Moreira Filho, 1993, p.7; González, 2000).

Dentro deste contexto, podemos verificar que, tradicionalmente, usavam-se as tabelas na comparação de vendas para justificar o estado real e/ou estimar valores aproximados. Mais recentemente, os modelos de preços hedônicos (regressão múltipla) tem sido utilizados para completar o método de comparação de vendas. Contudo, os dois métodos tem experimentado críticas das comunidades acadêmica e profissional. O primeiro método é, freqüêntemente, criticado por utilizar julgamentos subjetivos para determinar os ajustes necessários e também, por ser impreciso, tornando difícil para o avaliador obter dados seguros e comprovados. A regressão múltipla tem produzido, freqüêntemente, sérios problemas para a avaliação do estado real que resulta, primariamente, de estudos de multicolinearidade nas variáveis independentes e a partir de inclusões de propriedades “outlier” na amostra. Além disso, a colinearidade dentro dos dados pode tornar a regressão múltipla um modelo inadequado para um mercado que requer respostas rápidas e precisas. No entanto, a regressão é um método padrão aceitável para a avaliação de imóveis. (Worzala *et al*. 1995).

2.5 – AMOSTRAGEM EM MERCADO IMOBILIÁRIO

Quando se trabalha com o mercado imobiliário, qualquer que seja o segmento, terrenos urbanos, imóveis tipo apartamentos, imóveis tipo residências, etc., geralmente é impraticável a obtenção dos dados de

toda a população. Isto ocorre devido ao grande número de elementos na população, custos elevados para obtenção dos dados ou o grande período de tempo que se faz necessário. Assim, é conveniente trabalhar com uma amostra (Dantas, 1998, p.69).

A situação ideal para uma amostra é aquela onde cada elemento da população tem a mesma probabilidade de ser selecionado, ou seja, uma amostra do tipo probabilística aleatória. Em muitos casos, ainda, se faz necessário uma amostra do tipo probabilística aleatória estratificada. Esta última deve ser usada quando se tem, por exemplo, regiões e a população de indivíduos difere consideravelmente de uma para a outra.

Quando se trabalha com dados de mercado é muito difícil de se ter uma amostra estatisticamente ideal. Para não inviabilizar as inferências, deve-se evitar usar um banco de dados, sem a investigação do mercado no momento de realizar uma nova avaliação. Podem ter ocorrido mudanças no mercado e estas não poderiam deixar de serem captadas pela amostra, caso contrário a amostra seria tendenciosa. Ainda, a amostra deve ser equilibrada, por exemplo, quando uma categoria for exageradamente maior que as outras, acima de 70%, deve-se ajustar um modelo específico para tal categoria. A amostra deve ser formada por imóveis cujos preços, ou valores, são os praticados no mercado e com todas suas características físicas, locacionais e econômicas (Dantas, 1998, p.49).

O preço praticado é aquele que resulta de uma livre negociação entre o vendedor e o comprador. Este preço é representativo do mercado ou da população em estudo. Já os preços de oferta podem elevar o valor da média dos preços praticados no mercado, podendo servir como um indicador do limite superior dos preços de mercado. Contudo, podem fazer parte da amostra, desde que, atualizados e identificados. Por outro lado, não devem compor a amostra, os preços provenientes de desapropriações, transmissão “causa mortis”, transações entre parentes e outro (Dantas, 1998, p49-51).

2.6 - NÍVEIS DE RIGOR

Os níveis de rigor que caracterizam uma determinada avaliação de acordo com a precisão obtida no trabalho, são normatizados pela NBR 5676/90. O nível de rigor almejado numa dada avaliação relaciona-se diretamente com as informações extraídas do mercado, ou seja, a precisão do mercado será determinada por este nível que será, por sua vez, tanto maior quanto menor for a subjetividade presente na avaliação. O rigor de uma avaliação está condicionado à abrangência da pesquisa, à confiabilidade e adequação dos dados coletados, à qualidade do processo avaliatório e ao menor grau de subjetividade empregado pelo avaliador. Assim, os trabalhos avaliatórios podem, de acordo com a norma, ser classificados como de nível de rigor expedito, normal, rigoroso e rigoroso especial.

Na avaliação expedita o valor é obtido sem a utilização de qualquer instrumento matemático. Dessa forma, a ausência de rigor matemático determina que o valor seja atribuído através de escolha arbitrária, não caracterizando o aspecto técnico da avaliação, e bastando somente que o avaliador tenha bom nível de conhecimento de mercado.

A avaliação normal utiliza métodos estatísticos e requer exigências com relação à coleta e tratamento dos dados. Permite a homogeneização dos elementos e a eliminação estatística de dados discrepantes sempre que o número destes for maior ou igual a cinco.

Nas avaliações rigorosas, o trabalho deverá apresentar, através de metodologia adequada, isenção de subjetividade. O tratamento dos dados devem se basear em processos de inferência estatística que permitam calcular estimativas não tendenciosas do valor. O valor final da avaliação, resultado do tratamento estatístico adotado, deve estar contido em um intervalo de confiança fechado e máximo de 80%, desde que as hipóteses nulas sejam testadas ao nível de significância máximo de 5%.

A avaliação rigorosa especial caracteriza-se pelo encontro de um modelo estatístico o mais abrangente possível, ou seja, que incorpore o maior número de caracteres que contribuem para a formação do valor.

A função estimada da formação de valor deve ser eficiente e não tendenciosa, portanto, as hipóteses nulas da equação de regressão devem ser rejeitadas ao nível de significância máximo de 1%, e dos respectivos coeficientes ao nível de significância máximo de 10% unicaudal ou 5% em cada ramo do teste bicaudal. Devem ser analisadas as seguintes condições básicas referentes aos resíduos não explicados: normalidade, homocedasticidade, não auto-regressão e independência entre variáveis independentes.

Capítulo 3

MÉTODOS - REGRESSÃO LINEAR MÚLTIPLA

3.1 - INTRODUÇÃO

A origem do termo “Regressão” deu-se por Francis Galton, quando em um ensaio com pais e filhos ele estudou o relacionamento das alturas dos mesmos. A lei de regressão universal de Galton foi confirmada mais tarde por Karl Pearson, que através de um grande ensaio constatou que a altura média dos filhos de pais altos era inferior a altura de seus pais e que a altura média de filhos de pais baixos era superior a altura de seus pais, ou seja, ele concluiu que a altura tanto dos filhos altos como baixos tendem para a média de todos os homens (Gujarati, 2000, p.3).

Na atualidade, a interpretação da regressão é bem diferente. De modo geral pode-se dizer que a análise de regressão é o estudo de uma variável (a variável dependente) em função de uma ou mais variáveis (as variáveis independentes), com o objetivo de estimar e/ou prever a média populacional ou valor médio da variável dependente, utilizando valores observados por amostragem das variáveis independentes (Gujarati, 2000, p.9).

Atualmente a análise de regressão múltipla é uma das ferramentas ou métodos estatísticos utilizados com maior freqüência. É uma metodologia estatística para predizer valores de uma variável resposta (dependente) para uma coleção de valores de variáveis preditoras (independentes).

Em engenharia de avaliações, considera-se geralmente como variável dependente os preços à vista de mercado em oferta e efetivamente transacionados, e como variáveis independentes as características do imóvel decorrentes dos aspectos físicos e de

localização, bem como de aspectos econômicos. Observa-se que as variáveis independentes podem ser tanto de natureza quantitativa como qualitativa (Dantas, 1998, p.51,52).

A teoria econômica especifica tipicamente relações funcionais exatas entre variáveis. Porém, na realidade, não se verifica tal relação funcional exata. Isto diz que a teoria econômica deve ser ampliada com a introdução de elementos probabilísticos. Assim, a tarefa principal é administrar um ponto entre as relações exatas e as relações instáveis da realidade econômica (Goldberger, 1970, p.11-16).

3.2 - O MODELO

O modelo de regressão linear múltipla descreve uma variável dependente Y como função de várias variáveis regressoras ou independentes. Um modelo geral, com *p* variáveis regressoras, é dado por:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ \ldots \ + \beta_p X_{ip} + \varepsilon_i \quad (i=1,\ldots,n).$$

onde:

Y_i – representa as observações da chamada variável dependente, variável explicada ou variável resposta;

X_{ik} – são chamadas de variáveis independentes, variáveis explicativas, variáveis regressoras ou covariáveis (k = 1, 2, ..., p);

β_i – são os parâmetros da população;

ε_i – são os erros aleatórios

Os erros aleatórios representam os inúmeros fatores que, conjuntamente, podem interferir nas observações da variável dependente Y (Charnet *et al*., 1999, p.170).

A representação do modelo na forma matricial é $Y = X\beta + \varepsilon$, onde:

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} \qquad \mathbf{X} = \begin{bmatrix} 1 & X_{11} & \cdots & X_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & \cdots & X_{np} \end{bmatrix} \qquad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_p \end{bmatrix} \qquad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

A função de regressão do modelo, descrita em termos de valor esperado, é dada por:

$$E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ \ldots \ + \beta_p X_p$$

Nesse modelo, x_j é o valor fixo da variável regressora X_j, j=1,2,...,p. Os parâmetros β_j são denominados coeficientes de regressão. Podemos interpretar β_j como a mudança esperada em Y devido ao aumento de uma unidade em X_j, estando as outras variáveis X_k, $k \neq j$, fixas.

O coeficiente β_0 é o intercepto da superfície de resposta (regressão). Se a abrangência do modelo inclui (0, 0, ..., 0) então β_0 representa a resposta média E(Y) neste ponto. Em outras situações, β_0 não tem qualquer outro significado como um termo separado no modelo de regressão.

Um dos objetivos da análise de regressão é desenvolver uma equação que permita ao investigador estimar respostas para valores dados de variáveis preditoras. Para descrever a equação é necessário estimar os valores para os coeficientes de regressão β e a variância σ^2 do erro com os dados observados.

Os coeficientes de regressão podem ser estimados por vários métodos, um dos mais usados é o método de mínimos quadrados. Este método consiste em encontrar uma estimativa para os parâmetros de forma que a soma do quadrado dos erros seja mínima. Os estimadores gerados por este método são não viesados e consistentes (Neter e Wasserman, 1974, p.37,226).

O estimadores para o vetor de parâmetros β e para a variância σ^2 são dados, respectivamente, por:

$$\mathbf{b}=(\mathbf{X'X})^{-1}\mathbf{X'}Y \quad \text{e} \quad \mathbf{S}^2=(\mathbf{Y-Xb})'(\mathbf{Y-Xb})/(n-p-1).$$

As suposições exigidas para o modelo de regressão linear múltipla, além das estimativas dos parâmetros, são as seguintes:

1) as variáveis independentes são números reais sem perturbações aleatórias.

2) o número de observações, n, deve ser superior ao número de parâmetros, p, estimados.

3) os erros são variáveis aleatórias com as seguintes suposições:
- valor esperado zero - $E(\varepsilon_i) = 0$;
- variância constante - $Var(\varepsilon_i) = \sigma^2$;
- não correlacionados - $Cov(\varepsilon_i, \varepsilon_j) = 0$, $i \neq j$.

4) a distribuição dos erros é normal, $\varepsilon_i \approx N(0, \sigma^2)$. Como os erros são não correlacionados, pode-se afirmar, sob a hipótese de normalidade, que estes são independentes.

5) não deve existir nenhuma relação exata entre as variáveis independentes.

3.3 - INFERÊNCIA ESTATÍSTICA

Os parâmetros populacionais são estimados pontualmente a partir de uma amostra, porém é necessário obter informações sobre seu comportamento probabilístico. Este estudo é realizado através dos intervalos de confiança e testes de hipóteses.

3.3.1 - Intervalos de Confiança

O intervalo de confiança fornece informação sobre a precisão das estimativas. É o intervalo do qual pode-se 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 intervalos de confiança mais usuais em uma análise de regressão são descritos a seguir.

Intervalo de confiança para o parâmetro β_k: para o modelo onde os erros têm distribuição normal, o intervalo de confiança para β_k, é dado por

$$(b_k - t_{(1-\alpha/2\ ;\ n-p-1)} \cdot S(b_k)\ ;\ b_k + t_{(1-\alpha/2\ ;\ n-p-1)} \cdot S(b_k))$$

onde b_k é o estimador de β_k, $t_{(1-\alpha/2\ ;\ n-p-1)}$ é o valor da estatística t com significância α e (n-p-1) graus de liberdade e $S(b_k)$ é o desvio-padrão estimado de b_k. $S^2(b_k)$ é o k-ésimo elemento da diagonal principal da matriz:

$$\mathbf{S}^2(\mathbf{b}) = QME(\mathbf{X'X})^{-1}$$

Intervalo de confiança para valores médios preditos: o valor médio estimado para um caso (imóvel) i é dado por $\hat{Y}_i = \mathbf{X}_i'\mathbf{b}$, ($\mathbf{X}_i' = [\mathbf{1} \quad \mathbf{X_i}]$). O intervalo de confiança para o valor médio estimado é calculado por:

$$(\hat{Y}_i - t_{(1-\alpha/2\ ;\ n-p-1)} \cdot S(\hat{Y}_i)\ ;\ \hat{Y}_i + t_{(1-\alpha/2\ ;\ n-p-1)} \cdot S(\hat{Y}_i))$$

onde $\hat{Y}_k$ é o valor médio estimado para o caso i, $t_{(1-\alpha/2\ ;\ n-p-1)}$ é o valor da estatística t com significância α e (n-p-1) graus de liberdade e $S(\hat{Y}_i)$ é o desvio-padrão de $\hat{Y}_i$. $S^2(\hat{Y}_i)$ é dada por

$$S^2(\hat{Y}_i) = (QME)\,\mathbf{X}_i'(\mathbf{X'X})^{-1}\mathbf{X}_i = \mathbf{X}_i'\mathbf{S}^2(\mathbf{b})\mathbf{X}_i$$

Desta forma, é possível comparar o valor observado com o valor estimado e a precisão do ajuste.

Intervalo de confiança para valores preditos: o valor predito para um novo caso observado h é dado por $\hat{Y}_h = \mathbf{X}_h'\mathbf{b}$. E o intervalo de confiança para este novo caso observado é dado por

$$(\hat{Y}_h - t_{(1-\alpha/2\ ;\ n-p-1)} \cdot S(pred)\ ;\ \hat{Y}_h + t_{(1-\alpha/2\ ;\ n-p-1)} \cdot S(pred)).$$

Onde $\hat{Y}_h$ é o valor predito para o novo caso h, $t_{(1-\alpha/2\ ;\ n-p-1)}$ é o valor da estatística t com significância α e (n-p) graus de liberdade e S(pred) é o desvio-padrão do valor predito. S^2(predito) é dado por

$$S^2(\text{predito}) = QME(1 + \mathbf{X}_h'(\mathbf{X'X})^{-1}\mathbf{X}_h)$$

3.3.2 - Testes de Hipóteses

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. Os testes de hipóteses necessários em uma análise de regressão são descritos abaixo.

Teste de hipótese para a significância do modelo: este teste é usado para estabelecer se existe ou não alguma relação entre a variável dependente e o conjunto de variáveis independentes. Consiste em testar as seguintes hipóteses (Neter e Wasserman, 1974, p.228):

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_p = 0$$
$$H_1 : \beta_k \neq 0 \text{ para algum } k$$

A estatística do teste tem, sob H_0, a distribuição F com p e (n-p-1) graus de liberdade. A rejeição da hipótese H_0 indica a existência de regressão.

As quantidades necessárias para calcular o valor observado dessa estatística estão dispostas na tabela 3.1, denominada de tabela de análise de variância - ANOVA.

Tabela 3.1: Análise de Variância para testar a significância do modelo

Fontes de Variação	Soma de Quadrados	Graus de Liberdade	Quadrado Médio	F_0
Regressão	SQR	p	$QMR = \frac{SQR}{P}$	$\frac{QMR}{QME}$
Resíduo	SQE	n-p-1	$QME = \frac{SQE}{N-P-1}$	
Total	SQT	n-1	$QMT = \frac{SQT}{N-1}$	

onde:

SQR = **b'X'Y** - $n\bar{y}^2$, SQE = **Y'Y – b'X'Y** e SQT = SQE + SQR

A existência de uma relação de regressão, por si só, não garante que predições úteis podem ser feitas usando este modelo (Neter *et al.*, 1996, p.230). Este teste é apenas a primeira etapa na verificação de aceitação do modelo.

Teste de hipótese para o parâmetro β_k: Após a verificação de que pelo menos um dos parâmetros β_k é significativo, deve-se testar a significância de cada um deles, isto é, para cada parâmetro β_k (k=1,...,p), testam-se as hipóteses:

$$H_0 : \beta_k = 0$$
$$H_1 : \beta_k \neq 0$$

O teste para cada parâmetro é feito utilizando a estatística t de Student com (n-p-1) graus de liberdade, o desvio padrão amostral $S(b_k)$ e o estimador b_k. A estatística do teste é dada por

$$t^* = \frac{b_K}{S(b_K)}$$

Se $|t^*| \leq t_{(1-\alpha/2;n-p-1)}$, o teste não rejeita H_0; caso contrário o teste rejeita H_0 em favor de H_1. A rejeição de H_0 indica uma contribuição significativa da variável independente X_k no modelo.

Teste de hipótese para um subconjunto de parâmetros: Após o teste t sugerir as variáveis independentes a serem usadas na equação, é importante examinar se a variável dependente pode ser explicada pelas variáveis sugeridas tão adequadamente quanto por todas as variáveis. Para isto, testam-se as hipóteses:

$$H_0 : \beta_{q+1} = \beta_{q+2} = \ldots = \beta_{p-1} = 0,\ q<p$$
$$H_1 : \beta_k \neq 0, \text{ para algum } k=q+1, \ldots, p$$

onde q representa os coeficientes não usados na equação.

O teste é feito utilizando a estatística F com (p-q,n-p-1) graus de liberdade. A estatística do teste é dada por

$$F = \frac{(R_p^2 - R_q^2)/(p-q)}{(1-R_p^2)/(n-p-1)}.$$

Onde R_p^2 é o coeficiente de determinação obtido com o modelo cheio, com todas as p variáveis independentes e R_q^2 é o coeficiente de determinação obtido quando o modelo é ajustado para q variáveis.

Se $F \leq F_{(p-q;n-p-1)}$, o teste não rejeita H_0; caso contrário o teste rejeita H_0 em favor de H_1. A aceitação de H_0 indica que a variação da variável dependente é tão adequadamente explicada como o conjunto de todas as variáveis independentes (Chatterjee e Price, 1977, p.65).

3.4 - PODER DE EXPLICAÇÃO DO MODELO

O coeficiente de determinação, R^2, mede o quanto a variabilidade total dos dados é explicada pelo modelo de regressão. Quanto maior

R^2, mais a variação total de Y é reduzida pela introdução das variáveis preditoras. O coeficiente R^2 é dado por

$$R^2 = \frac{SQR_{eg}}{SQT} = 1 - \frac{SQE}{SQT}$$

Para a regressão linear múltipla, o coeficiente de determinação R^2 tende a aumentar à medida que mais variáveis regressoras são adicionadas no modelo. Este fato leva a um coeficiente que não mede mais a real explicação da variável independente Y.

Quando se deseja comparar diferentes modelos, muitos autores preferem usar o chamado coeficiente de determinação ajustado, com um ajuste realizado para os correspondentes graus de liberdade de SQE (soma do quadrado do erro) e SQT (soma do quadrado total), como definido abaixo (Draper e Smith, 1981, p.92):

$$R_a^2 = 1 - (1 - R^2)\left(\frac{n-1}{n-p}\right)$$

3.5 - RELACIONAMENTO ENTRE VARIÁVEIS

O estudo do relacionamento entre um conjunto de variáveis pode ser realizado aplicando diversas técnicas, desde os coeficientes de correlação de Pearson, de Spearman, de Kendall, até a chamada Análise Fatorial e a Análise de Regressão.

3.5.1 - Correlação

O coeficiente de correlação linear de Pearson é uma medida usada para estudo da relação linear existente entre duas variáveis X e Y, dada por:

$$r_{XY} = \frac{\sum (X_i - \overline{X})(Y_i - \overline{Y})}{\left[\sum (X_i - \overline{X})^2 \sum (Y_i - \overline{Y})^2\right]^{1/2}}$$

Este valor sempre está entre –1 e 1. Quanto mais próximo de 1 e –1 maior é a tendência de relação linear positiva e negativa respectivamente; e quando estiver próximo de zero indica ausência de correlação linear entre as variáveis (Dantas, 1998, p.114).

A matriz das correlações entre as variáveis independentes pode ser utilizada para orientar os primeiros passos para a construção da equação de regressão. Um valor do determinante da matriz das correlações das variáveis independentes próximo de zero é indicação de multicolinearidade (Dantas, 1998, p.132,133).

A regressão e a correlação estão intimamente relacionadas, porém são muito diferentes conceitualmente. O coeficiente de correlação mede a intensidade da associação linear entre duas variáveis aleatórias, enquanto a regressão tenta estimar ou prever o valor médio de uma variável aleatória com base nos valores fixados de outras variáveis fixadas. A análise de correlação trata duas variáveis simetricamente, não distingue a variável dependente e independente e supõe as duas aleatórias. Na regressão há uma assimetria em como as variáveis dependente e independente são tratadas. A variável dependente é suposta ser estatística, aleatória ou estocástica, isto é, ter uma distribuição de probabilidade (Gujarati, 2000, p.9).

A regressão mostra como as variáveis estão relacionadas e a correlação mostra o grau de relacionamento entre elas. O número fornecido pela correlação é um retrato instantâneo de quão próximo estão duas variáveis que variam juntas. Alguns economistas consideram a correlação uma técnica pouco poderosa, porém como a correlação e a regressão estão intimamente ligadas matematicamente, muitas vezes a correlação é um auxílio útil na regressão (Wonnacott e Wonnacott, 1978, p.98-102).

3.5.2 - Análise Fatorial de Correspondências

A Análise Fatorial de Correspondências é uma das técnicas multivariadas que permite examinar relações geométricas do cruzamento ou contingenciamento de variáveis categóricas, analisando a distribuição de massa de um conjunto de observações, tendo como princípios básicos, a proximidade geométrica e a redução de dimensionalidade. Não é um método para prova de hipóteses, mas sim uma técnica descritiva e explanatória, ou seja, nenhum teste de significado estatístico é costumeiramente aplicado aos resultados de uma análise fatorial de correspondência, que tem como princípio reproduzir, de forma simplificada, as informações de uma grande tabela de freqüência. Esta técnica permite estudar uma população de indivíduos descrita por variáveis que podem ser do tipo qualitativo, quantitativo ou uma mescla de ambos, desde que os dados contínuos sejam discretizados e restritos a valores positivos (Pereira, 2001, p.133).

A transformação de variáveis contínuas em qualitativas tem a finalidade de tornar homogêneo os conjuntos de dados que são compostos de variáveis numéricas e de variáveis qualitativas. Ainda se pode ter interesse em realizar uma codificação qualitativa até quando se dispõe de um conjunto de variáveis numéricas, sobre o qual se pode aplicar adequadamente a chamada análise de componentes principais. Uma análise fatorial de correspondência múltipla sobre as mesmas variáveis codificadas em classes, dá outra aproximação para os dados, pois permite exibir, possíveis relações não lineares entre as variáveis. Tais fenômenos são invisíveis nos resultados de uma análise de componentes principais, que não leva em conta, mais que relações lineares (Escofier e Pagès, 1992, p.7-25).

Para codificar uma variável contínua em classes, ou seja, recortar o seu intervalo de variação em subintervalos que definem outras modalidades, é necessário determinar o número de intervalos e seus limites. Diminuindo o excesso do número de classes, se agrupam indivíduos cada vez mais distintos e, por isto, perde-se muita

informação. Por outro lado, aumentando o número de casos corre–se o risco de se obter classes com pouca informação. A experiência mostra que não é útil superar o número de oito modalidades na codificação de variáveis quantitativas e que quatro ou cinco são suficientes (Escofier e Pagès, 1992, p.66-69).

3.6 - TRANSFORMAÇÕES DE VARIÁVEIS

As variáveis são um conjunto de medidas repetidas de um determinado objeto de estudo sendo que estas medidas podem ser realizadas em diferentes unidades, que levam a classificá-las como quantitativas ou qualitativas (Pereira, 2001, p.43,44).

O primeiro aspecto a ser analisado é com relação aos dados, estudando o tipo, comparando as grandezas e o comportamento entre as variáveis. Então deve-se preparar estes dados para a análise de regressão.

Algumas vezes se faz necessário algum tipo de transformação nas variáveis e também o uso de variáveis “dummy”. As transformações são necessárias quando ocorre a falta de linearidade, variância não-constante dos erros e não-normalidade dos erros; possibilitando a construção de modelos mais simples. Por razões práticas, modelos mais simples são mais fáceis de se estudar a validade e também de serem testados.

Geralmente transformações simples da variável dependente, das variáveis independentes ou de ambas possibilitam a construção de um modelo de regressão linear apropriado ao conjunto de dados transformados.

Quando se procura ajustar modelos a dados imobiliários, a transformação logarítmica é a preferida, pois as variáveis pertinentes a imóveis pertencem ao campo dos números reais positivos e os valores transformados também serão, assim a especificação logarítmica se adequa melhor na descrição dos preços das unidades em relação a seus respectivos atributos (Macedo, 1998; Dantas, 1998, p.143).

Outro aspecto é que a própria ABNT sugere a transformação logarítmica na variável resposta porque torna o modelo aditivo.

No entanto existem muitas possibilidades de transformações, e muitos modelos podem ser descritos. A escolha de qual a melhor transformação a se fazer exige um conhecimento do comportamento das variáveis em estudo. A razão principal para a realização das transformações é poder fazer uso de um modelo de regressão na forma mais simples em vez de uma forma mais complicada obtido com as variáveis originais (Draper e Smith, 1981, p.221).

3.6.1 - Linearidade

A linearidade ocorre quando os pontos permitem um ajuste através de um hiperplano. Isto pode ser investigado através do coeficiente de correlação calculado entre a variável dependente e cada variável independente, e ainda analisando os seguintes gráficos:

variável dependente versus variável independente: havendo várias variáveis regressoras, é recomendado o gráfico com cada uma delas, devendo-se observar se os pontos estão alinhados, conforme mostra a Figura 3.1.

resíduos versus variável independente: devendo observar se os pontos estão dispostos aleatoriamente, sem quaisquer tipos de tendência, conforme Figura 3.2, devendo também ser construído para cada variável independente.

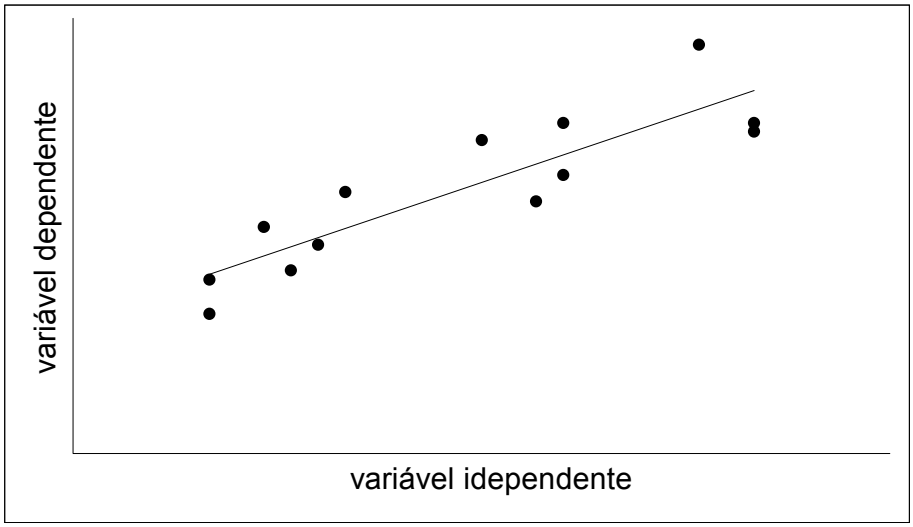

Figura 3.1: exemplo de relação linear pela plotagem da variável dependente versus variável independente.
Dados fictícios.

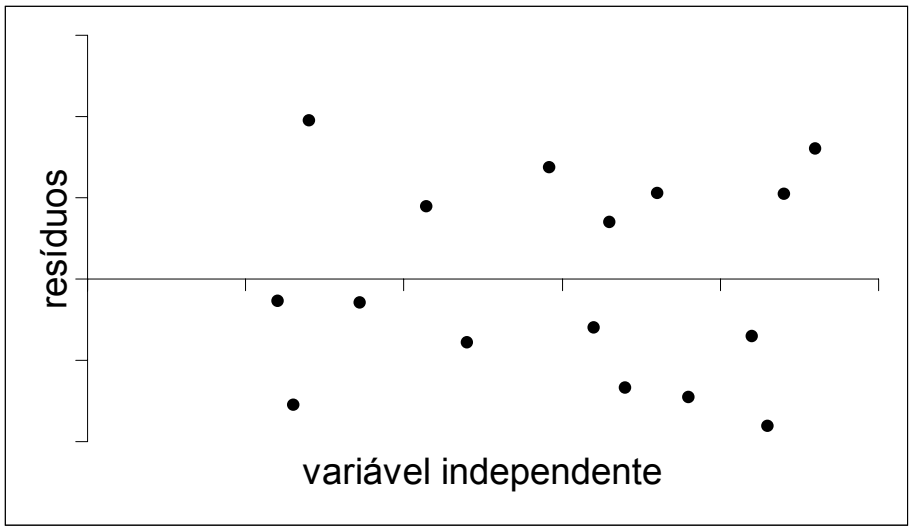

Figura 3.2: exemplo de relação linear pela plotagem dos resíduos versus variável independente.
Dados fictícios.

Quando ocorre a falta de linearidade, sendo a distribuição dos erros aproximadamente normal e com variância razoavelmente constante, transformações na variável independente devem ser testadas. Isto porque transformações na variável dependente, tal como a transformação por raiz quadrada, pode mudar o tipo de distribuição e

provocar uma diferença nas variâncias dos erros (Neter et al, 1996, p.126-128).

As transformações para uma variável independente X para a não-linearidade do modelo, dentre outras transformações, podem ser:

- Logarítmica, X' = logX (na base 10 ou e)
- Raiz quadrada, X' = $\sqrt{X}$
- Quadrática, X' = X^2
- Exponencial, X' = exp(X)
- Recíproca, X' = 1/X
- Exponencial negativa, X' = Exp(−X)

3.6.2 - Variância não-constante e não-normalidade dos erros

Quando se comprova a heterogeneidade da variância, as estimativas das variâncias dos estimadores dos parâmetros são tendenciosos, levando à valores incorretos das inferências. Neste caso há necessidade de estabilizar a variância, que pode ser feito através de transformações na variável resposta.

A suposição de normalidade dos erros deve ser satisfeita para que se possa calcular os intervalos de confiança e fazer inferências.

Uma variância não-constante e a não-normalidade dos erros aparecem freqüentemente ao mesmo tempo, e alguma transformação na variável resposta deve ser feita. É claro que uma transformação na variável resposta também pode resolver o problema de falta de linearidade de uma relação de regressão. Outras vezes, uma transformação simultânea da variável resposta e da variável preditora será necessária para obter uma relação de regressão linear (Neter *et al.*, 1996, p.129-132).

Exemplos de transformações da variável resposta Y para a variância não-constante e a não-normalidade dos erros são:

- Logarítmica, Y' = logY (na base 10 ou e)
- Raiz quadrada, Y' = $\sqrt{Y}$
- Quadrática, Y' = Y^2
- Recíproca, Y' = 1/Y
- Arco seno, Y' = arcsen $\sqrt{Y}$
- Box-Cox, Y'=Y^{λ}.

O procedimento de Box-Cox identifica automaticamente uma transformação da família de transformações de potência sobre Y. O parâmetro λ é determinado dos dados. Para λ=0 define-se a transformação logarítmica (Neter et al, 1996, p.129-132).

3.6.3 - Variáveis "Dummy"

As variáveis usadas nas equações de regressão podem não serem quantitativas. Isto geralmente ocorre quando se estuda o comportamento do mercado imobiliário, que além de ser caracterizado por variáveis quantitativas – como valores de venda ou locação e idade – é também caracterizado por variáveis qualitativas – como presença de elevador (sim ou não), situado na região A ou B ou C e padrão de acabamento (ótimo, bom, regular ou ruim) dentre outras. Estas variáveis são comumente chamadas de "variáveis dummy" (Dantas, 1998, p.157,158).

Para as variáveis que apresentam um aspecto dicotômico, como por exemplo a presença de elevador (sim ou não), relaciona-se determinado número as características. Geralmente usa-se o número *um* quando determinada característica está presente e o número *zero* em caso contrário (Moreira Filho, 1993, p.85-87).

As variáveis com mais de dois níveis, como por exemplo as regiões A, B ou C, podem ser trabalhadas como duas variáveis dicotômicas, sendo região A (sim ou não) e região B (sim ou não), estando a região C contemplada na análise quando se tem não para a região A e não

para a região B. Generalizando, tem-se que uma variável com p níveis pode ser reescrita como (p-1) variáveis dicotômicas.

Quando se faz uma regressão da variável dependente sobre as variáveis independentes “dummy”, os coeficientes de mínimos quadrados das variáveis “dummy” são as médias das celas em que estão tabulados (Johnston, 1974, p.239-241).

3.7 - MULTICOLINEARIDADE

Em análise de regressão linear múltipla, existe um freqüente interesse com relação a natureza e significância das relações entre as variáveis independentes e a variável dependente. Em muitas aplicações de administração e economia, freqüêntemente encontram-se variáveis independentes que estão correlacionadas entre elas mesmas e, também, com outras variáveis que não estão incluídas no modelo, mas estão relacionadas à variável dependente (Neter e Wasserman, 1974, p.339).

Define-se como multicolinearidade a existência de relações lineares entre as variáveis independentes. Quando a relação é exata tem-se o caso da multicolinearidade perfeita.

Na prática atual, raramente, encontramos variáveis independentes que são perfeitamente relacionadas. Este caso não traz problemas, pois é facilmente detectado e pode ser resolvido simplesmente eliminando uma ou mais variáveis independentes do modelo.

O interesse no que se refere a multicolinearide está nos casos em que ela ocorre com alto grau, isto é, quando duas variáveis independentes estão altamente correlacionadas ou quando há uma combinação quase linear entre um conjunto de variáveis independentes. Assim, a multicolinearidade é mais uma questão de grau do que de natureza (Kmenta, 1978, p.411-423).

O fato de muitas funções de regressão diferentes proporcionarem bons ajustes para um mesmo conjunto de dados é porque os coeficientes de regressão atendem várias amostras onde as variáveis

independentes são altamente correlacionadas. Assim, os coeficientes de regressão estimados variam de uma amostra para outra quando as variáveis independentes estão altamente correlacionadas. Isto leva a informações imprecisas a respeito dos coeficientes verdadeiros (Neter e Wasserman, 1974, p.344).

A multicolinearidade geralmente é causada pela própria natureza dos dados, principalmente nas áreas de economia com variáveis que representam valores de mercado. Algumas vezes a multicolinearidade pode também ocorrer devido a amostragem inadequada (Elian, 1998).

3.7.1 - Efeitos da Multicolinearidade

Efeito da Multicolinearidade nos Coeficientes de Regressão: quando as variáveis independentes são correlacionadas, o coeficiente de regressão de alguma variável independente depende de qual outra variável independente é incluída no modelo, pois adicionando ou deletando uma das variáveis independentes mudam-se os coeficientes de regressão. Assim, um coeficiente de regressão deixa de refletir os efeitos inerentes de uma particular variável independente sobre a variável dependente, mas reflete apenas um efeito parcial.

Note que o coeficiente de regressão de uma variável independente X_1 é inalterado quando uma variável independente X_2, não correlacionada com X_1, é adicionada no modelo de regressão, pois,

$$b_1 = \frac{\dfrac{\sum(x_{i1}-\bar{x}1)(y_i-\bar{y})}{\sum(x_{i1}-\bar{x}1)^2} - \left[\dfrac{\sum(y_i-\bar{y})^2}{\sum(x_{i1}-\bar{x}1)^2}\right]^{\frac{1}{2}} .r_{y2}.r_{12}}{1-r_{12}^2}$$

onde r_{y2} é o coeficiente de correlação entre as variáveis Y e X_2 e r_{12} é o coeficiente de correlação entre as variáveis X_1 e X_2.

Se X_1 e X_2 forem não correlacionadas, tem-se r_{12} =0, e, portanto, $b_1 = \dfrac{\sum(x_{i1}-\bar{x}1)(y_i-\bar{y})}{\sum(x_{i1}-\bar{x}1)^2}$, que é o coeficiente de X_1 na regressão simples de Y

sobre X_1. Logo, se X_1 e X_2 não são correlacionados, a adição de X_2 no modelo de regressão não muda o coeficiente de X_1.

Efeito da Multicolinearidade na Soma de Quadrados de Regressão: quando variáveis independentes são correlacionadas, não existe uma soma de quadrados única que pode ser atribuída a uma variável independente refletindo o efeito na redução da variação total em Y, ou seja, a soma de quadrado associada a uma variável independente varia, dependendo sobre que variável independente esteja incluída no modelo.

Efeito da Multicolinearidade nos Testes para os Coeficientes de Regressão: um abuso freqüênte nos modelos de regressão é observado ao se examinar a estatística t* para cada coeficiente de regressão. É possível que quando um conjunto de variáveis independentes esteja relacionado à variável dependente, obtendo todos os testes individuais sob os coeficientes de regressão, eles levarão à conclusão que os coeficientes são iguais a zero devido a multicolinearidade entre as variáveis independentes. Os coeficientes de regressão estimados individualmente podem não ser estatisticamente significativos, ainda que possa existir uma relação estatística entre a variável dependente e o conjunto de variáveis independentes (Matos, 2000, p.124-129).

Apesar das conseqüências dos efeitos da multicolinearidade, citados anteriormente, a alta Multicolinearidade não é, geralmente, um problema quando o propósito da análise de regressão é fazer inferências sobre a função resposta ou predições de novas observações, contanto que estas inferências sejam feitas dentro do âmbito das observações (Neter e Wasserman, 1974, p.345; Neter *et al.*, 1996, p.285,295).

Para fazer boas previsões o pesquisador deve ter confiança de que o caráter e o tamanho do relacionamento global se manterá de período para período. Note que, a questão de confiança é um problema

existente em todos os modelos de previsões, com ou sem a presença de multicolineariade (Chatterjee e Price, 1977, p.151-153).

Quando o relacionamento entre as variáveis independentes se mantém, no período previsto e no período da amostra, as previsões são corretas até mesmo fora da amostra (Judge *et al.*, 1988, p.859-861).

Nota-se, então, que quando os valores das variáveis independentes, para as quais se desejam as previsões, obedecem as mesmas dependências da matriz original, a multicolinearidade não é um problema. Como nas aplicações econômicas a multicolinearidade, quando existe, é uma característica da população, a estrutura permanece de amostra para amostra, não sendo um caso de uma amostra simplesmente infectada pela Multicolinearidade. Portanto, pode-se usar a equação de regressão estimada para fazer inferências sobre a função resposta ou predições de novas observações (Neter *et al.*, 1996, p.285-295).

3.7.2 - Detectando a Multicolinearidade

Existem muitas sugestões, ou métodos propostos, para detectar a multicolinearidade. Os mais comumente usados são:

coeficiente de correlação simples: é uma medida comumente usada no caso de duas variáveis independentes, sendo suficiente para detectar a colinearidade. Considera-se que um coeficiente de correlação maior que 0,80 ou 0,90 é indicativo de um problema sério de colinearidade. Porém, para mais de duas variáveis independentes, mesmo os coeficientes de correlação sendo baixos ainda pode existir a multicolinearidade, pois pares de correlações podem não dar visão de intercorrelacionamentos mais complexos entre três ou mais variáveis (Judge *et al.*, 1980, p.458,459).

determinante de (X'X): Se as variáveis independentes estão padronizadas, tal que (**X'X**) contém elementos que são os coeficientes

de correlação linear entre as variáveis independentes, então o determinante de (**X'X**) é um valor no intervalo [0;1]. Caso as variáveis independentes não estejam padronizadas, é melhor analisar o determinante da matriz de correlações entre as variáveis independentes, $\mathbf{R}_x$, que sempre assume valores no intervalo [0,1]. Um valor deste determinante próximo de zero é indicativo de multicolinearidade (Elian, 1998, p.125; Neter e Wasserman, 1974, p.347).

coeficiente de explicação do modelo, R^2: o R^2 tendo um valor alto, mas os coeficientes de correlação parcial tendo valores baixos, tem-se a indicação de multicolinearidade.

regressões auxiliares: se o valor de R^2, calculado da regressão de cada variável independente sobre as outras (k-1) variáveis independentes é alto, então há indicativo de multicolinearidade.

raízes características: Sejam λ_i , i=1,...,p, as raízes características de $\mathbf{R}_x$, tem-se que $\det(\mathbf{R}_x)=\Pi\lambda_i$. Baixos valores de uma ou mais raízes características, comparado com o maior valor, são indicativos de multicolinearidade. Um critério, bem eficiente, na quantificação da multicolinearidade, é a análise do valor de L, dado por

$$L=\lambda_{máx}/\lambda_{mín},$$

onde $\lambda_{máx}$ é o maior valor das raízes características e $\lambda_{mín}$ é o menor valor das raízes características. Se L < 100, considera-se não existir multicolinearidade, se 100 < L < 1000 existe multicolinearidade moderada e se L > 1000 há indicativo de multicolinearidade séria (Elian, 1998, p.131).

O gráfico dos resíduos: o gráfico dos resíduos versus cada variável independente, inclusive as variáveis que não fazem parte da equação de regressão, indicam a inexistência de correlacionamento quando os

pontos estiverem dispostos aleatoriamente, sem qualquer padrão. Caso contrário, se apresentarem algum tipo de tendência, há indicativo de correlacionamento.

3.7.3 - Soluções para o Problema de Multicolinearidade

A existência de multicolinearidade tendo sido detectada e considerada prejudicial, indica que o pesquisador deve procurar soluções para suavizar seus efeitos ruins. Várias medidas corretivas têm sido propostas, desde simples às mais complexas, para suavizar os efeitos provocados pela multicolinearidade (Elian,1988, p.131-134; Judge *et al*., 1980, p.464-468).

Remoção de variáveis - uma medida simples é remover uma ou várias variáveis independentes, pouco importantes no contexto geral, que venham a diminuir a multicolinearidade. Porém, esta ação não ajuda a avaliar os efeitos da variável independente, pois nenhuma informação é obtida à cerca da variável removida, e também porque o valor do coeficiente de regressão para a variável independente remanescente no modelo é afetada pelas variáveis independentes correlacionadas não incluídas no modelo.

Ampliação do tamanho da amostra - algumas vezes é possível adicionar algumas observações na amostra que elimina o padrão de multicolinearidade. Esta medida é usada quando o problema é causado por informação amostral inadequada. Porém, em administração e economia muitas variáveis independentes não podem ser controladas de forma que novas observações tenderão a mostrar o mesmo padrão de intercorrelação.

***Ridge Regression* -** outro critério, para o qual tem-se dado atenção, é *Ridge Regression* (Regressão em Cumeeira) que consiste no uso de

estimadores tendenciosos para os coeficientes. O estimador em crista é na verdade uma família de estimadores dados por:

$$\mathbf{b}(k)=(\mathbf{X'X}+k\mathbf{I})^{-1}\mathbf{X'Y},$$

onde k é um valor pequeno que deve ser escolhido a critério do pesquisador. Em geral, aumenta-se gradativamente o valor de k até que os estimadores dos coeficientes tornam-se estáveis, não variam. Se a escolha for k=0, tem-se o estimador de mínimos quadrados (Neter *et al.*, 1996, p.411-416; Elian, 1998, p.133; Draper e Smith, 1981, p.313-349).

Na presença de multicolinearidade sempre existe um valor de k para o qual os estimadores de *Ridge Regression* produzem um QME (quadrado médio do erro) menor do que o QME produzido pelos estimadores de mínimos quadrados ordinários. A dificuldade desta questão é que o valor de k varia de uma aplicação para outra e é desconhecido. Assim, embora exista este valor de k, não existe um caminho conhecido para obtê-lo, mesmo quando obtêm-se um k que produza um MSE menor do que o MSE de mínimos quadrados ordinários para um problema prático específico (Draper e Smith, 1981, p.315,316).

A função de regressão estimada pela *Ridge Regression* produz predições de novas observações que tendem a serem mais precisas do que as predições feitas pela função de regressão estimada pelo método de mínimos quadrados, quando as variáveis independentes são correlacionadas e a nova observação segue o mesmo padrão de multicolinearidade. Esta precisão na predição de novas observações é favorecida pela *Ridge Regression*, especialmente quando a multicolinearidade é forte (Neter et al., 1996, p.411-416).

Componentes principais – uma outra forma que pode ser utilizada para tratar o problema causado pela multicolinearidade é a técnica de componentes principais.

Esta técnica permite que todas as variáveis independentes participam de certa forma do modelo. Através desta técnica, é possível

reduzir um grande número de variáveis independentes em um número razoavelmente pequeno de novas variáveis independentes, que são chamadas de componentes e são determinadas pela combinação linear das variáveis originais. Estas novas variáveis (ou componentes principais) são não correlacionadas e são usadas para determinar o modelo de regressão.

O objetivo da análise de componentes principais é representar ou descrever um conjunto de variáveis por um outro conjunto menor de novas variáveis, sem perda significativa da informação original (Reis, 1997, p.255). A redução da dimensionalidade das variáveis consiste no fato de que as primeiras componentes principais possam explicar a maior parte da variabilidade total dos dados originais.

A análise de componentes principais permite um estudo detalhado da importância de cada variável, fornecendo a quantidade de explicação na componente principal e seu relacionamento com as demais variáveis. Sobre a análise de componentes principais pode ser visto em Escofier e Pagès (1990) e Bouroche e Saporta (1982).

3.8 - SELEÇÃO DE VARIÁVEIS REGRESSORAS

Um dos problemas mais difíceis e freqüêntes em análise de regressão é a seleção do conjunto de variáveis independentes para serem incluídas no modelo (Neter e Wasserman, 1974, p.371).

O investigador deve especificar o conjunto de variáveis independentes a ser empregado para descrever, controlar ou predizer a variável dependente.

Um problema muito difícil de relacionamento que aparece na seleção de variáveis é quando uma equação de regressão é construída com o objetivo de predição e envolve muitas variáveis. Talvez, muitas delas, contribuam pouco ou nada para precisão da predição. A escolha apropriada de algumas delas fornece a melhor predição, porém o problema é quantas e quais variáveis selecionar (Snedecor e Cochram, 1972, p.412,413).

Em alguns campos, a teoria pode ajudar na seleção das variáveis independentes a serem empregadas e na especificação da forma funcional da relação de regressão. Em tais campos, os experimentos podem ser controlados para fornecer dados sobre a base de que os parâmetros de regressão podem ser estimados e a forma teórica da regressão testada. Em muitos outros campos, entretanto, modelos teóricos são raros. Assim, os investigadores são freqüêntemente forçados a explorar as variáveis independentes para que possam realizar estudos sobre a variável dependente. Obviamente, tais conjuntos de variáveis independentes são grandes. Algumas das variáveis independentes podem ser removidas seletivamente. Uma variável independente pode não ser fundamental ao problema; pode estar sujeita a grandes erros de medidas; e pode efetivamente duplicar outra variável independente da lista. Outras variáveis independentes, que não podem ser medidas, podem ser deletadas ou substituídas por variáveis que estão altamente correlacionadas com estas.

Tipicamente, o número de variáveis independentes que permanece, após esta seleção inicial, ainda é grande. Posteriormente, muitas destas variáveis estarão altamente intercorrelacionadas. Portanto, o investigador geralmente desejará reduzir o número de variáveis independentes a serem usadas no modelo final. Existem várias razões para isto. Um modelo de regressão com um número grande de variáveis independentes é caro para se utilizar. Dessa forma, modelos de regressão com um número limitado de variáveis independentes são fáceis para se avaliar e estudar. Finalmente, a presença de muitas variáveis independentes altamente intercorrelacionadas, podem adicionar pouco ao poder de predição do modelo, enquanto retira suas habilidades descritivas e aumenta os erros de predição.

O problema então é como reduzir a lista de variáveis independentes de forma a obter a melhor seleção de variáveis independentes. Este conjunto precisa ser suficientemente pequeno para que a manutenção dos custos de atualização do modelo sejam manuseáveis e a análise facilitada, e ainda, deve ser grande o suficiente de forma que seja possível uma descrição, um controle e uma predição adequados.

Os procedimentos de procura para se encontrar o melhor conjunto de variáveis independentes devem ser empregados após o investigador ter estabelecido a forma funcional da relação de regressão, ou seja, se as variáveis dadas estão na forma linear, quadrática, etc; se as variáveis independentes são primeiramente transformadas, como por exemplo por transformação logarítmica; e se algum termo de interação foi incluído. Neste ponto, os procedimentos de procura são empregados para reduzir o número de variáveis independentes.

Existem muitos procedimentos de seleção, mas nenhum deles pode, comprovadamente, produzir o melhor conjunto de variáveis independentes. Não existe um conjunto ótimo de variáveis independentes, pois o processo de seleção das variáveis possui julgamentos subjetivos. Dentre os procedimentos, pode-se citar como os mais comumente usados: todas as regressões possíveis, backward, forward e stepwise.

Todas as regressões possíveis: este procedimento consiste em ajustar todas as possíveis equações de regressão. Após a obtenção de todas as regressões, deve-se utilizar os critérios para comparação dos modelos ajustados. Alguns critérios que podem ser usados são o R^2 (coeficiente de explicação), MSE (quadrado médio dos resíduos) e C_p (estatística de Mallows).

Para alguns conjuntos de variáveis, os três critérios podem levar para o mesmo “melhor” conjunto de variáveis independentes. Este não é o caso geral, pois diferentes critérios podem sugerir diferentes conjuntos de variáveis independentes. Daniel e Wood (1971, p.86) recomendam, no caso de um grande número de equações alternativas, o critério do erro quadrado total para caracterizar a equação. A principal desvantagem do procedimento de procura de todas as regressões possíveis é a quantidade de esforço computacional necessária. Já que cada variável independente potencial pode ser incluída ou excluída, gerando (2P - 1) regressões possíveis quando existem p variáveis independentes potenciais (Elian, 1998, p.139; Draper e Smith, 1981, p.296).

***stepwise* (passo a passo):** é, provavelmente, o mais amplamente usado dos métodos de pesquisa que não requerem a computação de todas as regressões possíveis. Ele foi desenvolvido para economizar esforços computacionais, quando comparado com a abordagem de todas as regressões possíveis, enquanto atinge um conjunto de variáveis independentes razoavelmente bom. Essencialmente, este método de pesquisa computa uma seqüência de equações de regressão, adicionando ou deletando uma variável independente em cada passo. A rotina de regressão *stepwise* permite que uma variável independente, trazida para dentro do modelo em um estágio anterior, seja removida subseqüentemente se ela não ajudar na conjunção com variáveis adicionadas nos últimos estágios. Esta rotina empregada, conduz a um teste para rastrear alguma variável independente que seja altamente correlacionada com variáveis independentes já incluídas no modelo. A limitação da procura da regressão *stepwise* é que ela presume a existência de um único conjunto ótimo de variáveis independentes e busca identificá-lo. Como notado anteriormente, não existe freqüêntemente um único conjunto ótimo. Outra limitação da rotina de regressão *stepwise*, é que ela algumas vezes surge com um conjunto de variáveis independentes razoavelmente fraco para predições, quando as variáveis independentes estão altamente correlacionadas (Draper e Smith, 1981, p.307-312).

Seleção f*orward*: este procedimento de procura é uma versão simplificada da regressão *stepwise*, omitindo o teste, se uma variável uma vez que tenha entrado no modelo deva ser retirada. Este procedimento considera, inicialmente, um modelo simples usando a variável de maior coeficiente de correlação com a variável dependente. Uma variável por vez é incorporada até que não haja mais inclusão, e as variáveis selecionadas definem o modelo.

Eliminação b*ackward*: este procedimento de procura é oposto à seleção *forward*. Ele começa com o modelo contendo todas as variáveis independentes potenciais. O procedimento de eliminação *backward*

requer mais computações que o método de seleção *forward*, já que ela começa com o maior modelo possível. Entretanto, ela tem uma vantagem de mostrar ao analista as implicações do modelo com muitas variáveis.

3.9 – CUIDADOS NO USO DO MODELO

É claro que o rastreamento de variáveis por um processo de seleção computadorizado é somente um passo na construção de um modelo de regressão. Uma vez que o conjunto de variáveis independentes tem sido identificado o modelo resultante necessita ser estudado.

Um teste formal para a falta de ajuste pode ser feito e a plotagem dos resíduos deve ser empregada para identificar a natureza da falta de ajuste, pontos discrepantes, e outras deficiências.

O processo de construção de um modelo de regressão requer repetidas análises sobre o conjunto de dados para checar se o modelo ajusta-se bem os dados. Um modelo ruim pode levar a predições errôneas. Este fato pode ocorrer devido a má escolha das variáveis independentes. Uma forma de medir as predições tendenciosas é deixar alguns dos dados originais fora dos cálculos para determinação do modelo e usá-los para realizar o ajustamento do poder preditivo do modelo (Neter e Wasserman, 1974, p.388).

3.10 - RESÍDUOS

Os resíduos são definidos como as n diferenças ei = Yi - $\hat{Y}_i$, i=1,...,n onde Yi é uma observação e $\hat{Y}_i$ é o valor ajustado. Observe que os resíduos ei são as diferenças entre o que é observado e o que é predito pela equação de regressão, isto é, a quantidade que a equação de regressão não tem capacidade para explicar. Assim pode-se pensar ei como os erros observados se o modelo está correto, lembrando que

as suposições impostas para erros, citadas na seção 3.1.1 deste capítulo, são: a independência, a média zero, a variância constante e que seguem a distribuição normal. Se o modelo ajustado está correto, os resíduos exibirão tendências que tendem a confirmar as suposições feitas, ou por fim, exibirão uma recusa das suposições (Elian, 1998, p.41,91).

Para examinar os resíduos e checar o modelo são utilizadas as ferramentas gráficas que são fáceis de construir e revelam claramente a validação das suposições. Neste sentido, apresenta-se aqui a metodologia de interpretação dos gráficos dos resíduos.

3.10.1 - Análise de Resíduos

Com a definição do modelo, deve-se realizar a análise dos resíduos, a fim de procurar evidências sobre eventuais violações das suposições de média zero, independência, homocedasticidade e normalidade.

média zero: a verificação de que os erros têm valor esperado zero pode ser investigado pelo gráfico da plotagem dos resíduos versus valores preditos, como mostra a figura 3.3, devendo os pontos estarem distribuídos em torno do zero.

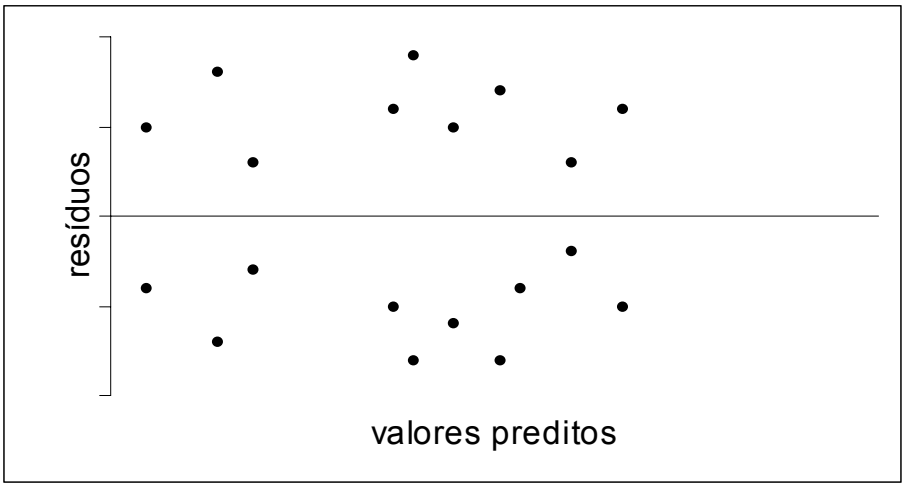

Figura 3.3: Exemplo de média zero.

Dados fictícios.

independência: a inexistência de autocorrelação dos erros pode ser investigada através do gráfico dos resíduos versus sequência no tempo (ou sequência de coleta de dados), pode-se verificar a independência dos resíduos quando eles se distribuem aleatoriamente, em torno de zero, conforme figura 3.4.

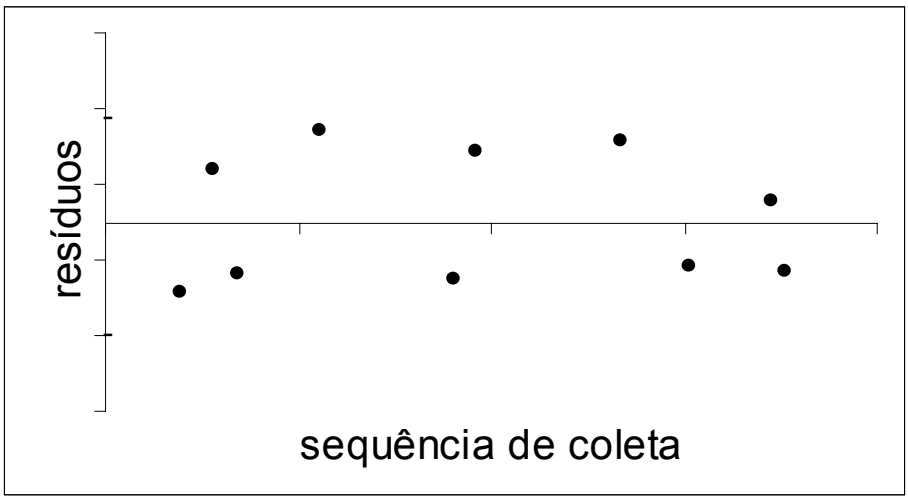

Figura 3.4: Exemplo de independência dos erros pela plotagem dos resíduos.
Dados fictícios.

Uma outra forma de verificação da existência de autocorrelação dos erros pode ser feita pela estatística de Durbin-Watson.

Utilizando-se da razão,

$$d = \frac{\sum_{i=2}^{n} (e_i - e_{i-1})}{\sum_{i=1}^{n} e_i^2}$$

e fazendo-se uso da distribuição (dos valores) tabelada por Durbin-Watson, testa-se a hipótese nula de que os erros são não correlacionados contra a hipótese alternativa de que os erros são correlacionados. O valor de d está estreitamente relacionado com o valor da correlação dos resíduos r. Tem-se que:

$$d \approx 2(1-r).$$

Se o valor de r é próximo de zero, então a estatística de Durbin-Watson é $d \approx 2$, que indica que os erros do modelo são não correlacionados (Hill et al., 1999).

Quando é detectada a autocorrelação dos erros, deve-se considerar a possibilidade que variáveis independentes importantes foram excluídas do modelo de regressão ou, ainda, há indicação de que o modelo é ruim. Isto leva à estimativas não confiáveis ou tendenciosas.

A validade do teste de Durbin-Watson depende dos erros terem média zero e variância constante e ainda as variáveis independentes não serem aleatórias. Isto ocorre quando valores da variável dependente defasadas aparecem como variáveis independentes (Hoffmann e Vieira, 1983, p.251,252).

O problema de autocorrelação entre os resíduos costuma acontecer quando as observações são realizadas ao longo do tempo, que não é o caso usual na amostragem de avaliação de imóveis.

variância constante: a suposição de variância constante ou homogeneidade de variância é verificada facilmente através dos gráficos dos resíduos versus variáveis independentes ou resíduos versus valores preditos. Quando o gráfico produz forma de megafone, implica que a variância não é constante, conforme mostra a figura 3.5, por outro lado, quando estão distribuídos aleatoriamente em torno de uma reta horizontal que passa pela origem, sem qualquer padrão, há indicação de variância constante.

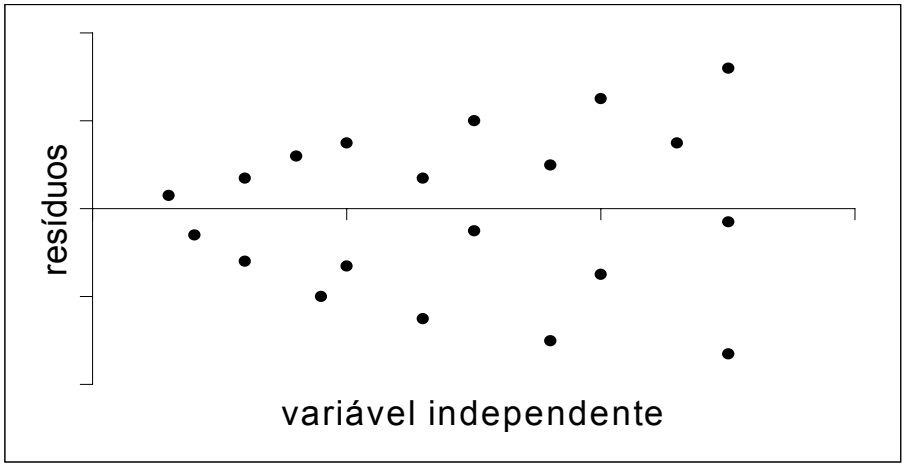

Figura 3.5: Exemplo de variância não-constante pela plotagem dos resíduos versus variável independente.
Dados fictícios.

distribuição normal: o gráfico construído pelos resíduos ordenados versus os respectivos valores teóricos da distribuição normal, mostra a normalidade dos erros quando os pontos se distribuem em torno de uma linha, conforme mostra a figura 3.6.

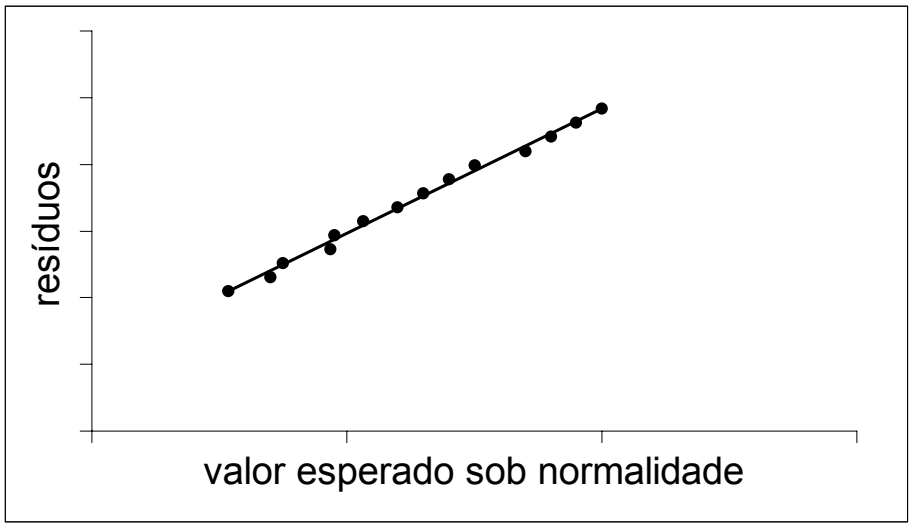

Figura 3.6: Exemplo de distribuição normal dos erros pela plotagem dos resíduos.
Dados fictícios.

Ainda, a normalidade dos erros pode ser investigada, utilizando-se os resíduos padronizados e verificando se aproximadamente 68% deles estão no intervalo [-1 ; 1], 90% no intervalo [-1,64 ; 1,64] e 95% no intervalo [-1,96 ; 1,96]. Pode-se, ainda, utilizar-se do histograma dos resíduos e verificar se este apresenta simetria e forma parecida com o gráfico da curva normal.

3.10.2 - Diagnóstico do Modelo

Através da análise dos gráficos de resíduos, prática esta indispensável, é possível a verificação da adequação do modelo, valores discrepantes e também a investigação da atuação das várias variáveis regressoras disponíveis (Neter et al., 1996, p.97-110).

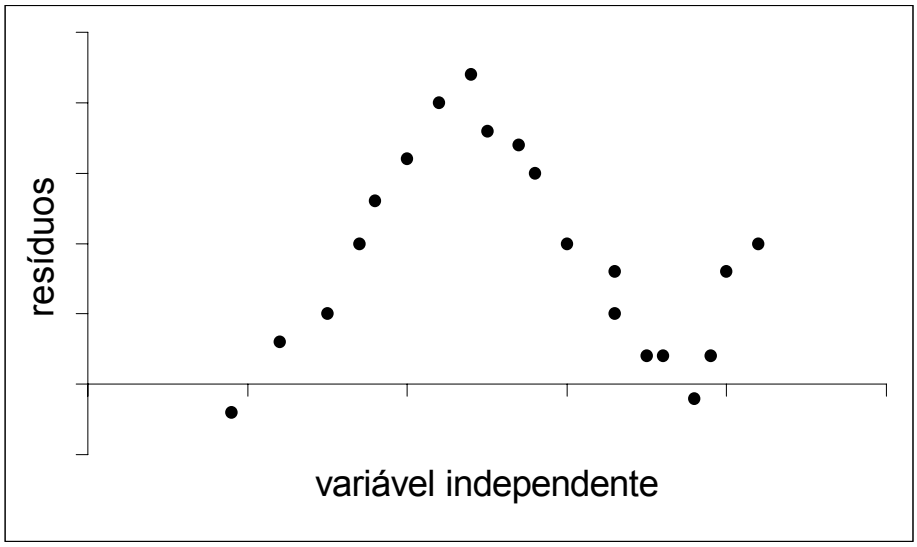

Figura 3.7: exemplo de não adequação do modelo pela plotagem dos resíduos versus variável independente.

Dados fictícios.

adequação do modelo: a verificação de que o modelo de regressão ajustado é adequado aos dados pode ser detectado através dos gráficos dos resíduos versus variáveis independentes ou resíduos versus valores preditos. Quando estes gráficos apresentarem algum tipo de padrão sistemático dos pontos, conforme mostra a figura 3.7, há indicação de que o modelo ajustado não é adequado.

valores discrepantes: os gráficos dos resíduos versus variáveis independentes ou resíduos versus valores ajustados, mostram a presença de pontos discrepantes ou "outliers". Considera-se um ponto discrepante quando é muito maior que o resto em valor absoluto e talvez se estiver afastado de zero por três ou mais desvios padrões, conforme mostra a figura 3.8 (Draper e Smith, 1981, p.152; Neter e Wasserman, 1974, p.106).

Um ponto discrepante pode ser retirado da análise se a causa de sua existência for devido a erro de medida ou de transcrição de valor. Caso contrário deve ser pesquisado e o investigador, conhecedor da natureza das variáveis, é quem deve dar a solução.

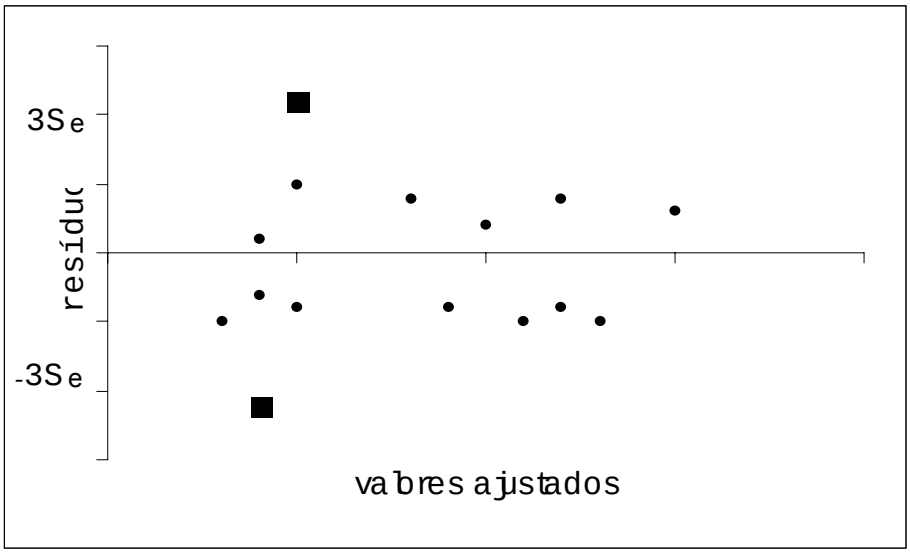

Figura 3.8: exemplo de valores discrepantes pela plotagem dos resíduos versus valores ajustados.
Dados fictícios.

omissão de variáveis independentes: a investigação de variáveis independentes importantes não incluídas no modelo pode ser feita pelo gráfico dos resíduos versus cada uma delas. Considera-se que variável deva ser incluída no modelo quando os pontos apresentarem uma relação linear, conforme a figura 3.9 (Neter *et al.*, 1996, p.109-111).

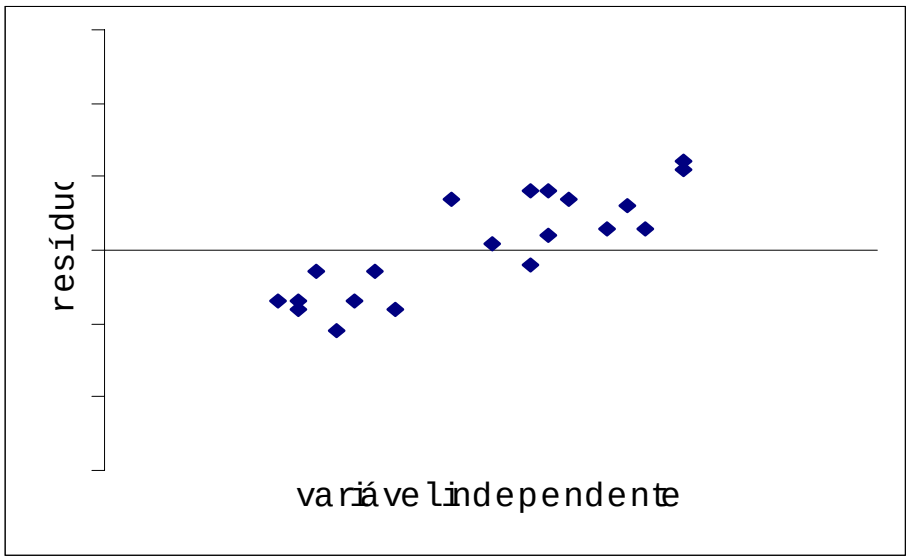

Figura 3.9: exemplo da omissão de variáveis regressoras pela plotagem dos resíduos versus cada variável independente não incluída no modelo.
Dados fictícios.

Nota-se que a análise dos resíduos é de fundamental importância na escolha de um modelo de regressão. Esta análise dos resíduos permite a verificação da validade ou de possíveis violações das suposições do modelo adotado. A validade das suposições é essencial para a obtenção dos estimadores, construção dos intervalos de confiança e testes de hipóteses.

Capítulo 4

ESTRATÉGIA PARA A CONSTRUÇÃO DO MODELO E APLICAÇÃO EM UM ESTUDO DE CASO

4.1 – ROTEIRO PARA A CONSTRUÇÃO DO MODELO

Construir um modelo de regressão linear múltipla para explicar o preço de um imóvel, demanda uma série de etapas a serem realizadas. As etapas, ou roteiro, que são necessárias na construção de um modelo, são apresentados a seguir.

4.1.1 – Identificação das Variáveis Independentes

Um dos aspectos mais importantes na avaliação de imóveis é a seleção das variáveis independentes que possam ser utilizadas na regressão, que são aquelas que tem influência na formação do preço.

As possíveis variáveis independentes (ou explicativas) que podem influenciar no preço de um imóvel, devem ser listadas a priori. A definição das variáveis explicativas preliminarmente, economiza tempo e diminui o custo de execução da pesquisa. As variáveis explicativas para a avaliação de imóveis são aquelas referentes a todas as características físicas e locacionais do imóvel. No entanto, dentre todas as características físicas e locacionais relacionadas a um imóvel, nem todas são relevantes à formação de seu preço.

De forma geral e preliminar, pode-se citar como relevantes à formação do preço, as seguintes variáveis explicativas: área total ou área útil, número de dormitórios, suíte, garagem, dependência de empregada, idade, elevador, estado de conservação, padrão de acabamento, região de valorização imobiliária, distância à escola,

acesso ao transporte urbano, etc. Esta lista de variáveis explicativas tende a variar de município para município, dependendo das características de cada um. Para a cidade de Criciúma – SC, a variável meio ambiente é relevante, pois Criciúma está localizada em área crítica de poluição, devido a exploração de carvão, curtume, cerâmica e metalurgia dentro do perímetro urbano. Em Canasvieiras – SC, município de praia, a variável distância ao mar é indispensável da lista de variáveis explicativas relevantes. Na avaliação do preço de residências no Colorado – Estados Unidas da América, a variável números de lareiras é relevante. Outras variáveis que podem vir a ser relevantes, dependendo da especificidade do município a que pertence a população a ser estudada, são: vista panorâmica, insolação, ventilação, infra-estrutura urbana, tamanho do terreno, etc. É responsabilidade do pesquisador, determinar o conjunto de variáveis relevantes, para que sua pesquisa atinja um resultado desejado.

4.1.2 – Levantamento dos Dados

O levantamento dos dados deve ser feito com muito cuidado, pois dele dependerá o sucesso da análise estatística. É importante a realização de um planejamento para a realização da coleta dos dados. Neste planejamento, o pesquisador deve definir o espaço físico, local onde está inserida a população a ser estudada, e o número de imóveis a serem pesquisados. Na maioria dos estudos é inviável utilizar todos os imóveis da população, isto por diversas razões, entre elas o tempo, o alto custo, a dificuldade na obtenção das informações e outras. Na avaliação de imóveis, geralmente se tem uma população muito grande, o levantamento das informações demanda muito tempo e existem muitos empecilhos, tais como; imóvel desocupado, ausência do morador e outros. Assim, na avaliação de imóveis deve-se usar o levantamento de dados por amostragem. Para garantir a validade da análise, o número de imóveis para compor a amostra deve ser fixado com base na precisão desejada nos resultados a serem obtidos.

A composição da amostra, ou seja, a seleção dos elementos que serão observados, deve ser feita sob uma metodologia adequada, de tal forma que os resultados da amostra sejam informativos para avaliar características de toda a população. Desta forma, na avaliação de imóveis, o que se deseja é uma amostra representativa de dados de mercado de imóveis com características semelhantes aos elementos da população.

No mercado de imóveis, é frequente a entrada de dados novos, por isso, deve se fazer um novo levantamento a cada nova avaliação para garantir a representação dos novos dados na amostra (Dantas, 1998, p.50).

4.1.3 – Transformações de Variáveis

As variáveis que são definidas em função das características e da localização de um imóvel, são do tipo quantitativas e qualitativas. Geralmente, estas variáveis necessitam de transformações para que possam ser realizadas as análises.

As variáveis qualitativas devem ser quantificadas através de uma codificação adequada. Em muitas situações são atribuídos apenas duas situações, tais como: pertence ao pólo de valorização A, existência de dependência de empregada, etc. Nestes casos atribui-se o valor 0 (zero) quando não tem a característica e 1 (um) caso contrário. Assim tem-se uma variável do tipo *dummy*, pronta para ser utilizada para análise. As variáveis que se referem as características de qualidade dos imóveis, como por exemplo a conservação do imóvel, (péssimo, regular, bom e ruim); classificação do imóvel (baixo, normal e alto) e outras, são casos que são resolvidos transformando a variável em novas variáveis do tipo *dummy*. No entanto, quando uma variável pode vir a gerar um número muito grande de modalidades, algumas vezes, ela pode ser definida por uma escala numérica, atribuindo-se pesos às modalidades, geralmente estes pesos são na ordem crescente, da situação menos favorável para a mais favorável. Esta forma é indicada

apenas quando o número de modalidades seria muito grande, para se ter um modelo mais simples e aumentar os graus de liberdade.

As variáveis quantitativas, geralmente, necessitam de transformações para resolver os problemas de falta de linearidade e assimetria. A falta de linearidade pode ser investigada através do gráfico de dispersão da variável dependente versus cada variável independente. Já a assimetria pode ser investigada utilizando-se de histogramas ou diagramas de caixas.

A transformação indicada para resolver os casos citados, principalmente o caso da assimetria positiva, natural para este tipo de dados, pois ocorre muito mais imóveis pequenos e de baixo custo do que grandes, é a transformação logarítmica, porém, não resolvendo, podem ser usadas outras transformações como as citadas nas seções 3.6.1 e 3.6.2 do capítulo 3.

4.1.4 – Análise Exploratória

Para o estudo de relacionamento entre as variáveis, primeiramente pode ser utilizado a correlação linear de Pearson para as variáveis quantitativas e as qualitativas, sendo que no segundo caso, elas devem ser transformadas em *dummy*. O coeficiente de correlação indica a existência ou não de relação linear entre as variáveis independentes e a variável dependente, informação necessária para uso da regressão linear. Estes coeficientes, quando apresentam valores altos entre as variáveis independentes, indicam a possível existência de multicolinearidade, e ainda, o valor do determinante da matriz destes coeficientes, quando é próximo de zero, também indica a existência de multicolinearidade.

Apesar do coeficiente de correlação linear de Pearson e do determinante da matriz destes coeficientes indicarem a existência da multicolinearidade, eles não a quantificam. Contudo, as raízes características da matriz dos coeficientes de correlações ainda podem ser usadas para quantificar a multicolinearidade.

Como o coeficiente de correlação linear de Pearson não é adequado para investigar a relação linear entre a variável dependente e variáveis independentes qualitativas, neste caso sugere-se utilizar o gráfico de dispersão ou teste F da ANOVA.

Quando se tem um conjunto com a presença de variáveis do tipo qualitativas, a análise fatorial de correspondências é adequada e pode ser usada para o estudo da associação entre as variáveis. Esta técnica permite conhecer quais as variáveis que estão fortemente (ou fracamente) associadas entre si. Isto apoia o pesquisador em duas situações: a primeira para indicar possível multicolinearidade; e a segunda, desde que a variável dependente esteja na análise, pode mostrar variáveis fortemente associadas à variável dependente. Este fato, agregado ao conhecimento subjetivo do pesquisador, sobre a importância de determinada variável na explicação da variável dependente, são suficientes para que tal variável seja incluída na equação de regressão. Ainda, esta ferramenta pode ser usada como um reforço, verificando se as variáveis significativas, incluídas na equação, pela análise de regressão, apresentam-se associadas à variável dependente, perante esta ferramenta.

4.1.5 – Construção do Modelo

Após o estudo do relacionamento e da possível existência de multicolinearidade, bem como as transformações necessárias, já se pode dar início à construção do modelo.

Quando têm-se um número grande de variáveis independentes e não existe multicolinearidade, pelo menos não forte, pode se utilizar, para a procura da melhor equação de regressão, o método *Stepwise*. Deve-se tomar cuidado nos valores atribuídos para entrada e saída de variáveis, é aconselhável usar o valor de entrada pouco maior que o de saída, para que no momento da comparação do valor da estatística não se corra o risco de um valor cair entre os valores de entrada e saída. O R^2 ajustado e o quadrado médio dos erros, QME, devem ser usados

para comparar os modelos.

Para o caso de existência de multicolinearidade forte, é indicado o uso da ferramenta *Ridge Regression* (regressão em cumeeira). Aqui é válido o uso da *Ridge Regression* com o método *Stepwise*, e usar o R^2 ajustado e o quadrado médio dos erros para comparar as equações. Deve-se buscar o valor de k que estabiliza os coeficientes e o conjunto de variáveis significativas, isto é, torna os valores das estimativas dos coeficientes constantes e mantém as mesmas variáveis significativas, para valores de k num determinado intervalo.

4.1.6 – Análise Crítica das Variáveis

O conhecimento subjetivo que o pesquisador retém sobre o assunto, especificamente sobre as variáveis, lhe permite uma análise critica das variáveis, tanto as que estão na equação quanto as que foram excluídas. Deve-se analisar as variáveis que foram excluídas mas que o pesquisador julgar relevantes na explicação da variável dependente, ou que se tenha interesse que tal variável esteja na equação. A inclusão de tais variáveis na equação não prejudica o poder de explicação do modelo, e ainda pode melhorar o resultado da predição de novas observações. O contrário também é válido, variáveis que foram incluídas na equação, mas que são consideradas irrelevantes pelo pesquisador, podem ser analisadas estatisticamente. Se mostrar que sua exclusão não prejudica o modelo, então esta pode ser excluída da equação.

A aplicação dos vários métodos na determinação da equação, citados no item anterior, proporcionam ao pesquisador um grande apoio nesta análise crítica, ao comparar os grupos de variáveis presentes em cada equação. Ainda, com apoio do R^2 ajustado e do quadrado médio do erro, pode-se analisar as alterações que a inclusão ou exclusão de algumas variáveis provocam na equação ajustada.

4.1.7 – Análise dos Resíduos

A investigação da adequação do modelo é uma etapa, do procedimento necessário na análise dos dados, tão importante quanto à sua construção. A plotagem dos resíduos é o instrumento usado para examinar o modelo. A análise gráfica dos resíduos é necessária para examinar o ajuste do modelo, ou seja, para confirmar se ele tem uma boa aproximação do verdadeiro sistema e para verificar se as suposições da regressão por mínimos quadrados não foram violadas (Montgomery, 1997, p.563-565).

É mais freqüente o uso dos resíduos padronizados por permitirem mais informações do que os resíduos comuns. Pode-se também, fazer uso dos resíduos "*Studentized*" ou ainda dos resíduos PRESS.

Os resíduos padronizados tem média zero e aproximadamente variância única. Ele é útil na procura de valores discrepantes. Em alguns conjuntos de dados os resíduos podem ter desvios-padrões que diferem grandemente e, como a padronização é feita utilizando o desvio padrão médio, nestes casos é recomendável o uso dos resíduos "*Studentized*", que tem variância constante 1 (um) na região onde o modelo é correto. No entanto, para grandes conjuntos de dados os resíduos padronizados e *Studentized* tem pequena diferença, oferecendo informação equivalente. Os resíduos PRESS também podem ser úteis no diagnóstico do modelo. Grandes resíduos PRESS indicam observações que geralmente são pontos muito influentes. Um resíduo comum muito diferente de um resíduo PRESS geralmente indica um ponto onde o modelo ajusta bem os dados, mas um modelo construído sem este ponto é um pobre preditor. Para maiores detalhes sobre este assunto consultar Montgomery (1997, p.563-565).

Os valores discrepantes, geralmente presentes, devem ser investigados, pois podem representar simplesmente um dado grande ou uma região do espaço da variável regressora onde o modelo ajustado é ruim. Quando existem os pontos discrepantes, eles podem ser comprovados pelo gráfico dos resíduos versus valores preditos. Neste caso, deve-se refazer toda a análise com a exclusão destes pontos. Se

os pontos discrepantes são em pequeno número, comparado com o tamanho da amostra, e não afetam os resultados, eles não precisam ser excluídos. Se são em pequeno número, mas afetam os resultados, então podem e devem ser retirados para realização da análise, porém, merecem uma análise crítica por parte do pesquisador. Sendo estes pontos em grande número, tal que a exclusão compromete a amostra e ou a análise, então o pesquisador deve reanalisar a coleta, a digitação dos dados e até mesmo realizar nova amostragem.

Os resíduos ainda devem ser usados para investigar a possível exclusão de variáveis independentes importantes para explicação da variável dependente, que ficaram fora da equação. Isto pode ser investigado com o gráfico dos resíduos versus cada variável independente não incluída na equação.

4.1.8 – Verificação da Aplicabilidade do Modelo

Um último passo que deve ser realizado antes de adotar o modelo para avaliação de imóveis, é verificar sua aplicabilidade.

Inicialmente, deve se fazer a análise de variância para testar a significância do modelo ajustado, no entanto, isto por si só não garante a qualidade das predições.

A qualidade do ajuste pode ser testada comparando os valores preditos com os valores observados. O ajuste é tão bom, quanto maior for a quantidade de valores preditos próximos dos valores observados, isto é, com pequeno erro de predição.

A calibração do modelo, ou capacidade de predição de novas observações, pode ser feita usando uma nova amostra e comparando os valores de predição com os valores observados. O percentual de predições nas faixas de erro indicarão boa ou má capacidade de predição do modelo.

4.2 – CASO EM ESTUDO

O presente estudo utiliza-se de uma base de dados constituída de 397 apartamentos da cidade de Criciúma, SC. As variáveis são do tipo quantitativas e qualitativas, representando as características do imóvel. Estes dados foram obtidos de Zancan (1995).

O *Software* utilizado para as análises estatísticas foi o *Statistica* 6.0, enquanto que para trabalhar com as transformações de variáveis e criação de variáveis *dummy*, utilizou-se do *Software Excel*.

4.3 – IDENTIFICAÇÃO E APRESENTAÇÃO DAS VARIÁVEIS

A variável dependente é o preço, que representa o valor de venda do imóvel em dólares. As variáveis independentes estão relacionadas, classificadas e descritas nos quadros 4.1 e 4.2.

Quadro 4.1: Descrição das variáveis independentes quantitativas

VARIÁVEIS	UNIDADES DE MEDIDA	DESCRIÇÃO
Área total (AT)	m2	identifica a área total do imóvel
Consumo de energia (CE)	kw	representa o consumo médio de energia, referente ao imóvel. No imóvel desocupado utilizou-se a média de consumo segundo a área residencial
Distância à escola (ES)	m	distância do imóvel à escola mais próxima
Acessibilidade (AC)	m	distância do imóvel ao ponto de ônibus mais próximo
Idade (ID)	anos	representa a idade do imóvel após a liberação do habite-se
Dormitórios (DO)	unidades	identifica o número de dormitórios
Meio Ambiente (MA)	-	representa o nível de poluição, obtido através de notas dadas por engenheiros de diversas áreas e calculadas as estatísticas descritivas

Quadro 4.2: Descrição das variáveis independentes qualitativas

VARIÁVEIS	CATEGORIAS	DESCRIÇÃO
Região Homogênea (RH)	1 à 11	regiões de mesma valorização imobiliária, levando em consideração a tipologia construtiva, infra-estrutura, topografia e idade das construções
Zona Fiscal (ZF)	1 à 16	zoneamento fiscal do município
Padrão de Entrada (PE)	baixo=1 normal=2 alto=3	padrão de vizinhança do imóvel
Classificação (CL)	baixo=1 normal=2 alto=3	padrão construtivo do imóvel
Conservação (CO)	péssimo=1 regular=2 bom=3 ótimo=4	Identifica o nível de conservação do imóvel
Garagem (GA)	Sem=0 com=1	identifica a existência de garagem
Suíte (SU)	sem=0 com=1	identifica a existência de suite
Dependência de empregada (DE)	sem=0 com=1	identifica a existência de dependência de empregada
Elevador (EL)	sem=0 com=1	identifica a existência de elevador
Pólos de valorização	pertence=1 não pertence=0	representam os seis pólos de valorização imobiliária: Bombeiro (PBO), Marista (PMA), Centenário (PCE), Michel (PMI), Praça Congresso (PPC) e Comerciário (PCO)

4.4 – LEVANTAMENTO DE DADOS

Para formação da amostra, Zancan (1995) escolheu imóveis do tipo apartamento, não novos, usando o banco de dados imobiliário de

Criciúma, SC, em confronto com o cadastro urbano do município.

Foram utilizados 380 imóveis para a determinação da equação de regressão e uma amostra de tamanho 17 foi retirada do conjunto inicial de dados para testar o ajuste do poder preditivo do modelo, conforme sugerem Neter e Wasserman (1974, p.388).

4.5 – TRANSFORMAÇÕES DE VARIÁVEIS

A existência de variáveis qualitativas e quantitativas no caso em estudo, fez necessário, primeiramente, a transformação de algumas variáveis, para realização do estudo do relacionamento entre elas.

As variáveis qualitativas precisaram ser transformadas em variáveis do tipo *dummy* para se calcular os coeficientes de correlação linear e fazer regressão. No entanto, as variáveis região homogênea (RH) e zona fiscal (ZF) foram classificadas com um número muito grande de categorias, o que levaria a um número muito grande de variáveis *dummy*. Assim, estas variáveis foram categorizadas novamente, conforme mostra o quadro 4.3, baseando-se na análise dos gráficos das distribuições de freqüência com diferentes números de categorias.

Quadro 4.3: Nova categorização das variáveis RH e ZF

VARIÁVEIS	CATEGORIAS	NOVAS CATEGORIAS
Região Homogênea (RH)	1 à 11	1 e 2 = 1 3 e 4 = 2 5 e 6 = 3 7 e 8 = 4 9, 10 e 11 = 5
Zona Fiscal (ZF)	1 à 16	1 e 2 = 1 3 e 4 = 2 5 e 6 = 3 7 e 8 = 4

Com estas novas categorias, estas variáveis puderam, coerentemente, serem transformadas em variáveis do tipo *dummy*, juntamente com as demais variáveis qualitativas: padrão de entrada (PE), classificação (CL) e conservação (CO), conforme o quadro 4.4.

Quadro 4.4: Transformação das variáveis qualitativas em *dummy*

VARIÁVEIS	CATEGORIAS
Região Homogênea 1 (RH1)	pertence=1 não pertence=0
Região Homogênea 2 (RH2)	pertence=1 não pertence=0
Região Homogênea 3 (RH3)	pertence=1 não pertence=0
Região Homogênea 4 (RH4)	pertence=1 não pertence=0
Zona Fiscal 1 (ZF1)	pertence=1 não pertence=0
Zona Fiscal 2 (ZF2)	pertence=1 não pertence=0
Zona Fiscal 3 (ZF3)	pertence=1 não pertence=0
Padrão de Entrada 1 (PE1)	pertence=1 não pertence=0
Padrão de Entrada 2 (PE2)	pertence=1 não pertence=0
Classificação 1 (CL1)	pertence=1 não pertence=0
Classificação 2 (CL2)	pertence=1 não pertence=0
Conservação 1 (CO1)	pertence=1 não pertence=0
Conservação 2 (CO2)	pertence=1 não pertence=0
Conservação 3 (CO3)	pertence=1 não pertence=0

Desta forma, passou a se ter dois conjuntos de dados: o primeiro com as variáveis quantitativas e as qualitativas categorizadas, ambas na forma original (totalizando 21 variáveis independentes); e o segundo com as variáveis quantitativas na forma original e as qualitativas transformadas em *dummy* (totalizando 30 variáveis independentes).

Para estudo da assimetria das variáveis quantitativas foram construídos diagramas de caixa. Os resultados são mostrados na Figura 4.1. Observa-se que a variável dependente preço é assimétrica positiva. As variáveis independentes área total (AT), consumo de energia (CE), distância à escola (ES), acessibilidade (AC), meio ambiente (MA) e idade (ID), assim como a variável dependente, também apresentaram-se assimétricas positiva.

Com a transformação logarítmica, Figura 4.2, a variável Ln(PR) tornou-se simétrica. Para as variáveis independentes apresentadas acima, também foi utilizada a transformação logarítmica, ficando Ln(AT), Ln(CE), Ln(ES), Ln(AC), Ln(MA) e Ln(ID). A Figura 4.2 mostra que para as variáveis Ln(AT), Ln(ES), Ln(AC) e Ln(MA) a transformação logarítmica amenizou o problema da assimetria. Para a variável consumo de energia e Ln(consumo de energia), além do diagrama de caixa, foi construído o histograma e o gráfico de dispersão, então foi possível perceber uma pequena melhora com a transformação logarítmica. Esta pequena melhora mais o fato de que os valores da variável consumo de energia diferem (são muito maiores) dos valores das demais variáveis, levam a concluir que se deve trabalhar com esta variável transformada, ou seja, Ln(CE). Quanto a variável idade, os estudos através do diagrama de caixa, histograma e gráfico de dispersão não mostraram haver qualquer melhora com a transformação logarítmica.

Assim, para os estudos posteriores, a variável dependente preço (PR) e as variáveis independentes área total (AT), consumo de energia (CE), distância à escola (ES), acessibilidade (AC), meio ambiente (MA) serão utilizadas com a transformação logarítmica e a variável idade (ID) na forma original.

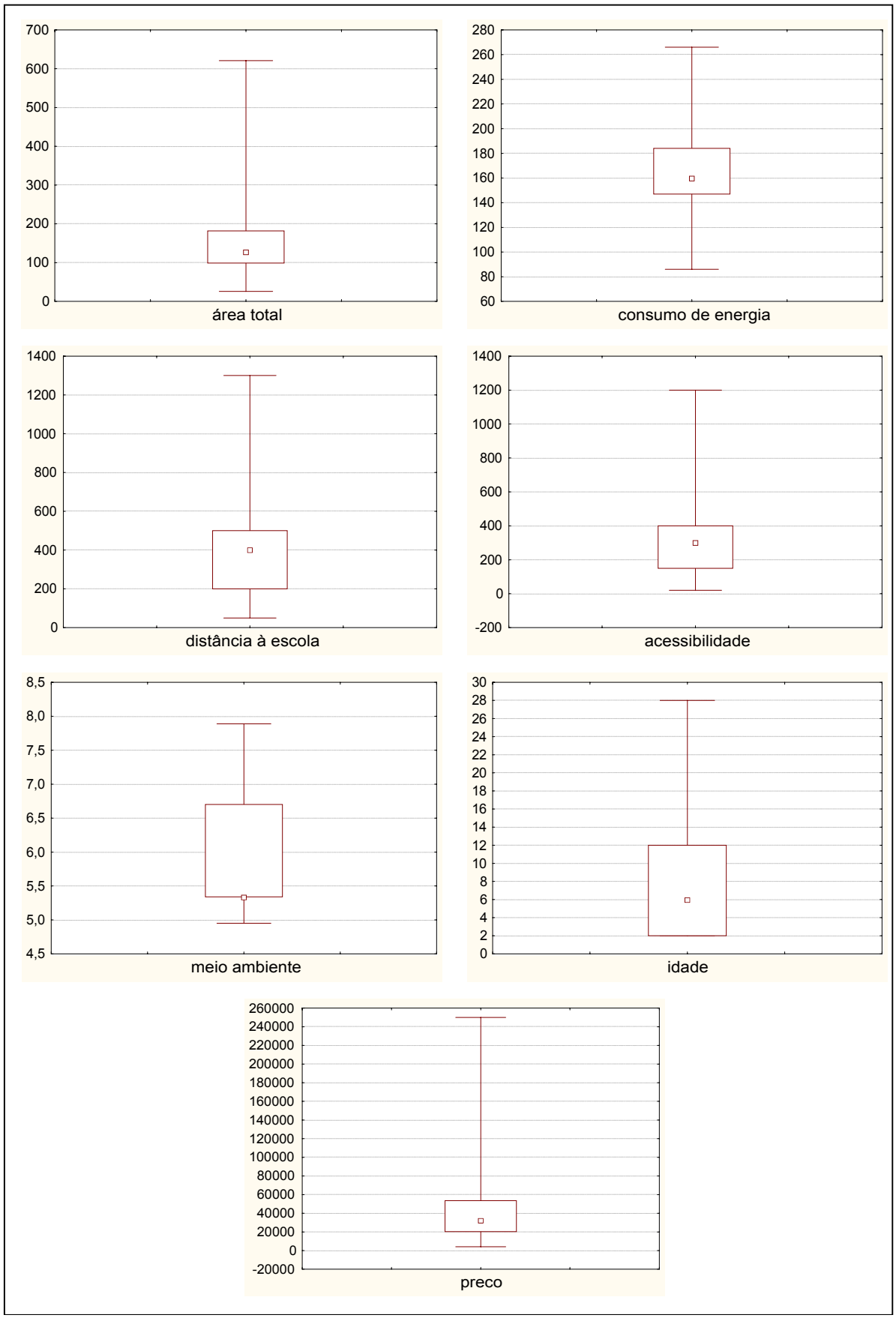

Figura 4.1: Diagramas de caixa das variáveis quantitativas

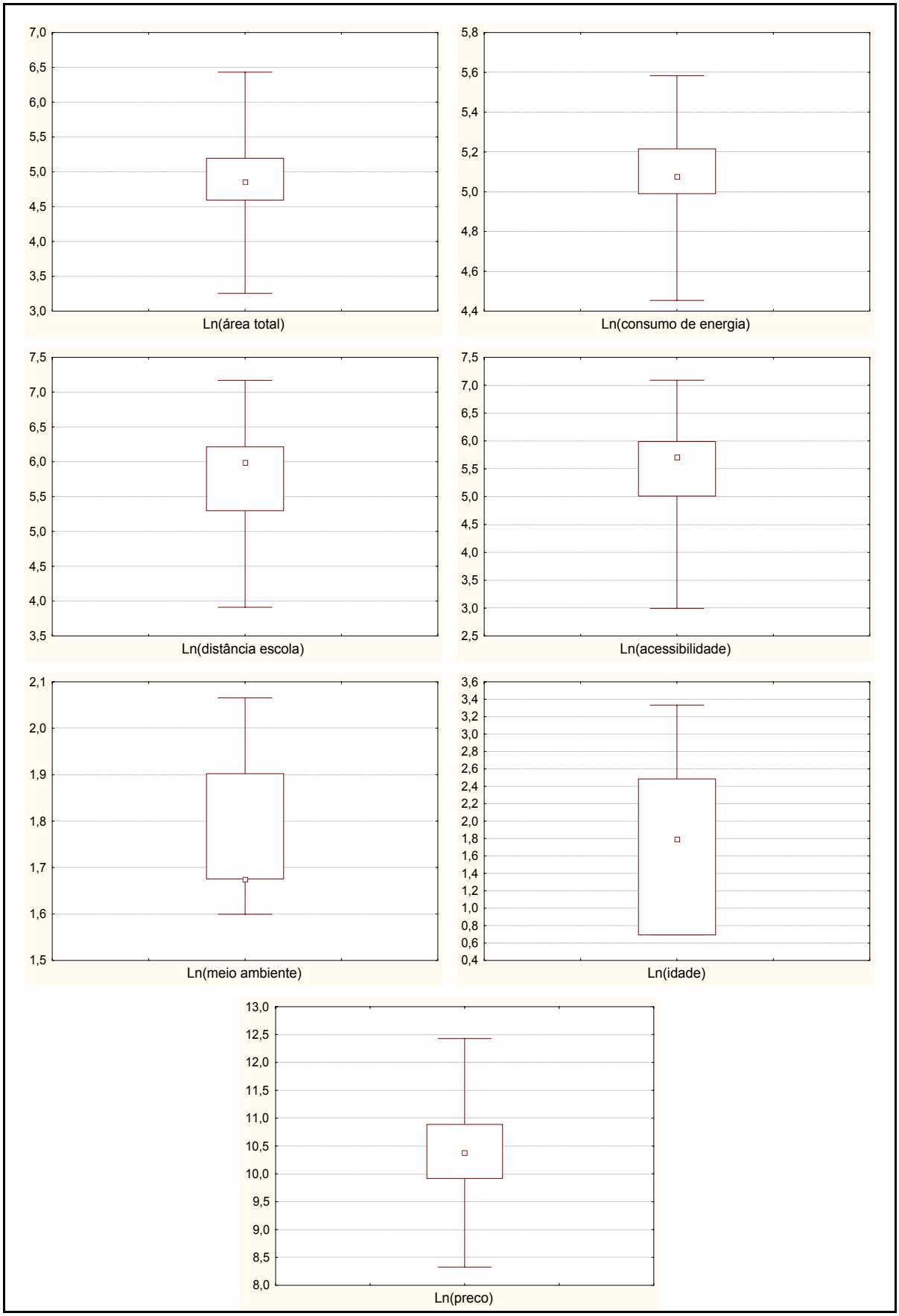

Figura 4.2: Diagramas de caixa das variáveis quantitativas com transformação logarítmica

4.6 – ANÁLISE EXPLORATÓRIA DAS VARIÁVEIS

Para a análise exploratória das variáveis pode-se usar os coeficientes de correlação linear de Pearson e gráficos de dispersão.

4.6.1 – Relação da Variável Dependente com as Independentes

Os coeficientes de correlação linear de Pearson entre a variável dependente preço e as variáveis independentes quantitativas, já com as transformações logarítmicas, são mostrados na tabela 4.1.

Os valores pequenos, mostram a falta de linearidade entre os pontos. Assim, não é viável um ajuste linear para estas variáveis, sendo, portanto, necessário alguma transformação, ou ainda, estas variáveis não são importantes na explicação do preço do imóvel. Observa-se que as variáveis Ln(área total), Ln(consumo de energia) e idade (ID), são as mais relacionadas linearmente com a variável preço. A Figura 4.3 mostra haver relacionamento linear forte do Ln(PR) com Ln(AT) e Ln(CE), fraco com (ID) e não haver qualquer tipo de relacionamento com as demais variáveis.

Tabela 4.1: Correlação linear (Pearson) entre a variável dependente ln(preço) e as variáveis independentes quantitativas

VARIÁVEIS INDEPENDENTES	**COEFICIENTES DE CORRELAÇÃO**	VALOR P
Ln(área total)-Ln(AT)	0,9432	0,000
Ln(consumo de energia)-Ln(CE)	0,5368	0,000
Ln(distância à escola)-Ln(ES)	-0,1182	0,021
Ln(acessibilidade)-Ln(AC)	0,2374	0,000
Ln(meio ambiente)-Ln(MA)	0,1309	0,011
Idade (ID)	-0,4920	0,000
Dormitórios (DO)	0,0243	0,637

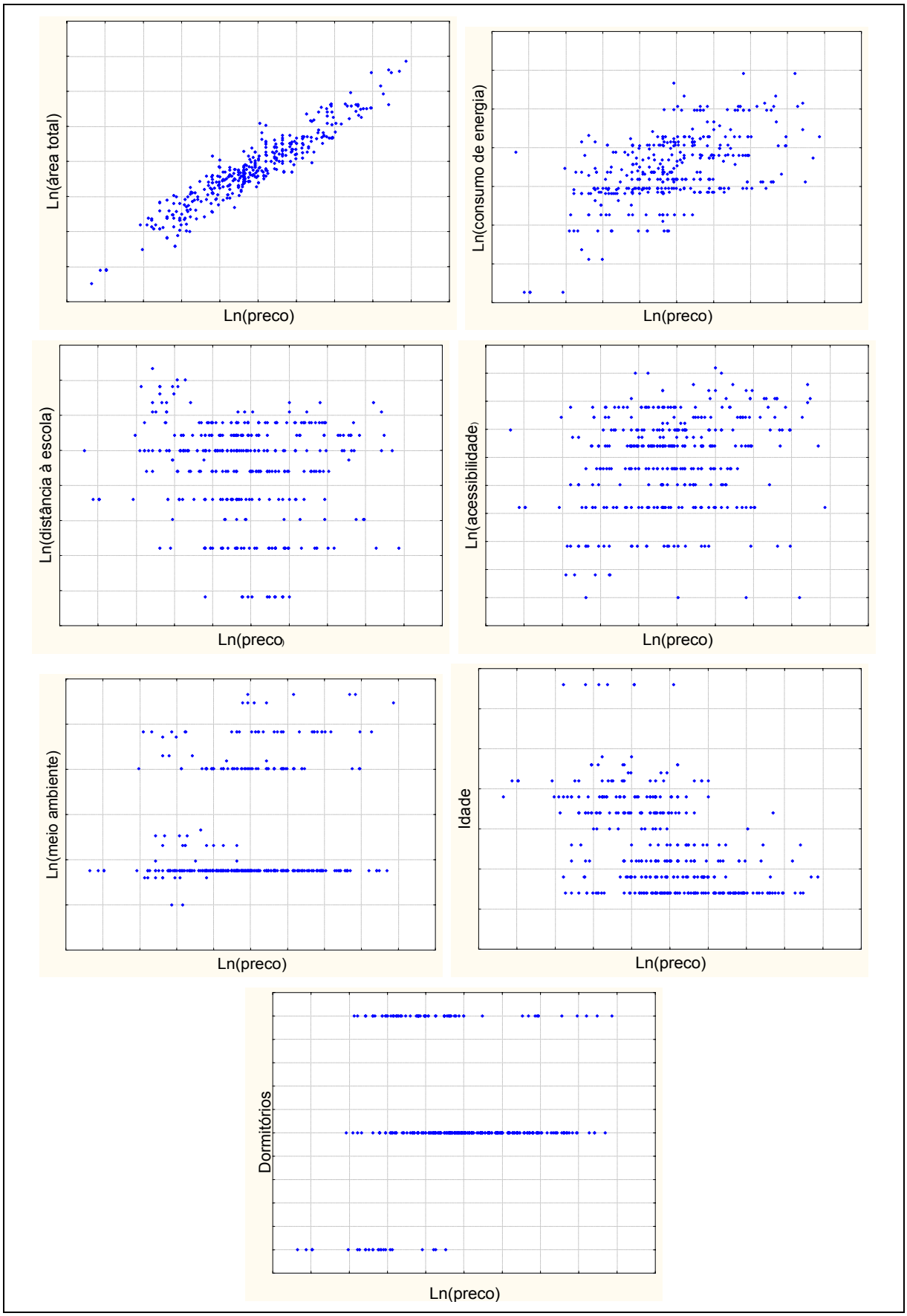

Figura 4.3: Gráficos de dispersão da variável dependente Ln(preço) versus variáveis quantitativas independentes

Para investigar possíveis relações entre a variável dependente e as variáveis independentes qualitativas e dummy, fez-se uso do teste F da ANOVA. A tabela 4.2 mostra os resultados. Se for considerado um nível de significância de 1%, já que n é consideravelmente grande, apesar de que a NBR 5679/90 permite 5%, duas destas variáveis podem ser excluídas da análise de regressão, estas variáveis são pólo bombeiro e pólo comercial.

Tabela 4.2: Teste F da ANOVA para a variável dependente ln(preço) versus variáveis independentes qualitativas

VARIÁVEIS INDEPENDENTES			VALORE DE F	VALOR P
VARIÁVEL	CATEGORIA	MÉDIA		
Região homogênea	1	9,3179	83	0,00
	2	9,8851		
	3	10,4112		
	4	10,8907		
	5	11,3853		
Zona fiscal	1	10,6224	7	0,00
	2	10,4024		
	3	10,2503		
	4	9,6617		
Padrão de entrada	1	9,4248	35	0,00
	2	10,3489		
	3	11,2978		
Classificação	1	9,6072	269	0,00
	2	10,4221		
	3	11,4839		
Conservação	1	9,6099	80	0,00
	2	9,9150		
	3	10,3226		
	4	10,9039		
Garagem	0	9,4479	199	0,00
	1	10,6129		
Suíte	0	9,7541	382	0,00
	1	10,8347		

Continuação ...

VARIÁVEIS INDEPENDENTES			VALOR DE F	VALOR P
VARIÁVEL	CATEGORIA	MÉDIA		
Dependência de empregada	0	10,0279	274	0,00
	1	11,0128		
Elevador	0	9,9466	32	0,00
	1	10,5083		
Pólo bombeiro	0	10,4527	4	0,04
	1	10,2453		
Pólo Marista	0	10,3540	34	0,00
	1	11,1481		
Pólo centenário	0	10,5711	34	0,00
	1	10,1231		
Pólo Michel	0	10,4872	46	0,00
	1	9,5626		
Pólo praça do congresso	0	10,2709	125	0,00
	1	11,3573		
Pólo comercial	0	10,4016	1	0,42
	1	10,4900		

Em face dos resultados obtidos nesta seção, pode-se decidir pela exclusão das variáveis independentes quantitativas Ln(distância à escola), Ln(acessibilidade), Ln(meio ambiente) e Dormitórios (DO) e das variáveis *dummy* pólo bombeiro (PBO) e pólo comercial (PCO). Note que, tabela 4.1, as variáveis Ln(distância à escola) e Ln(meio ambiente) não são significativamente correlacionadas a variável Ln(preço) a 1%, porém o coeficiente de correlação é baixo e a figura 4.3 confirma este fraco correlacionamento. Também, a variável Ln(acessibilidade), apesar de existir correlação com a variável Ln(preço), ela é baixa e a Figura 4.3 confirma esta falta de relacionamento. Estes fatos é que levaram a decisão de exclusão destas variáveis.

Assim, passa a se ter um conjunto com um número menor de variáveis independentes, terceiro conjunto, com 24 variáveis.

Ainda, a análise fatorial de correspondências foi usada como uma ferramenta de apoio ao estudo do relacionamento e importância das variáveis, pois esta técnica permite detectar as relações recíprocas, associações e oposições entre variáveis. Para realizar a análise fatorial de correspondências, as variáveis quantitativas contínuas foram discretizadas. O apêndice A, mostra a plotagem das variáveis utilizando-se os fatores 1, com poder de explicação de 40% do total da inércia; e o fator 2, com poder de explicação de 11% do total da inércia. Observa-se que as variáveis área total (AT), região homogênea (RH), classificação (CL), garagem (GA), conservação (CO), consumo de energia (CE), elevador (EL) e suite (SU), apresentaram-se associadas à variável dependente preço, indicando que estas variáveis podem ser importantes na determinação da equação de regressão. Ainda pode-se observar, através do apêndice A, que a análise fatorial de correspondências que as variáveis distância à escola (ES), acessibilidade (AC), meio ambiente (MA), Dormitórios (DO), pólo bombeiro (PBO) e pólo comercial (PCO) não têm nenhuma associação com a variável dependente preço (PR), indicando, também, como os resultados anteriores, a exclusão destas variáveis.

4.6.2 – Multicolinearidade

Para a investigação da possível existência de multicolinearidade, foi utilizado o valor do determinante da matriz das correlações, Rx. As correlações, apêndice B, foram calculadas usando o terceiro conjunto com 24 variáveis independentes.

Os altos valores das correlações indicam existência de possível multicolinearidade, porém para conjuntos com muitas variáveis, mesmo não havendo altas correlações, ainda pode haver multicolinearidade.

O valor do determinante da matriz das correlações foi de $3{,}87.10^{-7}$. Este valor, aproximadamente zero, é indicativo da existência de multicolinearidade, porém, isto não a quantifica.

Outro critério utilizado, para detectar e também quantificar, foi através do uso das raízes características da matriz Rx. Observe, na tabela 4.3, que existem várias raízes características com valor muito baixo, levando à indicação da existência de multicolinearidade, porém não se tem a quantificação desta multicolinearidade.

Tabela 4.3: Raízes características da matriz das correlações

NÚMERO	AUTOVALORES	VARIÂNCIA EXPLICADA (%)	VARIÂNCIA EXP. ACUM. (%)
1	10,7704	44,8767	44,8767
2	5,2213	21,7555	66,6322
3	2,3780	9,9083	76,5405
4	1,2013	5,0052	81,5458
5	0,8620	3,5916	85,1373
6	0,7009	2,9206	88,0579
7	0,5751	2,3962	90,4541
8	0,5157	2,1486	92,6027
9	0,4070	1,6960	94,2987
10	0,3693	1,5385	95,8373
11	0,2692	1,1215	96,9588
12	0,1882	0,7841	97,7429
13	0,1576	0,6569	98,3998
14	0,1086	0,4525	98,8523
15	0,0921	0,3839	99,2362
16	0,0629	0,2623	99,4984
17	0,0454	0,1891	99,6875
18	0,0405	0,1688	99,8563
19	0,0135	0,0561	99,9124
20	0,0092	0,0383	99,9507
21	0,0087	0,0361	99,9868
22	0,0024	0,0100	99,9968
23	0,0008	0,0032	100,0000

Para descrever o nível de multicolinearidade usou-se o critério descrito em Elian (1998, p.131), que consiste no cálculo do valor de L, dado por:

$$L=\lambda_{máx}/\lambda_{mín} = 10,7704/0,0008 = 13926.$$

Como o valor de L é maior que 1000, concluir-se, segundo Elian (1998, p.131), que existe multicolinearidade forte.

Assim, foi necessário buscar a determinação da equação de

regressão, utilizando métodos específicos para casos da existência de multicolinearidade.

4.7 – CONSTRUÇÃO DO MODELO

Para contornar o problema da multicolinearidade, o método escolhido foi o *Ridge Regression*, pois, neste caso de multicolinearidade forte, ele produz um QME (quadrado médio do erro) menor do que o QME produzido pelos estimadores de mínimos quadrados ordinários.

A análise de regressão foi realizada usando o terceiro conjunto de dados (24 variáveis independentes) com as variáveis qualitativas transformadas em variáveis *Dummy*, as quantitativas área total (AT) e consumo de energia (CE) com a transformação logarítmica e a quantitativa idade (ID) na forma original. A variável dependente preço foi usada com a transformação logarítmica.

Foram realizadas inúmeras simulações, com o método *Ridge Regression Stepwise Standard*, para identificar o melhor conjunto de variáveis a entrarem na equação.

O valor de k = 0,0001 estabilizou os coeficientes, o R^2 ajustado e o quadrado médio do erro, isto é, tornou os valores das estimativas dos coeficientes constantes, manteve as mesmas variáveis significativas, manteve praticamente inalterados os valores do R^2 ajustado e do quadrado médio do erro, para valores de k próximos de 0,0001.

Porém, a equação determinada apresentou-se incoerente para a avaliação de imóveis por apresentar o coeficiente da variável ln(consumo de energia) com sinal negativo. A variável consumo de energia representa a variável renda familiar e isto implica que uma família com maior renda familiar deve ter um imóvel de maior valor. Logo esta variável entrou na equação de forma incoerente e para corrigir tal incoerência, fez-se necessária a exclusão desta variável.

Análises posteriores mostraram a presença de pontos discrepantes, sendo dois pontos afastados mais de 3 desvios-padrões da média,

correspondentes às observações de ordem 28 e 311. Foram realizadas análises com a retirada dos dois pontos, havendo uma melhora no ajuste da equação, pois houve aumento do valor do R^2 ajustado e diminuição do valor do quadrado médio do erro. Como a amostra é consideravelmente grande, 380 imóveis, a redução de 2 observações não deve acarretar problemas. Assim, passou-se a trabalhar com este novo conjunto de 378 imóveis.

As variáveis significativas e os respectivos valores dos coeficientes são apresentados na tabela 4.4 e os valores das medidas referentes ao ajuste da equação são apresentados na tabela 4.5. Observa-se que os valores de R múltiplo e R^2 ajustado são consideravelmente altos, indicando alta correlação alta correlação da variável dependente com as variáveis independentes e alta explicação da variável dependente pelas variáveis independentes.

Tabela 4.4: *Ridge Regression* para determinação da equação

VARIÁVEIS	COEF. ESTIMADOS	VALOR $t_{(366)}$	VALOR P
Constante	44.06	5.839751	0.000
Região Homogênea 1	-7,26	-0,383533	0,000
Região Homogênea 2	-6,91	-0,298354	0,000
Região Homogênea 3	-5,39	-0,201594	0,000
Região Homogênea 4	-3,24	-0,121817	0,001
Classificação 1	-7,32	-0,330161	0,000
Classificação 2	-5,51	-0,183774	0,000
Conservação 1	-6,66	-0,239793	0,000
Conservação 2	-5,74	-0,166772	0,000
Conservação 3	-5,79	-0,126998	0,000
Dependência de empregada	2,32	0,051752	0,021
Ln(área total)	40,67	1,031288	0,000

Tabela 4.5: Medidas referentes ao ajuste da equação

MEDIDAS	VALORES
R múltiplo	0.9777
R^2	0,9559
R^2 ajustado	0,9546
Quadrado Médio do Erro	83.609.000,00

As variáveis quantitativas que no estudo realizado na seção 4.6.1, apresentaram-se relacionadas à variável dependente preço foram ln(AT), ln(CE) e idade. No entanto, a variável ID não foi significativa na análise de regressão, e, como não se tem argumentos que justifique sua inclusão na equação, ela foi excluída.

Em relação às variáveis qualitativas que, individualmente, foram consideradas pelo teste F da ANOVA serem relevantes para a explicação da variável preço, apenas as variáveis região homogênea, classificação, conservação e dependência de empregada foram significativas na análise de regressão. A variável dependência de empregada foi significativa ao nível de 2%, as demais ao nível de 1%. Dentre as variáveis não significativas, as variáveis suíte e garagem merecem atenção, porque além de apresentaram-se relevantes na relação com a variável preço, conforme o estudo realizado na seção 4.6.1, são valorizadas pelo mercado. Porém, em análises realizadas com estas variáveis incluídas na equação, não houve melhora das medidas do R^2 ajustado nem mesmo do quadrado médio do erro. Assim, estas variáveis não foram incluídas na equação final.

Portanto, a equação com o melhor R^2 ajustado e menor quadrado médio do erro foi obtida com as variáveis região homogênea, classificação, conservação, dependência de empregada e ln(área total).

A equação ajustada ficou da seguinte forma:

Preço = **exp {** 5,839751 **–** 0,383533 • Região Homogênea 1 **–** 0,298354 • Região Homogênea 2 **–** 0,201594 • Região Homogênea 3 **–** 0,121817 • Região Homogênea 4 **–** 0,330161 • Classificação 1 **–** 0,183774 • Classificação 2 **–** 0,239793 • Conservação 1 **–** 0,166772 • Conservação 2 **–** 0,126998 • Conservação 3 **+** 0,051752 • Dependência de Empregada **+** 1,031288 • ln(área total) **}**

Onde:

- Preço representa o valor de venda do imóvel em dólares;

- DE representa a variável indicadora de dependência de empregada;
- Ln(AT) representa a área total com transformação logarítmica;
- RH1, RH2, RH3 e RH4 representam as regiões homogêneas nas categorias 1, 2, 3 e 4 respectivamente (RH5 foi usada como padrão);
- CL1 e CL2 representam as classificações nas categorias 1 e 2 respectivamente (CL3 foi usada como padrão);
- CO1, CO2 e CO3 representam os níveis de conservação 1, 2 e 3 respectivamente (CO4 foi usada como padrão).

A análise de variância, tabela 4.6, mostra que rejeitamos a hipótese de não haver regressão, isto é, o modelo é significativo.

Tabela 4.6: Análise de variância para a significância da equação

FONTES DE VARIAÇÃO	SOMA DE QUADRADOS	GRAUS DE LIBERDADE	QUADRADO MÉDIO	F_0	P
Regressão	198,18	11	18,02	721,32	0,000
Resíduo	9,14	366	0,025		
Total	207,32	377			

4.8 – ANÁLISE DE RESÍDUOS

As suposições do modelo, agora, com a equação de regressão ajustada podem ser verificadas através da análise gráfica, utilizando-se os resíduos.

Variância Constante: a Figura 4.4 apresenta os pontos distribuídos aleatoriamente em torno da linha que passa pela origem. Esta disposição dos pontos indica que a suposição de variância constante é razoável.

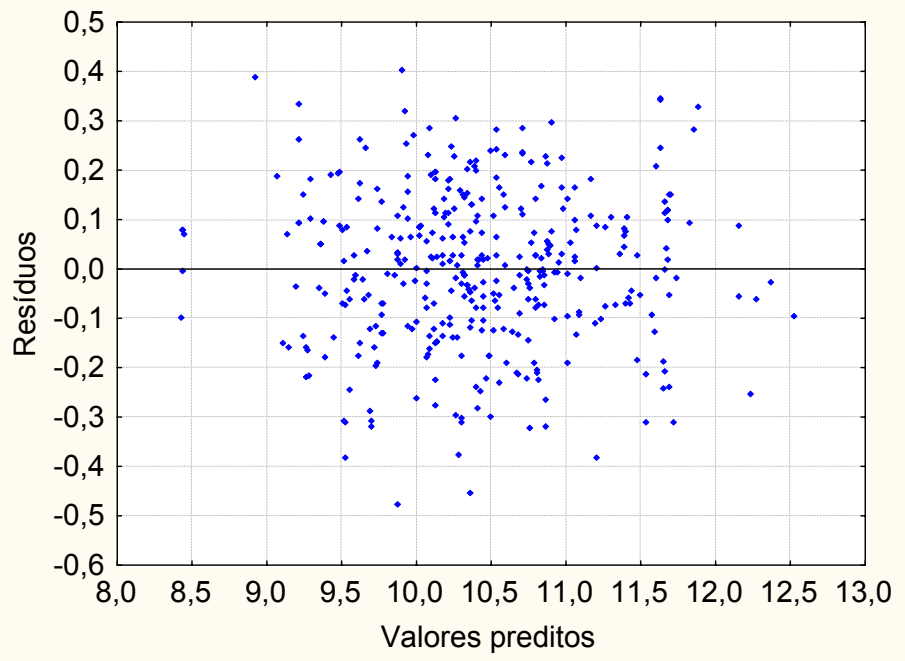

Figura 4.4: Variância constante – plotagem dos resíduos versus valores preditos.

Distribuição Normal: a Figura 4.5 mostra que os pontos estão dispostos sob uma linha reta, indicando a normalidade dos erros. A tabela 4.7 apresenta os percentuais de pontos exigidos, nas respectivas faixas, para se ter o atendimento da suposição de normalidade. O resultado obtido confirma a interpretação gráfica, pois os percentuais de resíduos exigidos dentro de cada faixa foram atendidos e, portanto, a suposição de normalidade foi atendida.

Tabela 4.7: Comparação da distribuição dos resíduos padronizados com a distribuição normal padrão

INTERVALOS	VALOR TEÓRICO	N. DE PONTOS	VALOR OBTIDO
[-1 ; 1]	68%	225	68%
[-1,64 ; 1,64]	90%	340	90%
[-1,96 ; 1,96]	95%	359	95%

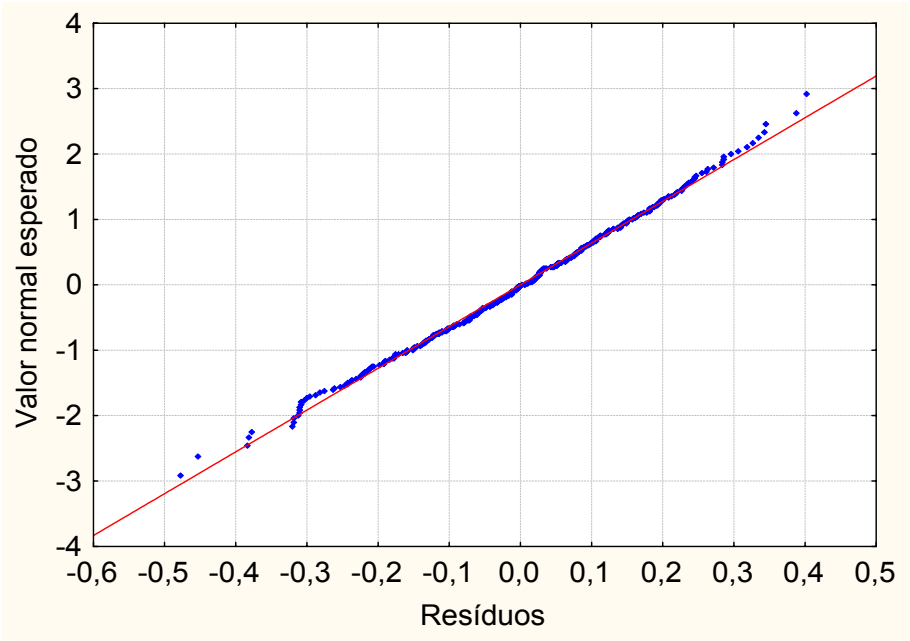

Figura 4.5: Normalidade – plotagem dos resíduos versus valores esperados pela distribuição normal.

Ainda, utilizando os resíduos da equação ajustada, é possível investigar a adequação da equação, a existência de valores discrepantes e a omissão de variáveis independentes importantes para a equação.

Adequação do modelo: a Figura 4.4 mostra que a distribuição dos pontos é aleatória em torno de zero, que indica uma escolha adequada da forma da equação matemática.

Foi realizada também a plotagem dos resíduos versus cada variável independente participante da equação, e nenhum dos gráficos apresentou qualquer padrão sistemático que levasse a suspeita de mal ajuste do modelo (ver apêndice C).

Valores discrepantes: Não apareceram resíduos padronizados maiores que três, ou seja, três desvios padrões afastados de zero (a média). Apareceram alguns valores afastados mais do que 2 desvios-padrões, conforme mostra a Figura 4.6, porém, não se caracterizam pontos discrepantes devido o tamanho da amostra ser grande no presente caso.

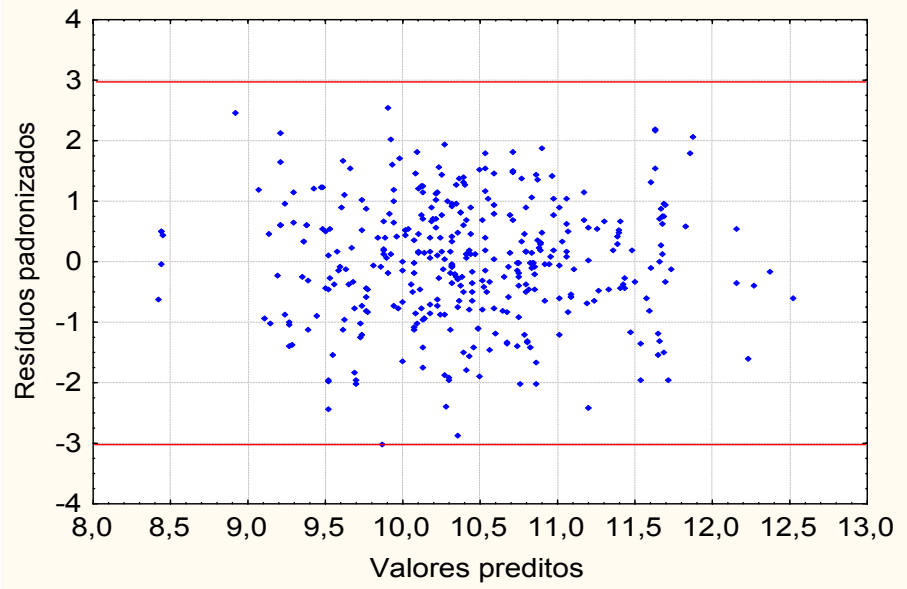

Figura 4.6: Valores discrepantes – plotagem dos valores preditos versus resíduos padronizados.

Omissão de variáveis independentes: os gráficos da plotagem dos resíduos versus cada variável independente não incluída na equação, são usados para investigar se alguma destas, deva ser incluída. As variáveis, que não participaram do modelo, mas pelas análises preliminares eram suspeitas de serem importantes, como suíte (SU), garagem (GA) e idade (ID), não o são, conforme pode ser visto nas Figuras 4.7, 4.8 e 4.9, respectivamente. Observe que não existe tendência que possa justificar a inclusão destas variáveis na equação.

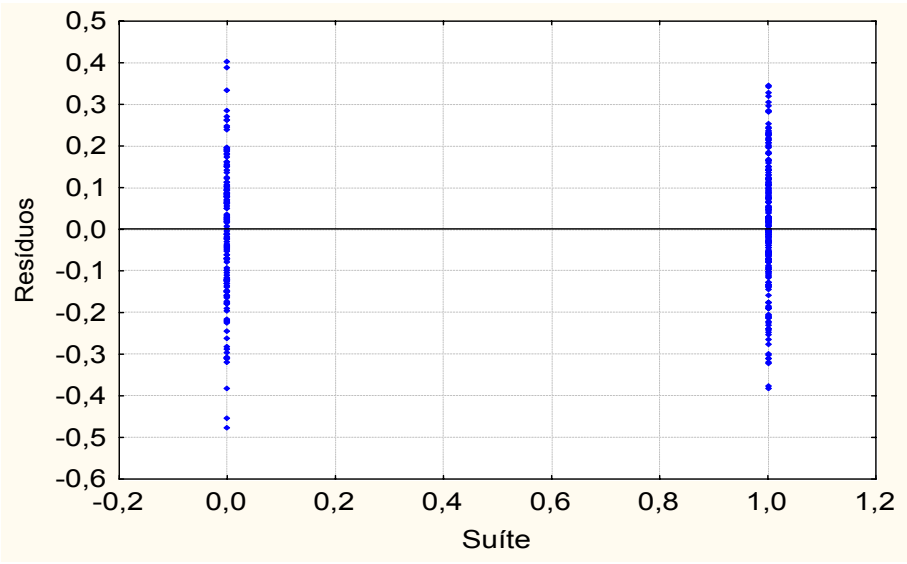

Figura 4.7: Omissão de variáveis – valores observados da variável suite versus resíduos

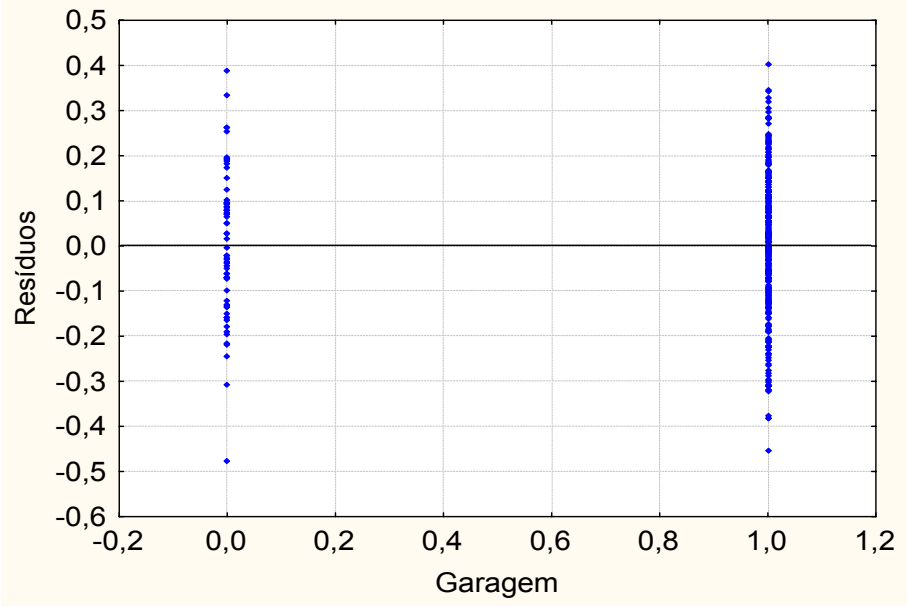

Figura 4.8: Omissão de variáveis – valores observados da variável garagem versus resíduos

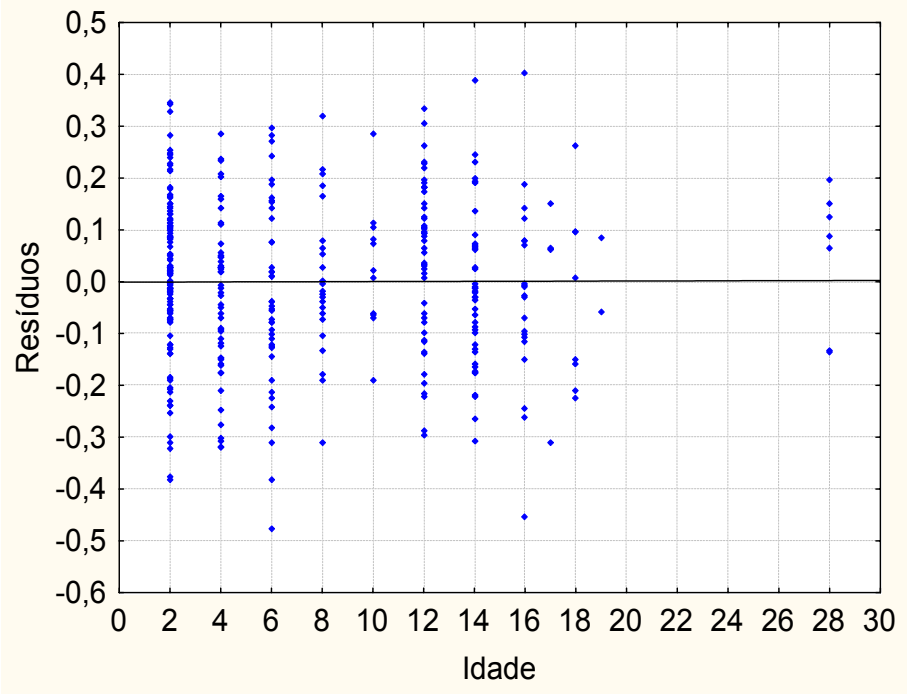

Figura 4.9: Omissão de variáveis – valores observados da variável idade versus resíduos

Quanto as demais variáveis, nenhuma delas apresentou qualquer padrão, apenas a variável padrão de entrada-1 (PE1) apresentou uma suave tendência, mas não o suficiente para ser incluída na equação

(ver apêndice D). Logo, pode-se dizer que o conjunto de variáveis incluídas na equação é o mais coerente.

4.9 – AVALIAÇÃO PRÁTICA DO MODELO CONSTRUÍDO

A construção de um modelo que satisfaz todas as suposições teóricas não é o suficiente para garantir a qualidade das predições. Uma avaliação prática do modelo mostrará a sua qualidade de ajuste e capacidade preditiva.

valores preditos: Os valores preditos pela equação ajustada, dentro das faixas de erros de 5%, 10%, 15%, 20%, 30% 40% e 50%, conforme tabela 4.8, mostra o ajuste da equação ajustada.

Tabela 4.8: Percentual de valores preditos nas faixas de 5% a 50%

FAIXAS DE ERRO	N. DE VALORES PREDITOS	% DE VALORES PREDITOS
5%	96	25,4%
10%	184	48,7%
15%	250	66,1%
20%	301	79,6%
30%	358	94,7%
40%	374	98,9%
50%	378	100%

Note que 4 imóveis tiveram valor de predição com erro superior a 40%, sendo que o maior erro individual foi de 49,5%.

Para melhor clareza e entendimento do que significam os resultados da tabela 4.8, tomou-se um imóvel de valor R$ 30.000,00 e os valores preditos nas respectivas faixas de erro são apresentadas na tabela 4.9.

Observe que, em se tratando de mercado imobiliário, especificamente valor de venda do imóvel, que tem muitos fatores não controlados influenciando, o valor de um imóvel de R$ 30.000,00 sendo

predito na primeira faixa é um excelente resultado, na segunda faixa um ótimo resultado, na terceira faixa um bom resultado e mesmo na quarta faixa o resultado é aceitável. Logo, baseado nos percentuais de valores preditos em cada faixa, pode-se concluir que a equação ajustada é boa preditora.

Tabela 4.9: Exemplo da variação do preço predito de um imóvel de valor R$ 30.000,00

FAIXAS DE ERRO	VALOR DE VARIAÇÃO	VARIAÇÃO DO VALOR PREDITO
5%	1.500,00	(28.500,00 ; 31.500,00)
10%	3.000,00	(27.000,00 ; 33.000,00)
15%	4.500,00	(25.500,00 ; 34.500,00)
20%	6.000,00	(24.000,00 ; 36.000,00)

Geralmente, nos estudos encontrados, os pesquisadores não apresentam dados que mostram a qualidade do ajuste. Nos estudos realizados por Worzala *et al*. (1995), são citados resultados entre 24% e 37% de imóveis dentro da faixa de 5%. Ainda citou erro de predição individual de 60,7%. Estas variações são, provavelmente, devido as diversas amostras que o autor usou, algumas com conjuntos mais homogêneos e, também, com amostras feitas em certas faixas de preços.

Nos estudos realizados por Zancan (1995), a equação construída proporcionou um ajuste inferior à equação construída neste trabalho, conforme mostra a tabela 4.10. Isto pode ter ocorrido devido a existência de multicolinearidade forte, pois a autora não usou técnicas para amenizar os problemas que ela causa. Ainda, a autora trabalhou com as variáveis qualitativas na forma categorizada, sem transformar em variáveis do tipo *dummy*. O modelo construído pela autora produziu erro de predição individual de 72%.

Os resultados citados, de Worzala *et al*. (1995) e Zancan (1995), apoiam a conclusão de que o modelo encontrado neste estudo de caso, com 25,4% dos imóveis na faixa de 5%, 99% na faixa de 40% e maior

erro individual de 49,5%, pode ser considerado bom, mesmo porque, a amostra usada não parece ter sido coletada adequadamente, de forma aleatória.

Tabela 4.10: Percentual de valores preditos nas faixas de 5% a 40%, pelo modelo construído por Zancan (1995)

FAIXAS DE ERRO	N. DE VALORES PREDITOS	% DE VALORES PREDITOS
5%	92	23,2%
10%	180	45,3%
15%	264	66,5%
20%	310	78,1%
30%	367	92,4%
40%	387	97,5%
50%	393	99,0%

valores de predição: o poder de predição da equação ajustada para novas observações, foi realizado usando os 17 imóveis separados inicialmente, conforme citado no item 4.4. Os resultados da predição são mostrados na tabela 4.11.

Observa-se, na tabela 4.11, que os valores preditos para os valores dos apartamentos representam de fato o que acontece nas negociações do mercado imobiliário, onde diversos fatores subjetivos sempre estão influenciando na negociação do preço e causando as variações mostradas na tabela 4.11. Observe que o maior erro individual foi de 28,3% e o erro percentual médio foi de 11,9%. Pode-se considerar 11,9% uma margem de erro aceitável no mercado imobiliário. Isto permite concluir que a equação ajustada está bem, ou razoavelmente bem ajustada para predição de novas observações.

Tabela 4.11: Predições de novas observações pela equação ajustada

PR-NOVO OBSERVADO	PREDIÇÃO	ERRO	% ERRO
20.907,65	22.151,79	1244,135	5,6%
46.156,24	46.376,50	220,2566	0,5%
100.000,00	89.924,72	10075,28	11,2%
17.000,00	22.015,09	5015,091	22,8%
21.200,00	20.662,20	537,799	2,6%
21.877,73	29.804,59	7926,857	26,6%
12.500,00	17.443,62	4943,623	28,3%
37.383,18	37.477,54	94,36189	0,3%
23.000,00	21.693,62	1306,385	6,0%
18.000,00	24.591,82	6591,824	26,8%
130.525,00	118.373,60	12151,37	10,3%
120.000,00	128.055,00	8054,975	6,3%
33.503,65	33.223,69	279,9579	0,8%
93.988,92	99.895,27	5906,349	5,9%
25.000,00	27.614,47	2614,472	9,5%
66.038,62	58.333,59	7705,027	13,2%
34.000,00	27.183,15	6816,846	25,1%
ERRO PERCENTUAL MÉDIO DA PREDIÇÃO			11,9%

4.10 – CONSIDERAÇÕES

Assim terminam todas as investigações relativas às suposições do modelo, ou seja, análise de resíduos e inferências sobre a equação ajustada. Todos os cálculos realizados, desde o estudo das variáveis, determinação da equação e gráficos para as análises e interpretações foram feitas através do *Software Statistica* 6.0, com apoio do *Software Excel*.

A estratégia proposta neste trabalho é factível, como ilustrado em um caso real, obtendo um modelo melhor do que o modelo obtido por outro autor com o mesmo banco de dados.

Capítulo 5

CONSIDERAÇÕES FINAIS

5.1 – CONCLUSÕES

O valor de um imóvel pode ser representado por uma equação de regressão linear múltipla, desde que se disponha de um banco de dados formado por uma amostra do tipo aleatória com informações de preço e das principais características dos imóveis.

O fato do conjunto de dados ser composto por uma mescla de tipo de variáveis, isto é, quantitativas contínuas e discretas e qualitativas categorizadas, não inviabiliza a determinação de uma equação de regressão do tipo linear múltipla que forneça bons valores de predição.

A Análise Fatorial de Correspondências, devido a participação de variáveis qualitativas, pode ser usada como uma ferramenta de apoio para detectar as associações entre as variáveis e, assim, indicando quais variáveis são importantes para participarem da equação de regressão.

A multicolinearidade forte está, geralmente, presente nos estudos referentes a avaliação de imóveis. Nestes casos, a aplicação dos métodos usuais pode excluir, da equação ajustada, variáveis importantes para a predição. Assim, a procura de uma equação que proporcione uma avaliação do tipo “rigorosa” ou “rigorosa especial”,

segundo a norma NBR-5676/90 da ABNT, estaria mascarada, ou seja, não estaria proporcionando a avaliação com o rigor desejado.

A *Ridge Regression* é uma técnica indicada para evitar a eliminação de variáveis importantes quando existe multicolinearidade forte, garantir a qualidade de predição da equação ajustada e o rigor da avaliação.

Usando a abordagem da *Ridge Regression*, foi possível construir a seguinte equação de regressão linear múltipla para a cidade de Criciúma:

$$\begin{aligned}\textbf{Preço} = \textbf{exp}\ \{\ & 5{,}839751 - 0{,}383533 \bullet \textbf{RH1} - 0{,}298354 \bullet \textbf{RH2} - \\ & 0{,}201594 \bullet \textbf{RH3} - 0{,}121817 \bullet \textbf{RH4} - 0{,}330161 \bullet \textbf{CL1} - \\ & 0{,}183774 \bullet \textbf{CL2} - 0{,}239793 \bullet \textbf{CO1} - 0{,}166772 \bullet \textbf{CO2} - \\ & 0{,}126998 \bullet \textbf{CO3} + 0{,}051752 \bullet \textbf{DE} + 1{,}031288 \bullet \textbf{Ln(AT)}\ \}\end{aligned}$$

Onde:

- Preço representa o valor de venda do imóvel em dólares;
- DE representa a variável indicadora de dependência de empregada;
- Ln(AT) representa a área total com transformação logarítmica;
- RH1, RH2, RH3 e RH4 representam as regiões homogêneas nas categorias 1, 2, 3 e 4 respectivamente (RH5 foi usada como padrão);
- CL1 e CL2 representam as classificações nas categorias 1 e 2 respectivamente (CL3 foi usada como padrão);
- CO1, CO2 e CO3 representam os níveis de conservação 1, 2 e 3 respectivamente (CO4 foi usada como padrão).

Esta equação atende todas as suposições teóricas para sua existência. O valor de R^2 ajustado é de 96%, mostrando que a equação ajustada proporciona uma explicação de 96% da variação do preço dos apartamentos estudados. Portanto, esta pode ser considerada uma boa equação de regressão e pode ser usada para fazer predições dos preços de outros apartamentos da cidade de Criciúma, considerando as mesmas condições dos apartamentos analisados.

5.2 – SUGESTÕES PARA NOVAS PESQUISAS

Construir uma equação de regressão linear múltipla utilizando-se de uma nova amostra aleatória.

Determinar outros modelos que não lineares, como por exemplo um modelo de Regressão Linear Generalizado.

Aplicação da estratégia proposta em um caso que possa se ter a sua aplicação prática, comparando os valores preditos pelo modelo com os valores efetivos de venda. Isto permitirá avaliar a qualidade efetiva do modelo com objetivo de predição.

_____ REFERÊNCIAS

ABNT (Associação Brasileira de Normas Técnicas). *Avaliação de imóveis urbanos* (NBR 5676 e NBR 502). Rio de Janeiro: ABNT, 1989.

AYRES, Antonio. **Como Avaliar Imóveis**. São Paulo: Editora Imobiliária S/C Lmtda, 1996.

BARBOSA FILHO, D. S. In: *Técnicas Avançadas de Engenharia de Avaliações*. Caixa Econômica Federal, 1988.

BOUROCHE, J. M.; SAPORTA, G. **Análise de Dados**. Rio de Janeiro: Zahar, 1982.

BUNCHAFT, G.; KELLNER, S. R. O. **Estatística sem Mistérios**. 2ª ed. V II. Petrópolis: Editora Vozes, 1999.

CHARNET, R., FREIRE, C., CHARNET, E.; BONVINO, H. **Análise de Modelos de Regressão Linear com aplicações**. Campinas: Unicamp, 1999.

CHATTERJEE, S.; PRICE, B. **Regression Analysis by Example**. USA: John Wiley & Sons, 1977.

DANIEL, C.; WOOD, T. E. **Fitting Equations to Data**. New York: John Wiley & Sons, Inc, 1971.

DANTAS, Rubens A**. Engenharia de Avaliações – Introdução à Metodologia Científica**. São Paulo: Pini, 1998.

DRAPER, N. R. & SMITH, H. **Applied Regression Analysis**. New York: Jhon Wiley & Sons, Inc, 1981.

ELIAN, Silvia N. **Análise de Regressão**. São Paulo: IME,1988.

ESCOFIER, B.; PAGÈS, J. **Análisis Factoriales Simples y Múltiples**. Bilbao: Servicio Editorial de la Universidad del Pais Vasco, 1992.

FIKER, José. **Avaliação de Imóveis Urbanos**. 5 ª ed., São Paulo: Pini, 1997.

GOLDBERGER, A. S. **Teoria Econométrica**. Madrid: Editorial Técnos, 1970.

GONZÁLEZ, M. A. S.; FORMOSO, C. T. *Análise Conceitual das Dificuldades na Determinação de Modelos de Formação de Preços Através da Análise de Regressão.* Engenharia Civil – UM, 8: 65-75, 2000.

HILL, R. C.; GRIFFITHS, W. E.; JUDGE, G. G. **Econometria**. São Paulo: Editora Saraiva, 1999.

HOFFMANN, Rodolfo; VIEIRA, Sônia**. Análise de Regressão: uma introdução à econometria**. São Paulo: HUCITEC, 1983.

IBAPE-SP (Instituto Brasileiro de Avaliações e Perícias de Engenharia). *Engenharia de Avaliações.* São Paulo: Pini, 1974.

JOHNSTON, John. **Métodos Econométricos**. São Paulo: Atlas, 1974.

JUDGE, G. G.; GRIFFITHS, E. W.; HILL, R. C.; LEE, T**. The Theory and Practice of Econometrics**. New York: John Wiley & Sons,1980.

JUDGE, G. G.; HILL, R. C.; GRIFFITHS, E. W.; LUTKEPOHL, H.; LEE, T. **Introduction to the Theory and Practice of Econometrics**. 2ª ed. USA: John Wiley & Sons, 1988.

KMENTA, Jan. **Elementos de Econometria**. São Paulo: Editora Atlas S/A, 1978.

MACEDO, P. B. R. *Hedonic Price Models with Spatial Effects: na Aplication to the Housing Market of Belo Horizonte, Brasil.* RBE Rio de Janeiro, 52(1): 63-81, Jan./Mar. 1998.

MATOS, Orlando C. de. **Econometria Básica**. São Paulo: Atlas S/A, 2000.

MONTGOMERY, Douglas C. **Design and Analysis of Experiments**. 4ª ed. USA: John Wiley, 1997.

MOREIRA, A. L. **Princípios de Engenharia de Avaliação**. São Paulo, Ed. Pini, 1990.

MOREIRA FILHO, I. I.;FRAINER, J. I.;MOREIRA, R. M. I.; MOREIRA, R. M. I. **Avaliação de Bens por Estatística Inferencial e Regressões Múltiplas**. Porto Alegre: Avalien, 1993.

MOSCOVITCH, S. K. *Qualidade de vida urbana e valores de imóveis: um estudo de caso para Belo Horizonte*. Nova Economia, número especial: 247-279, 1997.

NETER, J.; WASSERMAN, W. **Applied Linear Statistical Models**. Richard D. Irwin, Inc, Illinois, 1974.

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

PEREIRA, Júlio C. R. **Análise de Dados Qualitativos**. 3ª ed. São Paulo: Edusp, 2001.

REIS, Elizabeth. **Estatística Multivariada Aplicada**. Lisboa: Silabo, 1997.

ROSEN, S. *Hedonic prices and implicit markets: product differentiation in pure competition*. Journal of Political Economy, 22: 34-55, 1974.

SMITH, L. B.; ROSEN, K. T.; FALIS, G. *Recent development in economic models of housing markets.* Journal of Economic Literature, 26: 29-64, 1988.

SNEDECOR, G. W.; COCHRAN, W. G. **Statistical Methods**. 6ª ed., Iowa: Ames, 1972.

TRIVELLONI, Carlos A. P. *Metodologia para Avaliação em Massa de Apartamentos por Inferência Estatística e Técnicas de Análise Multivariada.* Dissertação de Mestrado. Curso de Pós Graduação em Engenharia de Produção, Universidade Federal de Santa Catarina. Florianópolis, 1998.

WONNACOTT, Ronald J.; WONNACOTT, Thomas H. **Econometria**. 2ª ed., Rio de Janeiro: Livros Técnicos e Científicos Editora, 1978.

WORZALA, E.; LENK M.; SILVA A. *An exploration of neural networks and it6s application to real estate valuation.* The Journal of Real Estate Research, 10 (2): 185-201, 1995.

ZANCAN, Evelise C. *Metodologia para Avaliação em Massa de Imóveis para Efeito de Cobrança de Tributos Municipais – Caso de Apartamentos da Cidade de Criciúma, Santa Catarina*. Dissertação de Mestrado. Curso de Pós-Graduação em Engenharia de Produção. Universidade Federal de Santa Catarina. Florianópolis, 1995.