Modelagem Estatística para Análise de Dados Imobiliários Completos e com Censura à Esquerda

AMANDA CRISTINA ESTEVAM

Universidade Federal de São Carlos
Centro de Ciências Exatas e de Tecnologia
Departamento de Estatística

Modelagem Estatística para Análise de Dados Imobiliários Completos e com Censura à Esquerda.

AMANDA CRISTINA ESTEVAM

ORIENTADORES
PROFª. DRª. VERA LÚCIA D. TOMAZELLA
PROF. DR. FRANCISCO LOUZADA NETO

Dissertação apresentada ao Departamento de Estatística da Universidade Federal de São Carlos - DEs/UFSCar, como parte dos requisitos para obtenção do título de Mestre em Estatística.

UFSCar - São Carlos/SP
Julho/2014

Ficha catalográfica elaborada pelo DePT da Biblioteca Comunitária da UFSCar

E79me

Estevam, Amanda Cristina.
Modelagem estatística para análise de dados imobiliários completos e com censura à esquerda / Amanda Cristina Estevam. -- São Carlos : UFSCar, 2014.
104 p.

Dissertação (Mestrado) -- Universidade Federal de São Carlos, 2014.

1. Estatística. 2. Modelos lineares (Estatística). 3. Seleção de modelos. 4. Software – GAMLSS - estatística. 5. Influência local. I. Título.

CDD: 519.5 (20ª)

UNIVERSIDADE FEDERAL DE SÃO CARLOS
Centro de Ciências Exatas e de Tecnologia
Programa de Pós-Graduação em Estatística
Via Washington Luís, Km 235 - C.P.676 - CGC 45358058/0001-40
FONE: (016) 3351-8292 – Email: ppgest@ufscar.br
13565-905 - SÃO CARLOS-SP - BRASIL

FOLHA DE APROVAÇÃO

Aluno(a) : Amanda Cristina Estevam

DISSERTAÇÃO DE MESTRADO DEFENDIDA E APROVADA EM 01/04/2014
PELA COMISSÃO JULGADORA:

Presidente _____
Profa. Dra. Vera Lucia Damasceno Tomazella (DEs-UFSCar/Orientadora)

1º Examinador _____
Prof. Dr. Francisco Louzada Neto (ICMC-USP/ Co-orientador)

2º Examinador _____
Ms. Guilherme Moraes Ferraudo (MONSANTO)

3º Examinador _____
Prof. Dr. Luis A. Milan (DEs-UFSCar)

”O saber a gente aprende com os mestres e os livros. A sabedoria, se aprende é com a vida e com os humildes”.
(Cora Coralina)

Aos meus pais,
Adriana Cristina Caporusso Estevam e **Carlos Roberto Estevam**,
pelo apoio, incentivo e carinho incondicionais.

Agradecimentos

A Deus por me dar sabedoria e força para concluir mais essa etapa da minha vida.

Aos meus pais, por serem exemplo de luta, força e caráter, sem dúvidas jamais teria conseguido chegar até aqui sem vocês. Muito obrigada por todos os conselhos, carinho, dedicação, companheirismo, proteção e principalmente por acreditarem em mim quando nem eu mesma era capaz.

Ao meu primeiro amigo e irmão, André, pois sei que não importa o que acontecer, sempre estaremos lá um para o outro.

Às minhas amigas guaribenses, Cinthia, Sâmia e Tati pela amizade de longa data e por sempre estarem prontas para aquele abraço apertado capaz de me revigorar e dar forças para continuar a lutar.

À minha família Rio Clarense, começando pelas minhas companheiras da rep DDA's, Gi e Welen, que mesmo nos meus momentos mais azedos estavam sempre lá sorrindo e me ajudando a superar as dificuldades de morar longe de casa. Às minhas amigas e companheiras de estudos Daiane e Carol e Bianca que com certeza, sem elas jamais teria conseguido finalizar minha graduação e consequentemente este mestrado. Ao meu amigo Jean, uma das pessoas mais peculiares que já conheci, sempre feliz e com uma sabedoria e humildade admiráveis.

À Luana, Eliete, Tica, Gabi, Carol e Dezão pelos inúmeros desabafos, fins de semana repletos de guloseimas, descontração e as mais diversas aventuras, cada um de vocês tem um lugar especial no meu coração. À She-ra que mesmo longe sempre se fez presente me incentivando e me dando forças para não desistir. Muito Obrigada!

Aos meus colegas de mestrado por todo o apoio, pelas inúmeras horas de estudos aos sábados e domingos, pelas conversas e conhecimentos trocados durante esse tempo, só tenho a agradecer, pois sem vocês não teria chegado nem ao final do primeiro semestre.

A todos que conheci e que de alguma maneira fizeram da minha estadia em São Carlos mais feliz e descontraída. Sejam nos churrascos das republicas Gato Preto, Veia Loca e Safra Boa, nas festas entre os períodos de provas ou nos inúmeros jogos universitários.

Às minhas companheiras de casa, Lívia, Maíra, Jéssica, Marina e Natália que também marcaram bons momentos da minha história e em especial aos meus queridos amigos Marta e Gil por toda a compatibilidade da nossa amizade.

Gostaria também de agradecer ao meu amigo, companheiro e namorado André, por me apoiar, ajudar e me mimar nos momentos bons e principalmente nos momentos ruins, sempre sabendo o que fazer e o que falar para me incentivar e me ajudar a superar os obstáculos.

Aos amigos que conquistei durante toda a minha vida, tendo ou não mantido contato, meu muito obrigada pelos momentos compartilhados, se hoje sou o que sou foi por cada um desses momentos.

Aos meus Orientadores, Vera e Neto, por acreditarem em mim, por todos os conselhos, dedicação e conhecimentos adquiridos ao longo dessa jornada. Pela paciência e compreensão com as minhas dificuldades e obrigada principalmente por me auxiliarem na realização deste trabalho.

Agradeço também ao prof. Milan e ao Guilherme por aceitarem participar desta banca, pelas sugestões e por toda a ajuda dedicada ao longo do desenvolvimento dessa dissertação.

À Isabel por toda a sua competência, sempre alegre e disposta a tirar todas as dúvidas e solucionar quaisquer problemas e imprevistos que surgiam. Aos demais professores e funcionários de pós-graduação pela oportunidade de aprender o novo e pelo excelente convívio.

À Capes pelo apoio financeiro para a realização deste trabalho.

Resumo

O mercado imobiliário possui um papel fundamental na economia do país e municípios atraindo diversos estudos e pesquisas que buscam explicar e interpretar as inúmeras transações realizadas, e principalmente, encontrar maneiras adequadas de determinar seu valor monetário. Geralmente a modelagem de dados imobiliários é feita por meio de modelos de regressão, especialmente os lineares e também, os modelos lineares generalizados (Nelder e Wedderburn,1972). Por se tratarem de dados com diferentes características, como heterocedasticidade, não normalidade e heterogeneidade, o uso desses modelos podem sofrer limitações, por isso torna-se adequada a utilização de modelos cada vez mais complexos, como por exemplo, os modelos aditivos generalizados para posição, escala e forma (*GAMLSS*) propostos por *Rigby & Stasinopoulos* (2005), que permitem que todas as estimativas dos parâmetros envolvidos no modelo sejam obtidas de forma paramétrica ou não-paramétrica. Neste contexto e com base em um conjunto de dados de lotes urbanos da cidade de São Carlos do ano de 2005 foi estimado a função empírica do valor de lotes abordando a classe de modelos lineares, modelos lineares generalizados e o *GAMLSS*. Alternativamente, considerando a existência de dois tipos de preços de imóveis: já vendidos (observados) e anunciados (censurados), foi proposto aos dados, a utilização da análise de sobrevivência considerando censura à esquerda e o *GAMLSS* no processo de estimação dos parâmetros. Foi realizado também um estudo de simulação e um estudo de influência local.

Palavras-chave: Engenharia de Avaliação, Modelos de Avaliação em Massa, Modelos Lineares Generalizados, GAMLSS, Censura à Esquerda, Influência Local.

Abstract

The real estate market has a key role in the country and counties economy attracting several studies and researches that explains and interpret the numerous transactions performed, and especially to find appropriate ways to define the monetary value. Usually the real estate data modeling is performed through regression models, especially the linear and also the generalized linear models (Nelder and Wedderburn , 1972) . Because these data has different characteristics such as heteroscedasticity , non-normality and heterogeneity , the use of these models can suffer limitations , so it is appropriate to use more and more complex models , such as generalized additive models for location , scale and shape *GAMLSS* (proposed by Rigby & Stasinopoulos (2005) , that allows all parameters of the response variable are modeled parametric or non parametric form . In this context and based on a dataset of urban land of São Carlos city in 2005 was estimated the empirical function the value of the land addressing the class of linear models , generalized linear models and the *GAMLSS*. Alternatively, considering the existence of two types of real estate prices: already sold (observed) and announced (censored), was proposed to the data, using the survival analysis considering censored left and the *GAMLSS* in the parameter estimation process. A simulation study and a study of local influence was also performed.

Keywords: Engineering Appraisal, Mass Valuation Models General Linear Models, GAMLSS, Local Influence, Censored Left.

Sumário

Lista de Figuras

Lista de Tabelas

Capítulo

1

Introdução

1.1 Motivação

Nos últimos anos, o Brasil vem apresentando um crescimento significativo no mercado de imóveis, proporcionado principalmente pelas políticas de incentivo ao crédito imobiliário e diversos programas habitacionais oferecidos pelo governo e instituições financeiras.

Ao atribuir um valor ao imóvel, é possível melhorar o direcionamento de políticas tributárias, os ajustes fiscais, regularização imobiliária, disputas judiciais, atualização patrimonial, compra e venda, entre outras finalidades. As diversas transações e a expressiva quantidade de recursos envolvidos fazem com que este mercado esteja fortemente relacionado com o progresso de qualquer país, por isso a importância do mesmo. No âmbito social, a habitação proporciona *status*, conforto e segurança para o indivíduo, sendo um dos bens mais importantes adquiridos ao longo da vida, o mais caro, e o de maior vida útil. Porém, por se tratar de um bem com características específicas como, por exemplo, segurança, rentabilidade, diversidade e volatilidade, e apresentar um comportamento diferenciado dos mercados de outros bens, o mercado imobiliário é considerado um dos setores mais complexos da economia.

Desta maneira, a adequada determinação técnica do valor de um imóvel é fundamental para a tomada de decisão e alternativas de investimentos em diversos segmentos da sociedade e em muitos órgãos governamentais e privados. Com o mercado brasileiro aquecido e o avanço constante desde 2005, diversos estudos e pesquisas buscam explicar e interpretar as diversas transações e principalmente encontrar maneiras adequadas de determinar seu valor monetário. Assim, cabe a Engenharia de Avaliações, enquanto ciência do valor, responder questões importantes como quais são as preferências do mercado, quais variáveis interferem na formação do

preço, quanto custa produzir o bem avaliado e principalmente estipular através de técnicas e conhecimentos específicos de diversas áreas da ciência, um valor ao imóvel ou de um direito sobre ele.

Os primeiros trabalhos sobre técnicas de engenharia de avaliação no Brasil datam do início do século XX, e em 1952, com o departamento da Caixa Econômica Federal, foi criada a primeira norma sobre avaliação de imóveis. Os trabalhos publicados na época eram baseados na utilização de fatores de homogeneização determinísticos e em fórmulas empíricas que não davam muita segurança aos avaliadores. Em 1974, com o engenheiro Domingos de Saboya Barbosa Filho, surgiu a Metodologia de Pesquisa Científica aplicada à Engenharia de Avaliação com o intuito de substituir as técnicas usadas até então. O maior desenvolvimento da engenharia de avaliações se deu na década de 90, com a introdução da metodologia científica no trabalho avaliatório.

De acordo com Dantas (2005), a aplicação da metodologia mais adequada para realização de um trabalho avaliatório depende fundamentalmente das condições mercadológicas com que se defronta o avaliador, pelas informações coletadas neste mercado, bem como pela natureza do serviço que se pretende desenvolver. As avaliações devem ser realizadas com base em normas técnicas da ABNT-Associação Brasileira de Normas Técnicas.

Para a formação de preços, utiliza-se principalmente a metodologia de regressão múltipla e a localização é considerada um dos atributos mais importantes na modelagem de valorização de imóveis, porém, não é quantificada diretamente, mas através de variáveis *proxy*, como a fixação espacial (imobilidade), a acessibilidade, as características da vizinhança, renda média e distância ao centro comercial-cultural (González & Formoso, 2000). Dessa forma, imóveis próximos com características semelhantes, possuem valores influenciados pelas construções ao redor. O imóvel apresenta uma vida útil considerável, e ao longo de sua existência seu preço pode sofrer diversas modificações, uma vez que os aspectos espaciais e as transformações do ambiente, como construções de escolas, avenidas e indústrias são fundamentais para sua valorização ou desvalorização. Para aprofundamento de técnicas, definições e métodos utilizados em engenharia de avaliações ver Dantas (2005).

Na literatura é possível encontrar vários trabalhos utilizando modelos de preços hedônicos[1]. González & Formoso (2000), por exemplo, fazem uma excelente revisão literária das dificuldades da determinação de modelos de formação de preços através da análise de regressão. Nota-se também, que na maioria dos estudos envolvendo a precificação de imóveis, a regressão linear é abordada, como o caso de Aguirre & Macedo (1996) que adota a transformação Box-Cox para a variável resposta considerando o mercado imobiliário de Belo Horizonte. Biderman (2001) que avalia a demanda por imóveis novos em São Paulo e Angelo & Fávero (2003), que determi-

[1]Hedônico: palavra de origem grega que significa prazer, ou seja, o prazer ao se adquirir um bem está relacionado intimamente com os atributos nele contido. A teoria de preços hedônicos, para a formação do preço de um bem, considera quais atributos são mais relevantes, utilizando-se da análise de regressão, na qual os preços dos produtos são regredidos em função de suas características. Foi criada por Lancaster (1966) e consolidada por Rosen (1974), o qual a introduziu pela primeira vez em um contexto de mercado. (Ângelo & Fávero (2003), González & Formoso (2000), Florencio (2010))

nam o preço de apartamentos paulistanos através da faixa de renda utilizando-se da regressão log-linear.

Trabalhos utilizando modelos lineares generalizados também foram propostos, por exemplo, em Dantas & Cordeiro ((a) 1988 e (b) 2001) que empregam técnicas de validação cruzada e em Barbosa & Bidurin (1991), que recomendam as distribuições gama ou lognormal para o conjunto analisado.

Já algumas das utilizações de técnicas semi-paramétricas e não paramétricas podem ser encontradas em Anglin & Gencay (1996) e Gencay & Yang (1996) que usam dados de serviços de listagem de Windsor, Canada e em Martins-Filho e Bin (2003), o qual evidencia a vantagem dos modelos não paramétricos considerando a estimação do valor de comercialização de casas de Multnomah County, Oregon-USA. Na literatura brasileira algumas poucas pesquisas envolvem essa técnica, podendo ser encontradas apenas em Neto (2006) que utilizou redes neurais artificiais, em Florencio (2010) que abordou a metodologia GAMLSS em dados do mercado imobiliário de Aracaju-SE, para enfatizar a superioridade dos modelos semi-paramétricos com relação aos paramétricos na valorização dos bens imobiliários e em Florencio (2012) que incorpora os efeitos da correlação espacial nos modelos *GAMLSS* em daados de Aracaju-SE.

Como mencionado por Florencio (2010), a estimação da equação hedônica não é trivial, visto que a teoria não determina sua forma funcional nem as variáveis relevantes para a sua estimação. Com todos os obstáculos das técnicas de precificação dos imóveis e o avanço tecnológico dos computadores, a busca por modelos cada vez mais complexos e realistas se faz necessário.

Assim, visto que na literatura brasileira, a utilização de modelos não paramétricos e a não consideração da diferença entre os preços dos imóveis ofertados e vendidos é escassa, acredita-se que a utilização dos modelos aditivos generalizados para posição, escala e forma (*GAMLSS*) levando em consideração a censura à esquerda existente na avaliação dos imóveis ofertados possa trazer resultados satisfatórios nos dados em estudo.

1.2 Objetivos

As técnicas convencionais para a avaliação imobiliária enfrentam problemas especialmente pelo desconhecimento da forma funcional que descreve a relação entre os valores e devido às particularidades existentes nos dados, afetando diretamente a construção de modelos que estimem de forma satisfatória os preços imobiliários. Considerando ainda, que os dados em estudo são referentes ao ano de 2005 e o mercado imobiliário é bastante dinâmico, o qual é influenciado pela situação econômica do momento, pela oferta e procura, e subjetividade das partes envolvidas que na maioria das vezes, expressa puramente o interesse dos vendedores, torna-se ainda mais difícil encontrar resultados realistas.

Assim, este trabalho tem por objetivo principal a criação de uma equação representativa da precificação de lotes urbanos da cidade de São Carlos.[2] O estudo é realizado em duas partes. A primeira consiste em abordar modelos cuja transação já tenha sido efetuada utilizando o modelo de regressão linear, o modelo linear generalizado e os modelos aditivos generalizados para posição, escala e forma. Já na segunda parte, os imóveis em oferta também são estudados, caracterizando os dados como censurados à esquerda. Nesta última etapa é novamente utilizada a metodologia *GAMLSS*. O interesse do trabalho está em verificar qual modelo, do ponto de vista estatístico, apresenta resultados mais precisos e realistas.

1.3 Organização

O presente trabalho está dividido em seis capítulos. No Capítulo 2 é apresentado uma revisão bibliográfica das metodologias utilizadas na análise de dados imobiliários, entre elas encontram-se o modelo de regressão linear, os modelos lineares generalizados (*Nelder & Wedderburn*, 1972) e a teoria dos modelos aditivos generalizados para posição escala e forma (GAMLSS) propostos por *Rigby & Stasinopoulos* (2005). Também será abordada a teoria de influência local proposta por Cook (1986) para modelos lineares generalizados.

Com o objetivo de verificar a diferença entre o valor do imóvel vendido e ofertado, no Capítulo 3 encontram-se alguns dos principais conceitos de análise de sobrevivência. Para a modelagem dos dados censurados foi apresentada a distribuição Weibull e o modelo de regressão Weibull.

No Capítulo 4, para testar as propriedades assintóticas dos estimadores, é realizado um estudo de simulação utilizando os modelos lineares, os lineares generalizados e os modelos aditivos generalizados para posição, escala e forma com e sem censura .

No Capítulo 5 as metodologias apresentadas nos Capítulos 2 e 3 são aplicadas a um conjunto de dados de lotes urbanos da cidade de São Carlos, interior do estado de São Paulo, no ano de 2005. Finalmente no Capítulo 6 encontram-se as conclusões e propostas futuras. No Apêndice encontra-se o código no *R* para o ajuste com censura utilizando a classe de modelos *GAMLSS*.

[2]Fundada em 1857, localiza-se no interior do estado de São Paulo e conta com uma população flutuante de mais de vinte mil graduandos e pós-graduandos (São Carlos, 2013), intensificando ainda mais a atuação do setor imobiliário.

Capítulo

2

Metodos Estatísticos via Regressão Linear, Modelos Lineares Generalizados e GAMLSS

2.1 Introdução

A metodologia científica na avaliação territorial não está relacionada simplesmente com a coleta de dados, mas com o processo avaliatório como um todo, no qual é possível definir as seguintes etapas:

1. Conhecimento do objeto da pesquisa;
2. Planejamento (preparação da pesquisa);
3. Trabalho de Campo (coleta de dados);
4. Processamento e análise dos dados;
5. Interpretação e explicação dos resultados (o modelo);
6. Redação do relatório de pesquisa (o laudo de avaliação).

Quando a Engenharia de Avaliações segue todas essas fases e é capaz de encontrar modelos que expliquem, de maneira satisfatória, a variabilidade observada nos preços do mercado que se estuda, pode ser considerada como uma ciência: a ciência do Valor.

O presente capítulo tem por objetivo definir e explicar brevemente algumas das possíveis formas de modelar dados imobiliários, ou seja, encontrar a relação entre os preços que são

praticados no mercado e as diversas características que influenciam na formação dos mesmos através da teoria de regressão.[1]

2.2 Modelo de Regressão Linear (MRL)

Um modelo de regressão é uma ferramenta estatística que utiliza o relacionamento existente entre duas ou mais variáveis de maneira que uma delas possa ser descrita ou o seu valor estimado a partir das demais. O modelo linear considera que essa relação é feita a partir de uma função linear. O caso mais simples é quando se trabalha com uma única variável independente X e a variável resposta Y e denomina-se **Modelo de Regressão Linear Simples**, o qual é representado por:

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \tag{2.1}$$

com $i = 1, ..., n$ onde Y_i é o valor da variável resposta da i-ésima observação, x_i corresponde ao valor da variável explicativa da i-ésima observação, β_0 é o intercepto, β_1 é o coeficiente angular, os quais são parâmetros desconhecidos que serão estimados a partir da amostra, e ε_i é o erro.

Algumas considerações sobre o erro:

- Distribuição não especificada: ε_i é o termo de erro aleatório com $E(\varepsilon_i) = 0$; $Var(\varepsilon_i) = \sigma^2$;
- Erros com distribuição Normal: acrescenta-se às suposições acima que os erros tem distribuição normal e são independentes;
- $E(Y_i) = \beta_0 + \beta_1 x_i$ e $Var(Y_i) = \sigma^2$.

Quando mais de uma variável independente é necessária para explicar a variabilidade dos preços, generaliza-se o caso linear simples e tem-se a **Regressão Linear Múltipla (Matricial)** dada por:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{2.2}$$

$$\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \dots & x_{1p-1} \\ 1 & x_{21} & \dots & x_{2p-i} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & \dots & x_{np-1} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}, \tag{2.3}$$

onde $\boldsymbol{Y}$ é a variável dependente, composta de um vetor (n $\times$ 1) de observações tomadas em cada uma das n áreas; X é uma matriz (n $\times$ p) com (p-1) variáveis explicativas tomadas também nas n áreas; $\boldsymbol{\beta} = (\beta_1, ...\beta_p)^T$ é um vetor (p$\times$ 1) com os coeficientes de regressão e $\boldsymbol{\varepsilon}$ é um vetor (n $\times$ 1) de erros aleatórios.

[1] A análise de regressão originou-se com Gauss em trabalhos de astronomia no período de 1809 a 1821.

Como hipótese tem-se que os erros ε_i do modelo são: independentes, não correlacionados com a variável dependente, tem variância constante e apresentam distribuição normal com média zero. Do ponto de vista de avaliação de imóveis em massa, esta hipótese é irrealista.

2.2.1 Estimação dos parâmetros

2.2.1.1 Método dos Mínimos Quadrados

Na regressão linear, o método mais comum para a estimação dos parâmetros é o método dos mínimos quadrados (MMQ). O MMQ busca encontrar os valores dos $\beta^{'}s$ minimizando a soma dos quadrados dos erros (SQE).

O **caso simples** é dado por:

$$\text{SQE} = \sum \epsilon_i^2 = \sum (Y_i - \beta_0 - \beta_1 x_i)^2. \tag{2.4}$$

Para minimizar 2.4, basta derivá-la em relação aos parâmetros β_0 e β_1, igualar as derivadas parciais a zero e resolver o sistema de equações. Os estimadores dos mínimos quadrados $\widehat{\beta_0}$ e $\widehat{\beta_1}$, são então dados por:

$$\widehat{\beta_0} = \overline{Y} - \widehat{\beta_1}\overline{x}, \tag{2.5}$$

$$\widehat{\beta_1} = \frac{\sum_{i=1}^{n}(x - \overline{x})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(x - \overline{x})^2}, \tag{2.6}$$

sendo $\overline{Y}$ e $\overline{x}$ as médias amostrais de $\boldsymbol{Y}$ e $\boldsymbol{x}$.

Para o **caso múltiplo**,

$$\text{SQE} = \sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2} - ... - \beta_p x_{ip})^2. \tag{2.7}$$

Na forma matricial tem-se:

$$\begin{aligned}\text{SQE} &= \sum_{i=1}^{n} \varepsilon_i^2 = (\boldsymbol{Y} - \boldsymbol{X\beta})^T(\boldsymbol{Y} - \boldsymbol{X\beta}) \\ &= \boldsymbol{Y}^T\boldsymbol{Y} - \boldsymbol{Y}^T\boldsymbol{X\beta} - \boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{Y} + \beta^T\boldsymbol{X}^T\boldsymbol{X\beta},\end{aligned} \tag{2.8}$$

com $\boldsymbol{Y}^T\boldsymbol{X\beta} = \boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{Y}$, pois o produto resulta em um escalar. Derivando SQE em relação a $\boldsymbol{\beta}$,

$$\frac{\partial SQE}{\partial \boldsymbol{\beta}} = -2\boldsymbol{X}^T\boldsymbol{Y} + 2\boldsymbol{X}^T\boldsymbol{X\beta}.$$

Igualando a zero e substituindo $\boldsymbol{\beta}$ por $\widehat{\boldsymbol{\beta}}$, tem-se

$$(\boldsymbol{X^T}X)\widehat{\boldsymbol{\beta}} = \boldsymbol{X}^T\boldsymbol{Y}.$$

Como a matriz $(\boldsymbol{X^T X})$ é inversível, ou seja, não singular pois seu determinante é diferente de zero, conclui-se que

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}. \tag{2.9}$$

Logo, tem-se que o modelo ajustado é:

$$\widehat{\boldsymbol{Y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^TY = \boldsymbol{H}\boldsymbol{Y}, \tag{2.10}$$

onde $\boldsymbol{H}$ é denominada *matriz do ajuste*, a qual transforma o vetor de respostas $\boldsymbol{Y}$ no vetor de valores ajustados $\widehat{\boldsymbol{Y}}$. A matriz do ajuste é idempotente, simétrica e possui posto completo.

Propriedades dos estimadores dos mínimos quadrados

- $E[\widehat{\boldsymbol{\beta}}] = \boldsymbol{\beta}$;
- Assumindo que os ϵ_i não são correlacionados e possuem a mesma variância, ou seja, $cov[\epsilon_i, \epsilon_j] = \delta_{ij}\sigma^2$, então, $Var[\boldsymbol{\epsilon}] = \sigma^2\boldsymbol{I}$, e mais,

$$Var[\boldsymbol{Y}] = Var[\boldsymbol{Y} - \boldsymbol{X\beta}] = Var[\boldsymbol{\epsilon}]. \tag{2.11}$$

  Assim,

$$Var[\widehat{\boldsymbol{\beta}}] = \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}; \tag{2.12}$$

- A distribuição de $\boldsymbol{Y}$ é normal n-variada com $E(\boldsymbol{Y}) = \boldsymbol{X\beta}$ e $Cov(\boldsymbol{Y}) = \sigma^2\boldsymbol{I}$;
- A distribuição de $\widehat{\boldsymbol{\beta}}$ é normal p-variada com $E(\widehat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$ e $Cov(\widehat{\boldsymbol{\beta}}) = \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}$;
- O estimador não viciado para σ^2 é dado por: $\widehat{\sigma^2} = \frac{(\boldsymbol{Y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}})^T(\boldsymbol{Y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}})}{n-p} = \frac{SQR}{n-p} = QMR$ sendo SQR conhecido como *a soma de quadrado dos resíduos*, que mede a discrepância entre o vetor de observações e os valores ajustados e QMR é o *quadrado médio residual*;
- $\frac{SQR}{\sigma^2}$ tem distribuição qui-quadrado χ^2_{n-p} com n-p graus de liberdade.

2.2.1.2 Intervalo de Confiança

Para realizar testes de hipóteses e construir intervalos de confiança, as suposições são de que os erros são independentes e identicamente distribuídos. Considerando p covariáveis no modelo, sabe-se que $\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}$ possui distribuição $N_p(O, \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1})$ e portanto seu intervalo de confiança ao nível de $100(1-\alpha)\%$ para a média $\mu(\boldsymbol{a})$, com $\boldsymbol{a} \in R^P$ é dado por:

$$\boldsymbol{a}^T\widehat{\boldsymbol{\beta}} \pm \sigma\sqrt{c_\alpha}\left\{\boldsymbol{a}^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{a}\right\}, \tag{2.13}$$

para qualquer $\boldsymbol{a} \in R^p$, onde c_α é tal que $P(\chi^2 \leq c_\alpha) = 1 - \alpha$.

Esse intervalo de confiança é um intervalo com $100(1-\alpha)\%$ de confiança individualmente, porém não é simultâneo. Para solucionar esse problema foram propostos alguns intervalos de confiança simultâneos como os de Bonferroni, Máximo Módulo e Scheffé. Para maiores informações sobre estes intervalos consultar Demétrio e Zocchi (2007).

2.2.1.3 Teste de Hipótese

Os testes de hipóteses são usados para verificar se o ajuste do modelo proposto aos dados é adequado, mais ainda, para verificar a significância da regressão. Para isso, testa-se as hipóteses $H_0 : \beta_1 = \beta_2 = ... = \beta_{p-1} = 0$ versus $H_a = \beta_i \neq 0$ para pelo menos um $i = 1, ..., p-1$. Um dos métodos usados para testar essa significância é a *análise de variância (ANOVA)*, a qual decompõe a variabilidade total de Y em dois componentes e essa variação pode ser medida por meio de somas de quadrados. A partir dessas informações, tem-se:

$$\sum_{i=1}^{n}(Y_i - \overline{Y}_i)^2 = \sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2 + \sum_{i=1}^{n}(\widehat{Y}_i - \overline{Y}_i)^2, \tag{2.14}$$

ou seja,

$$SQT = SQE + SQR, \tag{2.15}$$

sendo SQT a soma de quadrados total, SQR a soma de quadrados da regressão que representa a variação em $\boldsymbol{Y}$ explicada pela regressão ajustada e SQE a soma de quadrados dos resíduos que é a variação não explicada pela regressão.

A análise de variância consiste em testar as hipóteses H_0 e H_a utilizando a estatística teste:

$$F_{\text{calculado}} = \frac{\text{SQR}/(p)}{\text{SQE}/(n-p-1)} = \frac{\text{QMR}}{\text{QME}} \sim F_{(p,n-p-1)}, \tag{2.16}$$

onde $\text{QMR} = \text{SQR}/p$ e $\text{QME} = \text{SQE}/(n-p-1)$, denominados *Quadrado Médio da Regressão* e *Quadrado Médio dos Resíduos* respectivamente.

Dessa maneira, se $F_{\text{calculado}} > F(p, n-p-1, 1-\alpha)$, rejeita-se a hipótese nula de H_0 ao nível de $100(1-\alpha)\%$ de confiança ou de $100\alpha\%$ escolhido.

2.2.2 Seleção do modelo

Além de utilizar o maior R^2, a seleção de modelos pode ser feita através do AIC, *forward*, *stepwise*, *backward*, entre outras. A seguir algumas delas serão brevemente comentadas.

Forward

O modelo inicial é composto apenas pelo intercepto e adiciona-se uma variável por vez. A primeira variável selecionada é aquela com maior correlação com a resposta. Supondo que essa

variável seja x_1, calcula-se a estatística F. Seja F_a o menor nível crítico calculado para um dado valor α crítico. A variável entra no modelo se a estatística F for maior do que F_a.

Considerando que x_1 foi selecionado para o modelo, encontra-se uma variável com maior correlação com a resposta considerando a presença da primeira variável no modelo. Supondo que a maior correlação parcial com y seja x_2, se o valor da estatística for maior do que F_a, x_2 é selecionado para o modelo. O processo é repetido até que não seja incluída nenhuma variável no modelo.

Backward:

Considera-se inicialmente o modelo com todas as variáveis e por etapas elimina ou não cada uma delas. A decisão de retirada da variável é tomada baseando-se em testes F parciais, que são calculados para cada variável como se ela fosse a última a entrar no modelo. Para cada variável explicativa calcula-se a estatística F. O menor valor das estatísticas F parciais calculadas é então comparado com o F crítico, F_b, calculado para um dado valor α crítico. Se o menor valor encontrado for menor do que F_b, elimina-se do modelo a covariável responsável pelo menor valor da estatística F parcial.

Ajusta-se novamente o modelo, agora com as $p-1$ variáveis. As estatísticas F parciais são calculadas para esse modelo e o processo é repetido. Quando a menor estatística F parcial não for menor do que F_b termina-se a eliminação das variáveis.

Stepwise:

É uma mistura do forward com o backward. O modelo inicial é composto apenas pelo intercepto, após a inclusão de duas variáveis, verifica-se se a primeira permanece no modelo. O processo continua até que nenhuma variável seja retirada ou incluída no modelo.

Akaike (AIC)

Este método proposto por Akaike (1974) busca encontrar um modelo que seja parcimonioso, ou seja, que apresente um número reduzido de parâmetros sem perder o bom ajuste. O objetivo é encontrar o modelo com o menor valor para a função:

$$AIC = -2L(\widehat{\beta}) + 2p, \tag{2.17}$$

em que $L(\widehat{\beta})$ é o logaritmo da verossimilhança e p o número de parâmetros.

2.2.3 Teste de diagnóstico

Dentre as diversas etapas que se tem na escolha de um modelo de regressão, a análise de diagnóstico é utilizada para verificar a qualidade do modelo escolhido identificando, por exemplo, a existência de pontos discrepantes, a adequação da distribuição da variável resposta e a detecção de observações influentes.

Uma das propostas mais inovadoras na área de diagnóstico em regressão, segundo Paula (2004), foi apresentada por Cook (1986) que propõe avaliar a influência conjunta das observações sob pequenas perturbações no modelo ou nos dados, conhecida como influência local a qual será detalhada na seção 2.5.

Análise de resíduos

Para verificar se um modelo de regressão é adequado utiliza-se um conjunto de técnicas baseadas nos resíduos $\widehat{r}_i = y_i - \widehat{y}_i$, a análise de resíduos, a qual procura medir a discrepância entre o valor observado e o valor ajustado da i-ésima observação.

Para testar a normalidade dos resíduos utiliza-se testes como Shapiro-Wilk, Kolmogorov-Smirnov, entre outros, e caso os resíduos sigam uma distribuição normal é possivel ver graficamente através de um diagrama de dispersão, dos quantis dos resíduos e dos quantis de uma distribuição normal, aproximadamente uma reta.

A seguir serão apresentadas técnicas para detecção de outliers e de pontos influentes.

Verificação de *Outliers*

Para detectar *outliers* são utilizadas medidas como a diagonal principal da matriz do ajuste 2.10 e resíduos studentizados.

- A diagonal principal da matriz H, $(\widehat{h}_{ii})$, identifica outliers entre as covariáveis X. O vetor de observação i é considerado um *outlier* se $h_{ii} > 2p/n$.
- Os resíduos studentizados $\widehat{e}_i^* = e_i/\sqrt{\widehat{\sigma}^2(1-h_{ii})}$, onde $\widehat{\sigma}^2$ é a estimativa da variância, identificam *outliers* na í-ésima variável resposta. Na prática, a i-ésima observação é considerada *outlier* se $e_i^* > 3$.

Após a identificação de *outliers*, é importante verificar se são pontos influentes ou não.

Identificação de pontos influentes

Para pontos influentes a verificação é feita utilizando medidas como DFFITS, DFBETAS e D'Cook.

- A medida DFFITS $(DFFITS_i)$ é calculada para saber se a i-ésima observação é um ponto influente e é dada pela seguinte fórmula:

$$DFFITS_i = \frac{Y_i - \widehat{Y_{i(i)}}}{\sqrt{\widehat{\sigma}^2_{(i)}(1-hii)}} = \frac{d_i}{\sqrt{\widehat{\sigma}^2_{(i)}(1-h_{ii})}},$$

  onde $\widehat{Y_{i(i)}}$ é o i-ésimo valor ajustado calculado sem o i-ésimo vetor de observações, d_i conhecido como resíduo deletado, e $\widehat{\sigma}^2_{(i)}$ é a estimativa da variância calculada sem o i-ésimo vetor de observações. Na prática, a i-ésima observação é considerada ponto influente se $|DFFITS_i| > 2$

- O $DFBETAS_{k(i)}$ identifica se o i-ésimo vetor de observações está afetando o coeficiente de X_k e é dado pela seguinte fórmula:

$$DFBETAS_{k(i)} = \frac{\widehat{\boldsymbol{\beta}_k} - \widehat{\boldsymbol{\beta}}_{\boldsymbol{k(i)}}}{\sqrt{QME_{(i)}c_{kk}}}, \ \ k = 1, ..., p$$

onde c_{kk} é o k-ésimo elemento da diagonal de $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$. Na prática, a i-ésima observação é considerada ponto influente para $\widehat{\boldsymbol{\beta}_k}$ se $|DFBETAS_{k(i)}| > 2$

- A distância de Cook, ou D'Cook (D_i) mede a influência do vetor de observações i sobre todos os n valores ajustados $\widehat{Y}_i$. D_i é definido como:

$$\widehat{D}_i = \frac{(Y_i - \widehat{Y_{i(i)}})^T(Y_i - \widehat{Y_{i(i)}})}{p\widehat{\sigma}^2} = \frac{\widehat{e}_i^2}{p\widehat{\sigma}^2}\left[\frac{\widehat{h}_{ii}}{(1-\widehat{h}_{ii})^2}\right],$$

onde o último membro da igualdade tem distribuição $F_{(p,n-p)}$. Se $\widehat{D}_i > F_{(0,5,p,n-p)}$, ele é considerado um ponto influente.

Técnicas gráficas

Existem também técnicas gráficas que embora ditas informais por apresentarem subjetividade, são bastante utilizadas. As mais usadas são:

- gráficos dos resíduos padronizados r_i versus a ordem das observações para detectar observações aberrantes;
- gráficos dos resíduos padronizados versus os valores ajustados $\widehat{y}_i$, o qual indica homocedasticidade da variância dos erros se os pontos estiverem distribuídos aleatoriamente entre as duas retas $y = \pm 2$, paralelas ao eixo horizontal, sem exibir uma forma definida.

2.3 Modelos Lineares Generalizados (MLG)

Durante muitos anos os modelos lineares foram utilizados para descrever a maioria dos fenômenos aleatórios. Mesmo quando o fenômeno sob estudo não apresentava uma resposta para a qual fosse razoável a suposição de normalidade, tentava-se algum tipo de transformação no sentido de encontrar a normalidade procurada. (Paula, 2004)

Com o avanço da estatística e as dificuldades apresentadas na modelagem de dados imobiliários ao se utilizar a regressão linear, pois dados de imóveis muitas vezes não apresentam normalidade e homocedastidade, pressupostos necessários para a aplicação da teoria, a enganharia de avaliações, para solucionar esse problema, foi em busca de outros métodos de regressão, como por exemplo, os modelos lineares generalizados (MLGs) propostos por Nelder e Wedderburn (1972).

Os MLGs são uma extensão natural dos modelos de regressão linear e permitem aumentar as opções para a distribuição da variável resposta (as pertencentes à família exponencial) alem

de dar uma maior flexibilidade para a relação funcional entre a média da variável resposta e o preditor linear η. A seguir serão apresentados a definição, resultados relacionados a estimação, testes de hipóteses, intervalos de confiança e métodos de diagnósticos dos MLGs.

Família Exponencial

A classe da família exponencial foi proposta independentemente por Koopman, Pitman e Darmois em estudos de propriedades de suficiência estatística. Seu conceito foi introduzido por Fisher, no entanto, foi com o trabalho pioneiro de Nelder & Wedderburn em 1972 que teve destaque na área de regressão. A seguir sua definição.

Família Exponencial Uniparamétrica

A função de (probabilidade ou densidade) na forma canônica é dada pela fórmula:

$$f(x;\theta) = h(x)\text{exp}[\eta(\theta)t(x) - b(\theta)], \tag{2.18}$$

onde h(x), $\eta(\theta)$, t(x), $b(\theta)$ assumem valores reais e as funções $\eta(\theta)$, t(x), $b(\theta)$ não são únicas. Outra característica da família exponencial é que o suporte, $\{x, f(x,\theta) > 0\}$ não pode depender do parâmetro (θ), assim, a distribuição uniforme em $(0,\theta)$ não pertence à família exponencial.

Utilizando o teorema da fatoração de Neyman-Fisher, a estatística t(x) é suficiente para θ.

Família Exponencial Multiparamétrica

A família exponencial multiparamétrica de dimensão k é caracterizada por uma uma função de (probabilidade ou densidade) da forma:

$$f(x;\boldsymbol{\theta}) = h(x)\text{exp}[\sum_{i=1}^{k} \eta_i(\boldsymbol{\theta})t_i(x) - b(\boldsymbol{\theta})], \tag{2.19}$$

onde $\boldsymbol{\theta}$ é um vetor de parâmetros de dimensão k e as funções $h(x)$, $\eta_i(\boldsymbol{\theta})$, $t_i(x)$, $b(\boldsymbol{\theta})$ assumem valores reais. Pelo teorema da fatoração, o vetor $\boldsymbol{T} = [T_1(\boldsymbol{X}), ..., T_k(\boldsymbol{X})]^T$ é suficiente para o vetor de parâmetros $\boldsymbol{\theta}$.

Definição

Um MLG é definido por uma distribuição de probabilidade que pertence à família exponencial para a variável resposta (componente aleatório), um conjunto de variáveis independentes descrevendo a estrutura linear do modelo (componente sistemático) e uma função de ligação entre o componente aleatório e o sistemático. A análise de dados através do MLG é bastante flexível, pois para uma mesma estrutura linear pode-se obter vários modelos dependendo do componente aleatório e da função de ligação escolhida (Dantas, 2005).

Suponha $Y_1, ..., Y_n$ variáveis aleatórias independentes pertencentes à família exponencial, com médias $\mu_1, ..., \mu_n$, ou seja, $E(Y_i) = \mu_i$, $\phi > 0$ o parâmetro de dispersão e θ_i denominado parâmetro canônico. Então, a função de densidade de Y_i é dada por:

$$f(y;\theta_i,\phi) = \text{exp}[\phi^{-1}\{y\theta_i - b(\theta_i)\} + c(y,\phi)]. \tag{2.20}$$

Seja $V(\mu_i) = d\mu_i/d\theta$ a função de variância que depende apenas de μ_i. Então, $E(Y_i) = \mu_i = b^{'}(\theta_i)$ e $Var(Y_i) = \phi V(\mu_i) = \phi b^{''}(\theta_i)$.

O parâmetro natural θ_i pode ser expresso por:

$$\theta_i = \int V_i^{-1} d\mu_i = q(\mu_i). \tag{2.21}$$

A função de variância desempenha um papel importante na família exponencial, uma vez que a mesma caracteriza a distribuição, isto é, dada a função de variância tem-se uma classe de distribuições correspondentes e vice-versa. (Paula, 2004)

O **componente aleatório** de um MLG corresponde ao conjunto de variáveis aleatórias independentes $Y_1, ...Y_n$ com distribuição de probabilidade pertencente à família exponencial. A escolha da distribuição que será atribuída à variável resposta deve estar de acordo com a natureza dos dados tais como: natureza contínua ou discreta, intervalo de variação e assimetria.

O **componente sistemático** é definido por:

$$g(\boldsymbol{\mu}) = \boldsymbol{\eta},$$

em que $\boldsymbol{\eta} = \boldsymbol{X\beta}$ é o preditor linear, $\boldsymbol{X} = (x_1, ..., x_p)^T$ a matriz das covariáveis, $\boldsymbol{\beta} = (\beta_1, ...\beta_p)^T$, $p < n$ é um vetor de parâmetros desconhecidos a serem estimados e g(.) é uma função monótona e diferenciável que relaciona o componente aleatório ao sistemático conhecida como **função de ligação**.

Distribuição Gama

Para a variável aleatória $Y \sim \text{Ga}(\nu, \nu/\mu)$, com distribuição gama e parâmetros de forma ν e de escala ν/μ, tem-se que sua função densidade de probabilidade é dada pela seguinte equação:

$$f(y) = \frac{1}{\Gamma(\nu)} \left(\frac{\nu}{\mu}\right)^{\nu} y^{\nu-1} exp\left(-\frac{\nu}{\mu} y\right), \tag{2.22}$$

onde $y > 0$ e $\Gamma(\nu) = \int_0^\infty x^{\nu-1} e^{-x} dx$.

A função densidade de probabilidade pode ser escrita também na forma da família exponencial (Turkman e Silva, 2000):

$$f(y; \theta, \phi) = \exp\left\{\nu(\theta y + \ln(-\theta)) + (\nu - 1)\ln y + \ln\Gamma(\nu) + \nu \ln\nu\right\}. \tag{2.23}$$

Dessa forma, tem-se que:

- $\phi = \frac{1}{\nu}$,
- $\theta = -\frac{1}{\mu}$,
- $b(\theta) = -\ln(-\theta)$,
- $c(y, \phi) = (\nu - 1)\ln y - \ln\Gamma(\nu) + \nu \ln\nu$,
- $b^{'}(\theta) = -\frac{1}{\theta}$,

- $b''(\theta) = \frac{1}{\theta^2}$,
- $E(Y) = b'(\theta) = \mu$,
- $Var(Y) = \phi b''(\theta) = \frac{\mu^2}{\nu}$.

A função de ligação canônica para distribuição gama é denominada recíproca, onde o componente sistemático $\eta = \frac{1}{\mu}$. No entanto, existem outras funções de ligações adequadas para distribuição gama, como por exemplo, a função de ligação logarítmica. Quando $\mu > 0$,

$$\log(\mu) = \eta = X\beta \longmapsto \mu = \exp(X\beta) > 0. \tag{2.24}$$

O modelo gama é usado na análise de dados contínuos não-negativos que apresentam uma variância crescente com a média e mais, fundamentalmente, quando o coeficiente de variação dos dados for aproximadamente constante. É, também, aplicado na estimação de componentes de variância de modelos com efeitos aleatórios, e como uma distribuição aproximada de medições física, tempos de sobrevivência, etc. (Cordeiro e Demétrio, 2008)

2.3.1 Estimação

O método utilizado será o de máxima verossimilhança, um método que tem muitas propriedades ótimas, tais como, consistência e eficiência assintótica. (Cordeiro e Demétrio, 2008)

Para estimar os parâmetros $\beta's$ do modelo, considera-se inicialmente o vetor escore, vetor que é formado pelas derivadas parciais de primeira ordem do logaritmo da função de verossimilhança. Considerando o parâmetro de dispersão ϕ fixo, o logaritmo da função de verossimilhança apenas de $\boldsymbol{\beta}$ é definido como sendo

$$l(\boldsymbol{\beta}) = \frac{1}{\phi}\sum_{i=1}^{n}[y_i\theta_i - b(\theta_i)] + \sum_{i=1}^{n} c(y_i, \phi) \tag{2.25}$$

em que $\theta_1 = q(\mu_i)$, $\mu_i = g^{-1}(\eta_i)$ e $\eta_1 = \sum_{i=1}^{p} x_{ir}\beta_r$.

Nelder e Wedderburn desenvolveram um algoritmo para a estimação dos parâmetros $\boldsymbol{\beta}$ através da máxima verossimilhança baseado em um método semelhante ao de Newton-Raphson, conhecido como *Método de Escore de Fisher*. Este método consiste em resolver o sistema $U(\beta_j) = \frac{\partial l(\beta)}{\partial \beta_j} = 0$, onde $j = 1, ..., p$, $U(\boldsymbol{\beta})$ é a função Escore e $l(\boldsymbol{\beta})$ é a log-verossimilhança de $\boldsymbol{\beta}$. Pela regra da cadeia, tem-se que o elemento típico é dado por

$$U_r = \frac{1}{\phi}\sum_{i=1}^{n}(y_i - \mu_i)\frac{1}{V_i}\frac{d\mu_i}{d\eta_i}x_{ir}, \tag{2.26}$$

com $r = 1, ..., p$, $\mu_i = b'(\theta_i)$ e $\frac{d\mu_i}{\theta_i} = V_i$.

Assim, tem-se o seguinte processo iterativo

$$\boldsymbol{\beta}^{m+1} = \boldsymbol{\beta}^m + (\boldsymbol{K})^{-1}\boldsymbol{U}^m,$$

sendo $\boldsymbol{\beta}^m$ e $\boldsymbol{\beta}^{m+1}$ os vetores de parâmetros estimados nos passos m e $m+1$, $\boldsymbol{U}^m$ o vetor escore avaliado no passo m e $\boldsymbol{K}$ a matriz de informação esperada de Fisher, cujos elementos típicos são dados por

$$k_{r,s} = -E\left[\frac{\partial^2 l(\boldsymbol{\beta})}{\partial\beta_r \partial\beta_s}\right].$$

De 2.26, pode se escrever $k_{r,s}$ como

$$k_{r,s} = E(U_r U_s) = \phi^{-2}\sum_{i=1}^{n} E(Y_i - \mu_i)^2 \frac{1}{V_i^2}\left(\frac{d\mu_i}{d\eta_i}\right)^2 x_{ir}x_{is}$$

ou

$$k_{rs} = \phi^{-1}\sum_{i=1}^{n} w_i x_{ir} x_{is},$$

com peso $w_i = \frac{1}{V_i}\left(\frac{d\mu_i}{d\eta_i}\right)^2$. Assim, a matriz de informação de Fisher, $\boldsymbol{K}$, tem a forma

$$\boldsymbol{K} = \phi^{-1}\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X}, \tag{2.27}$$

onde $\boldsymbol{W} = \text{diag}\{w_1, ..., w_n\}$ é uma matriz diagonal de pesos que capta a informação sobre a distribuição e a função de ligação usadas. Com isso, o vetor escore $\boldsymbol{U} = \boldsymbol{U}(\boldsymbol{\beta})$ pode ser reescrito na forma

$$\boldsymbol{U} = \frac{1}{\phi}\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{G}(\boldsymbol{y} - \boldsymbol{\mu}),$$

com $\boldsymbol{G} = \text{diag}\{d\eta_1/d\mu_1, ..., d\eta_n/d\mu_n\} = \text{diag}\left\{g'(\mu_1), ..., g'(\mu_n)\right\}$.

Substituindo $\boldsymbol{U}$ e $\boldsymbol{K}$ em 2.27, tem-se que

$$\boldsymbol{\beta}^{(m+1)} = (\boldsymbol{X}^T\boldsymbol{W}^{(m)}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{W}^{(m)}\boldsymbol{z}^{(m)}, \tag{2.28}$$

com $\boldsymbol{z} = \boldsymbol{\eta} + G(\boldsymbol{y} - \boldsymbol{\mu})$ e $\boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{\beta}$.

Para ξ suficientemente pequeno, um dos critérios de convergência é dado por:

$$\sum_{r=1}^{p}\left(\frac{\beta_r^{m+1} - \beta_r^m}{\beta_r^m}\right) < \xi.$$

Intervalos de Confiança

De acordo com (Paula, 2004) uma banda assintótica de confiança de coeficiente $1-\alpha$ pode ser construída para $\mu(\boldsymbol{z}) = g^{-1}(\boldsymbol{z}^T\boldsymbol{\beta})$, para qualquer $\boldsymbol{z} \in R^p$. Assintóticamente temos que $\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta} \sim N_p(\boldsymbol{0}, \phi^{-1}(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X})^{-1})$. Logo, uma banda assintótica para o preditor linear $\boldsymbol{z}^T\boldsymbol{\beta}$ é dada por:

$$\boldsymbol{z}^T\widehat{\boldsymbol{\beta}} \pm \sqrt{\phi^{-1}c_\alpha}\left\{\boldsymbol{z}^T(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{z}\right\}^{-1/2},$$

para qualquer $\boldsymbol{z} \in R^p$, em que c_α é tal que $P\left\{\chi_p^2 \leq c_\alpha\right\} = 1 - \alpha$. Aplicando a transformação $g^{-1}(.)$ pode-se encontrar uma banda assintótica de confiança de coeficiente $1 - \alpha$ para $\mu(\boldsymbol{z})$, dada por:

$$g^{-1}\left[\boldsymbol{z}^T\widehat{\boldsymbol{\beta}} \pm \sqrt{\phi^{-1}c_\alpha}\left\{\boldsymbol{z}^T(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{z}\right\}^{-1/2}\right],$$

para qualquer $\mathbf{z} \in \mathbf{R^p}$, onde $\boldsymbol{z}$ é um vetor $p \times 1$ que varia livremente por R^p. As quantidades $\boldsymbol{W}$ e ϕ devem ser estimadas consistentemente. (Paula, 2004)

Teste de Hipótese

Função Desvio

Assim como nos modelos lineares, nos MLGs busca-se encontrar um modelo parcimonioso que explique bem os dados sem tornar a interpretação complexa. Para verificar a adequabilidade do ajuste de um MLG com p parâmetros, utiliza-se uma medida de discrepância proposta por Nelder & Wedderburn (1972), com expressão dada por:

$$S_p = 2(\widehat{l_n} - \widehat{l_p}),$$

sendo $\widehat{l_n}$ e $\widehat{l_p}$ os máximos do logaritmo da função de verossimilhança para os modelos saturado (modelo completo com n parâmetros) e corrente (sob pesquisa), respectivamente.

Do logaritmo da função de verossimilhança obtém-se:

$$\widehat{l_n} = \phi^{-1}\sum_{i=1}^{n}[y_i\widetilde{\theta_i} - b(\widetilde{\theta_i})] + \phi^{-1}\sum_{i=1}^{n}c(y_i,\phi)$$

e

$$\widehat{l_p} = \phi^{-1}\sum_{i=1}^{n}[y_i\widehat{\theta_i} - b(\widehat{\theta_i})] + \phi^{-1}\sum_{i=1}^{n}c(y_i,\phi),$$

sendo $\widetilde{\theta} = q(y_i)$ e $\widehat{\theta_i} = q(\widehat{\mu_i})$ as estimativas de máxima verossimilhança do parâmetro canônico sob os modelos saturado e corrente, respectivamente.

Então tem-se,

$$S_p = \phi^{-1}D_p = 2\phi^{-1}\sum_{i=1}^{n}[y_i(\widetilde{\theta_i} - \widehat{\theta_i}) + b(\widehat{\theta_i}) - b(\widetilde{\theta})], \tag{2.29}$$

em que S_p e D_p são denominados de desvio escalonado e desvio.

Pode-se ainda escrever S_p como

$$S_p = \frac{1}{\phi}\sum_{n=1}^{n}d_i^2,$$

onde d_i^2 é o *componente do desvio* que mede a diferença das funções de logverossimilhança observada e ajustada da observação i.

Na prática, o deviance S_p é comparado com o valor crítico da distribuição $\chi^2_{n-p,\alpha}$ (com $n-p$ graus de liberdade e α de significância). Caso $S_p \leq \chi^2_{n-p,\alpha}$ o modelo proposto é aceito ao nível α de significância. O deviance também é utilizado para comparar modelos encaixados (com mesma função de ligação e distribuição).

Teste da Razão de Verossimilhança:

Nos modelos lineares generalizados as estatísticas utilizadas para testar as hipóteses dos parâmetros são baseadas na teoria de máxima verossimilhança e os testes mais utilizados são o teste da razão de verossimilhança, de Wald e Escore. Quando a hipótese refere-se a um único coeficiente β_j, aconselha-se usar o teste de Wald, enquanto que o teste da razão de verossimilhança é preferido para hipóteses relativas a vários coeficientes. A seguir será apresentado apenas o teste da razão de verossimilhança, para mais informações consultar Demétrio (2003).

Suponha ϕ conhecido e deseja-se testar apenas um subconjunto do vetor de parâmetros desconhecidos β. Seja $\boldsymbol{\beta} = (\boldsymbol{\beta_1}^T, \boldsymbol{\beta_2}^T)^T$ uma partição do vetor de parâmetros $\boldsymbol{\beta}$, em que $\boldsymbol{\beta_1}$ tem dimensão q e é o vetor de interesse, e $\boldsymbol{\beta_2}$ tem dimensão (p-q). Sejam as hipóteses:

$$H_0 : \boldsymbol{\beta}_1 = \boldsymbol{\beta}_{1,0} \times H : \boldsymbol{\beta}_1 \neq \boldsymbol{\beta}_{1,0}$$

.

onde $\boldsymbol{\beta}_{1,0}$ é um vetor especificado para $\boldsymbol{\beta}_1$. Seja $\widehat{\boldsymbol{\beta}} = (\widehat{\boldsymbol{\beta}_1}^T, \widehat{\boldsymbol{\beta}_2}^T)^T$ o estimador de máxima verossimilhança (EMV) de $\boldsymbol{\beta}$ sem restrição, $\widetilde{\boldsymbol{\beta}} = (\boldsymbol{\beta}_{1,0}^T, \widetilde{\boldsymbol{\beta}_2}^T)^T$ o EMV restrito de $\boldsymbol{\beta}$ em que $\widetilde{\boldsymbol{\beta}_2}$ é o EMV de $\boldsymbol{\beta}_2$ sob H_0.

Então, compara-se os valores do logaritmo da função de verossimilhança maximizada sem restrição ($l(\widehat{\boldsymbol{\beta}_1}, \widehat{\boldsymbol{\beta}_2})$) e sob H_0 ($l(\boldsymbol{\beta}_{1,0}, \widetilde{\boldsymbol{\beta}_2})$), ou em termos de desvio, $D(\boldsymbol{y}; \widehat{\boldsymbol{\mu}})$ e $D(\boldsymbol{y}; \widetilde{\boldsymbol{\mu}})$ em que $\widetilde{\boldsymbol{\mu}} = g^{-1}(\widetilde{\boldsymbol{\eta}})$ e $\widetilde{\boldsymbol{\eta}} = \boldsymbol{X}\boldsymbol{\beta}$. A estatística para do teste é dada por:

$$w = 2[l(\widehat{\boldsymbol{\beta}_1}, \widehat{\boldsymbol{\beta}_2}) - l(\boldsymbol{\beta}_{1,0}, \widetilde{\boldsymbol{\beta}_2})] = \phi^{-1}[D(\boldsymbol{y}; \widetilde{\boldsymbol{\mu}}) - D(\boldsymbol{y}; \widehat{\boldsymbol{\mu}})] \tag{2.30}$$

Para amostras grandes, rejeita-se H_0 a um nível de $100\alpha\%$ de significância, se $w > \chi^2_{q,1-\alpha}$.

2.3.2 Diagnósticos

De acordo com Demétrio e Cordeiro (2008), as técnicas usadas em MLGs são parecidas com as do modelo de regressão linear. No entanto, no MLG deve-se utilizar a variável dependente ajustada $\boldsymbol{z}$, e o preditor linear estimado $\widehat{\boldsymbol{\eta}}$, ao invés dos vetores y e $\widehat{\mu}$ usados na verificação da pressuposição de linearidade para o modelo linear. Outra modificação é a estimativa consistente do parâmetro de dispersão ϕ e a matriz projeção $\boldsymbol{H}$ que é definida por:

$$\boldsymbol{H} = \boldsymbol{W}^{1/2}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{W}^{1/2}. \tag{2.31}$$

Nota-se que $\boldsymbol{H}$ depende das variáveis explanatórias, da função de ligação e da função de variância, dificultando a interpretação da medida de leverage. Demonstra-se que

$$\boldsymbol{V}^{-1/2}(\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}) \cong \boldsymbol{H}\boldsymbol{V}^{-1/2}(\boldsymbol{Y} - \boldsymbol{\mu}),$$

sendo $\boldsymbol{V} = \text{diag}\{V(\mu_i)\}$, o que mostra que $\boldsymbol{H}$ mede a influência em unidades studentizadas de $\boldsymbol{Y}$ sobre $\widehat{\boldsymbol{\mu}}$. (Cordeiro e Demétrio, 2008)

A seguir serão apresentados alguns dos resíduos utilizados em MLGs.

Resíduo Ordinário:

O resíduo mais simples e é dado por:

$$\widehat{r}_i = y_i - \widehat{\mu}_i. \tag{2.32}$$

Resíduos de Pearson:

O resíduo de Pearson é definido por:

$$\widehat{r}_i^P = \frac{y_i - \widehat{\mu}}{\widehat{V}_i^{1/2}}. \tag{2.33}$$

Componentes do Desvio:

O componente do desvio é dado pela seguinte expressão

$$d(y_i; \widehat{\mu}_i) = \text{sinal}(y_i - \widehat{\mu}_i)\sqrt{2}\left[y_i(q(y_i) - q(\widehat{\mu}_i)) - b(q(y_i)) + b(q(\widehat{\mu}_i))\right]^{1/2}, \tag{2.34}$$

em que a função $q(.)$ relaciona o parâmetro θ com a média μ e a função $b(.)$ varia de acordo com o componente aleatório.

O componente do desvio padronizado é dado por

$$\widehat{t}_{D_i} = \frac{\phi^{1/2} d(y_i; \widehat{\mu}_i)}{\sqrt{(1 - \widehat{h_{ii}})}},$$

em que $\widehat{h}_{ii}$ são os termos da diagonal principal da matriz chapéu H.

Pontos Influentes:

Para identificar pontos influentes, calcula-se a distância de Cook (D_{c_i}) aproximada que é dada por

$$\widehat{D}_{c_i} = \frac{(y_i - \widehat{\mu}_i)^2 \widehat{h_{ii}}}{\widehat{V}_i(1 - \widehat{h_{ii}})^2}, \tag{2.35}$$

onde $\widehat{h}_{ii}$ é o i-ésimo elemento da diagonal da matriz H. Diz-se que uma observação y_i é influente quando a distância de Cook obtida daquele valor se distancia muito dos demais valores calculados. Pode-se verificar então fazendo o gráfico da distância de Cook versus o índice.

Tipos de Gráficos:

São basicamente os mesmos gráficos para o modelo de regressão linear. A construção do gráfico dos valores ajustados transformados versus os resíduos deve apresentar resíduos em torno de zero com amplitude constante para que indique um bom ajuste.

De acordo com Paula (2004), as técnicas mais recomendadas para os MLGs são os gráficos $\widehat{t}_{D_i}$ contra a ordem das observações, contra os valores ajustados e contra as variáveis explicativas.

2.4 Modelos Aditivos Generalizados para Posição, Escala e Forma (GAMLSS)

Como observado por Dantas (2005), embora a regressão linear tradicional seja muito usada para modelar dados imobiliários, pode apresentar resultados inconsistentes pois muitas vezes os dados não apresentam normalidade e não atendem os pressupostos necessários da teoria. Devido aos avanços da estatística, a engenharia de avaliações também se adaptou e passou a utilizar outros métodos de regressão, como por exemplo, os modelos lineares generalizados (MLGs), que embora aumentem as opções para a distribuição da variável resposta (as pertencentes à família exponencial) não permitem que esta tenha uma grande assimetria e curtose, além de modelar apenas a média.

A fim de suprir as restrições acima, *Rigby & Stasinopoulos* (2005) propuseram os modelos aditivos generalizados para posição, escala e forma (*GAMLSS*), uma nova classe de modelos de regressão (semi)paramétricos, que permitem que todos os parâmetros da variável resposta sejam modelados de forma linear ou não-linear.

Lembrando que em um modelo paramétrico, a forma das funções que relaciona as variáveis preditoras e a variável resposta são conhecidas, exceto por um número finito de parâmetros desconhecidos e pode ser representada pela seguinte forma:

$$y_i = f(x_i, \beta_1, ..., \beta_k) + \epsilon_i,$$

onde $i = 1, ...n$, e y_i é a i-ésima observação da variável resposta, $\boldsymbol{\beta}^T = (\beta_1, ..., \beta_k)$ é o vetor de parâmetros a ser estimado e ϵ_i tem média zero e variância σ^2.

Já no modelo não paramétrico, a forma funcional de f não é conhecida e de acordo com Florencio (2010), estima-se uma função média sem referência a uma forma funcional previamente estabelecida, precisando apenas escolher o espaço de funções apropriado.

A seguir será explicado brevemente sobre a estimação e inferência dos modelos GAMLSS.

2.4.1 Modelos aditivos

Os *modelos aditivos* são considerados uma extensão natural dos modelos lineares, os quais mantém o efeito aditivo na relação entre a variável resposta e as variáveis explicativas podendo incluir funções arbitrárias, não necessariamente lineares, e são expressos por

$$y_i = \sum_{j=1}^{p} f_j(x_{ij}) + \epsilon_i, \tag{2.36}$$

em que $y^T = (y_1, y_2, ..., y_n)$ é um vetor $n \times 1$ de respostas e a i-ésima linha da matriz $\mathbf{X}$ é $\mathbf{X}_i = (x_{i1}, x_{i2}, ..., x_{ij})$ um vetor $p \times 1$ de variáveis explicativas, com $E(\epsilon_i) = 0$ e $Var(\epsilon_i) = \sigma$. Nota-se que ao considerar $f_j(x_{ij}) = x_{ij}\beta_j$, tem-se um caso particular do modelo linear.

Este modelo é não paramétrico e cada função f_j, com $j = 1, ..., p$ é estimada através de suavizadores. Quando os preditores combinam formas paramétricas de algumas variáveis explicativas, (g), com termos não-paramétricos de outras, $(k - g)$, são denominados *semiparamétricos* e expressos por

$$y_i = \beta_1 x_{i1} + ... + \beta_g x_{ig} + f_1\left(x_{i,g+1}\right) + ... + f_{k-g}\left(x_{i,g+k}\right) + \epsilon_i.$$

Suavizadores

As funções de suavização, ou suavizadores, de acordo com Lima et al. (2001), são ferramentas que descrevem a variação da média de uma variável Y como função de uma ou mais variáveis não estocásticas $X_1, .., X_k$. Entre os muitos métodos para estimação de modelos não paramétricos (ver Härdle, 1990), destaca-se o Spline, o qual será definido a seguir.

Spline:uma função por partes "emendadas"por nós (knots), na qual cada parte é um polinômio contínuo com um certo grau de derivadas também contínuas.

As funções splines estão associadas a uma partição do intervalo $[a, b]$ onde se pretende trabalhar:

$$I = a = x_0 < x1 < ...x_{m-1} < x_m = b.$$

Em cada subintervalo (x_{i-1}, x_i), com $i = 1, ..., m$ as splines são polinômios de um determinado grau m. Esse procedimento produz um polinômio por partes $s(x)$, que pode ser utilizado para aproximar a função procurada. Existe uma relação entre o grau dos "pedaços "dos polinômios e a ordem das derivadas exigidas nos pontos da partição. Assim, devem ser impostas algumas restrições na definição geral das splines para garantir a continuidade e suavidade de $s(x)$, como a colocação dos nós (knots) e os pontos de ligação entre os polinômios. (Cunha, 2000)

Definição 2.4.1 *Uma função $s(x)$ é chamada de spline de grau m, associada a uma partição do intervalo $[a, b]$, se:*

- *$s(x)$ é um polinômio de grau m em cada subintervalo (x_{k-1}, x_k), $k = 1, ..., m$;*
- *$s(x)$ tem m–1 derivadas contínuas em cada x_k, e, portanto em [a, b].*

O conjunto das funções $S_d(x_0, ..., x_m)$ é um espaço linear e recebe o nome de espaço spline (spline space) cujos elementos são funções splines.

2.4.2 Definição

A classe de modelos GAMLSS é dita semi-paramétrica pois exige uma distribuição paramétrica para a variável resposta ao mesmo tempo que permite que os parâmetros da

distribuição e das funções das covariáveis sejam modelados através de funções de suavização não-paramétricas. As opções para a distribuição da variável resposta são bastante variadas, existindo cerca de 40 tipos diferentes no GAMLSS, com um, dois, três ou até quatro parâmetros.

Considere $y^T = (y_1, y_2, ..., y_n)$ o vetor de observações da variável resposta. Sua função (densidade) de probabilidade condicionada a $\boldsymbol{\theta_i}$ é $f(y_i|\boldsymbol{\theta}_i)$, onde $\boldsymbol{\theta}_i^T = (\theta_{i1}, \theta_{i2}, ..., \theta_{ip})$ é um vetor de p parâmetros que é relacionado aos efeitos das variáveis explicativas sobre a variável resposta através de funções de ligação denominadas *smoother*, as quais podem assumir diversas formas. (Paiva, Freire e Cecatti, 2008).

Seja uma função de ligação $g_k(\cdot)$, para $k = 1, ..., p$ que relaciona o *k*-ésimo parâmetro θ_k às variáveis explanatórias e efeitos aleatórios por meio de um modelo aditivo dado por:

$$g_k(\boldsymbol{\theta}_k) = \boldsymbol{\eta}_k = \boldsymbol{X}_k\boldsymbol{\beta}_k + \sum_{j=1}^{J_k} \boldsymbol{Z}_{jk}\boldsymbol{\gamma}_{jk}, \tag{2.37}$$

em que

- $\boldsymbol{\theta}_k^T = (\theta_{1k}, \theta_{2k}, ..., \theta_{nk})$ vetor $n \times 1$;
- $\boldsymbol{\eta}_k^T = (\eta_{1k}, \eta_{2k}, ..., \eta_{nk})$ vetor $n \times 1$;
- $\boldsymbol{\beta}_k^T = \left(\beta_{1k}, \beta_{2k}, ..., \beta_{J'_k k}\right)$ é um vetor de parâmetros de tamanho J'_k;
- $\boldsymbol{X}_k$ é uma matriz de delineamento conhecida de tamanho $n \times J'_k$;
- $\boldsymbol{Z}_{jk}$ é uma matriz de delineamento fixa e conhecida de ordem $n \times q_{jk}$;
- $\boldsymbol{\gamma}_{jk}$ é uma variável aleatória q_{jk}-dimensional com distribuição $\boldsymbol{\gamma}_{jk} \sim N\left(\mathbf{0}, \boldsymbol{G}_{jk}^{-1}\right)$;
- $\boldsymbol{G}_{jk}^{-1}$ é uma inversa (generalizada) de uma matriz simétrica $\boldsymbol{G}_{jk} = \boldsymbol{G}_{jk}(\boldsymbol{\lambda}_{jk})$ de dimensão $q_{jk} \times q_{jk}$ que pode depender de um vetor de hiperparâmetros $\boldsymbol{\lambda}_{jk}$.

O modelo acima é denotado por *GAMLSS aditivo semiparamétrico*. Conforme comentado por Florencio (2010), como na maioria das situações práticas são requeridos no máximo quatro parâmetros para a distribuição, os quais são caracterizados por μ (posição), σ (escala), ν (assimetria) e τ (curtose) e denominados parâmetros de posição, escala e forma (os dois últimos) respectivamente, para estimá-los têm-se as seguintes equações:

$$g_1(\boldsymbol{\mu}) = \boldsymbol{\eta}_1 = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \sum_{j=1}^{J_1} \boldsymbol{Z}_{j1}\boldsymbol{\gamma}_{j1}, \tag{2.38}$$

$$g_2(\boldsymbol{\sigma}) = \boldsymbol{\eta}_2 = \boldsymbol{X}_2\boldsymbol{\beta}_2 + \sum_{j=1}^{J_2} \boldsymbol{Z}_{j2}\boldsymbol{\gamma}_{j2}, \tag{2.39}$$

$$g_3(\boldsymbol{\nu}) = \boldsymbol{\eta}_3 = \boldsymbol{X}_3\boldsymbol{\beta}_3 + \sum_{j=1}^{J_3} \boldsymbol{Z}_{j3}\boldsymbol{\gamma}_{j3}, \tag{2.40}$$

$$g_4(\boldsymbol{\tau}) = \boldsymbol{\eta}_4 = \boldsymbol{X}_4\boldsymbol{\beta}_4 + \sum_{j=1}^{J_4} \boldsymbol{Z}_{j4}\boldsymbol{\gamma}_{j4}. \tag{2.41}$$

2.4.2.1 Casos particulares do modelo *GAMLSS*

Modelo linear completamente paramétrico

Se $J_k = 0$, não há termos aditivos associados aos parâmetros da distribuição. Então temos o caso do modelo linear completamente paramétrico dado por:

$$g_k(\boldsymbol{\theta}_k) = \boldsymbol{\eta}_k = \boldsymbol{X}_k\boldsymbol{\beta}_k.$$

Modelo aditivo semiparamétrico linear

Se $\boldsymbol{Z}_{jk} = \boldsymbol{I}_n$, em que $\boldsymbol{I}_n$ é uma matriz identidade de ordem $n \times n$ e $\gamma_{jk} = \boldsymbol{h}_{jk} = h_{jk}(\boldsymbol{x}_{jk})$ para todas as combinações de j e k, tem-se:

$$g_k(\boldsymbol{\theta}_k) = \boldsymbol{\eta}_k = \boldsymbol{X}_k\boldsymbol{\beta}_k + \sum_{j=1}^{J_k} h_{jk}(\boldsymbol{x}_{jk}),$$

em que X_{jk}, são vetores de tamanho n para $j = 1, 2, ..., J_k$ e $k = 1, 2, ..., p$. A função h_{jk} é uma função desconhecida da variável explanatória $\boldsymbol{X}_{jk}$ e $\boldsymbol{h}_{jk} = h_{jk}(\boldsymbol{x}_{jk})$ é um vetor que avalia a função h_{jk} em $\boldsymbol{x}_{jk}$.

Modelo aditivo semiparamétrico não-linear

Pode-se estender o modelo aditivo semiparamétrico linear para permitir a inclusão de termos não-lineares na modelagem dos k parâmetros da distribuição na forma:

$$g_k(\boldsymbol{\theta}_k) = \boldsymbol{\eta}_k = h_k(\boldsymbol{x}_k, \boldsymbol{\beta}_k) + \sum_{j=1}^{J_k} h_{jk}(x_{jk}),$$

em que h_k para $k = 1, 2, ..., p$ são funções não lineares e $\boldsymbol{X}_k$ é uma matriz de covariáveis conhecida de ordem $n \times J_k$.

Modelo paramétrico não-linear

Se $J_k = 0$, então o modelo se reduz a um *GAMLSS* paramétrico não-linear expresso por:

$$g_k(\boldsymbol{\theta}_k) = \boldsymbol{\eta}_k = h_k(\boldsymbol{X}_k, \boldsymbol{\beta}_k).$$

Modelo paramétrico linear

Finalmente, se $h_k\left(\boldsymbol{X}_k, \boldsymbol{\beta}_k\right) = \boldsymbol{X}_k\boldsymbol{\beta}_k$, para $i = 1, 2, ..., n$ e $k = 1, 2, ..., p$, então, se reduz ao modelo paramétrico linear.

O *GAMLSS* é utilizado para modelar uma família de distribuições mais amplas que apenas as da família exponencial. A única restrição do GAMLSS implementado no R é que a $\log f\left(y_i|\mu_i, \sigma_i, \nu_i, \tau_i\right)$ e a primeira derivada em relação a cada parâmetro de $\boldsymbol{\theta}$ tem que ser computável. (Florencio, 2010)

Para maiores informações sobre quais distribuições e parâmetros estão disponíveis no pacote *GAMLSS* (*R Development Core Team 2007*) para modelagem, bem como os termos aditivos e funções, vide *Rigby* & *Stasinopoulos* (2005) e o *help* do R.

Preditor Linear

Como no modelo GAMLSS 2.37 pode haver componentes paramétricos, $\boldsymbol{X}_k\boldsymbol{\beta}_k$, $k = 1, ..., p$ e aditivos, $\boldsymbol{Z}_{jk}$ no preditor linear, a seguir alguns comentários a respeito de cada um deles.

Termos Paramétricos: o componente paramétrico pode conter termos lineares e de interação, fatores, polinômios, polinômios fracionários (Royston & Altman, 1994) e polinômios segmentados (com nós fixos) para as variáveis exploratórias.

Termos Aditivos: podem modelar uma série de termos, como termos de suavização e efeitos aleatórios. Alguns dos mais usados são *splines* cúbicos, *splines* de penalização, polinômios fracionários e *loess*. Para uma leitura mais detalhada desses termos aditivos, vide Cleveland et al. (1992) e Hastie & Tibshirani (1990(a), 1993(b)). Alguns termos aditivos disponíveis no *GAMLSS* estão descritos no Quadro 1.

Quadro 1: Alguns termos aditivos disponíveis no *R*.

Termo aditivo	Comando no *R*
Cubic splines	cs( )
Varying coefficient	vc( )
Penalized splines	ps( )
Loess	lo( )
Fractional polynomials	fp( )
Power polynomials	pp( )
non-linear fit	nl( )
Ramdom effects	random( )
Random effects	ra( )
Random coefficient	rc( )

2.4.3 Estimação

Seja $\boldsymbol{\gamma}_{jk}$ com distribuição $N(\mathbf{0}, \boldsymbol{G}_{\boldsymbol{jk}}^{\mathbf{-1}})$, em que $\boldsymbol{G}_{jk}^{-1}$ é uma matriz simétrica generalizada inversível $(q_{jk} \times q_{jk})$ dependente dos hiperparâmetros $\boldsymbol{\lambda}_{jk}$. Se $\boldsymbol{G}_{jk}$ não tiver inversa, então é compreendida a priori como uma função de densidade imprópria proporcional a $\exp\left(\frac{1}{2}\boldsymbol{\gamma}_{jk}^{T}\boldsymbol{G}_{jk}\boldsymbol{\gamma}_{jk}\right)$. (Rigby & Stasinopoulos (2005), Florêncio (2010)).

Assim, no processo de estimação do modelo 2.37, o vetor de parâmetros $\boldsymbol{\beta}_k$ e os parâmetros de efeitos aleatórios $\boldsymbol{\gamma}_{jk}$, para $j = 1, 2, ..., J_k$ e $k = 1, 2, 3, 4$ são estimados na estrutura de *GAMLSS* (para valores fixos dos hiper-parâmetros de suavização $\boldsymbol{\lambda}_{jk}$'s) maximizando a função de verossimilhança penalizada ℓ_p dada por:

$$\ell_p = \ell - \frac{1}{2}\sum_{k=1}^{p}\sum_{j=1}^{J_k}\boldsymbol{\gamma}_{jk}^{T}\boldsymbol{G}_{jk}\boldsymbol{\gamma}_{jk}, \tag{2.42}$$

onde $\ell = \sum_{i=1}^{n} \log f\left(y_i|\theta^i\right)$ é a função de log-verossimilhança.

Na ausência de termos aditivos, os parâmetros $\boldsymbol{\beta}s$ são estimados da função de verossimilhança maximizada, ou seja, $\ell_p = \ell$.

Algoritmos

Nos modelos completamente paramétricos, maximizar a função de verossimilhança penalizada ℓ_p é equivalente a maximizar a função de verossimilhança ℓ. No R, a maximização de ℓ_p pode ser feita através de dois algoritmos que são escolhidos dentro da função *gamlss( )* na opção *method*.

Os algoritmos implementados são *CG* e *RS*. O primeiro é uma generalização do algoritmo de *Cole* e *Green* (1992) que utiliza a primeira derivada e o valor esperado ou aproximado das derivadas de segunda ordem e das derivadas cruzadas da função de log-verossimilhança com relação aos parâmetros da distribuição. O segundo é uma generalização do algoritmo usado

por *Rigby & Stasinopoulos* no ajuste da média e da dispersão de modelos aditivos e não utiliza o valor esperado das derivadas cruzadas. Dessa forma é usado quando os parâmetros θ são ortogonais e o valor esperado das derivadas cruzadas é zero.

2.4.4 Seleção do modelo e Diagnósticos

A seleção do modelo pode ser feita através do desvio global ajustado (Global Deviance-GD), que é definido como menos duas vezes o logaritmo da função de verossimilhança calculado para os valores estimados dos coeficientes do modelo. É um critério objetivo que compara dois modelos encaixados utilizando o teste da razão de verossimilhança.

Para modelos não encaixados, utiliza-se o critério de informação de Akaike generalizado (GAIC) que penaliza sobre ajustes adicionando aos desvios globais ajustados uma penalidade fixa $\#$, ou seja, $GAIC(\#) = GD + \#df$, onde df corresponde aos graus de liberdade efetivos no modelo. O critério de informação de Akaike (Akaike Information Criterion - AIC; Akaike, 1974) e o critério bayesiano de Schwarz (Schwarz Bayesian Criterion -SBC; Schwarz, 1978) são casos especiais do critério $GAIC(\#)$, e correspondem a $\# = 2$ e $\# = log(n)$, respectivamente. (Florencio, 2010)

Além dos critérios *AIC* e *SBC*, será realizada a comparação entre os modelos utilizando o "pseudo coeficiente de determinação" ($pseudo - R^2$), o qual é calculado pela expressão:

$$pseudo - R^2 = (corr(\text{valores observados, valores ajustados}))^2.$$

A análise dos resíduos pode ser feita através da função *plot()* que produz quatro gráficos de resíduos quantílicos normalizados e um sumário de medidas resumo da distribuição dos resíduos. Com isso, se um modelo se ajustar bem aos dados, os seus resíduos verdadeiros deverão apresentar distribuição aproximadamente normal padrão, mesmo quando a distribuição do modelo não é normal (DUNN e SMITH, 1996).

Um dos diferenciais no diagnóstico do GAMLSS são os gráficos *worm-plots* que foram introduzidos por Van Burren & Fredriks (2001) e consistem em uma ferramenta de diagnóstico para análise dos resíduos que pode ser expresso como um único gráfico com o intervalo todo da variável explicativa ou em diferentes regiões (intervalos) como uma coleção de gráficos *detrended qq-plot*[2]. Se os pontos estão situados entre as curvas elípticas, ou seja, na região de aceitação, o ajuste do modelo é considerado adequado.

2.5 Influência Local

Para compreender as características de um conjunto de dados, muitas vezes utiliza-se de modelos estatísticos que tentam extrair informações aproximadas de situações mais complexas obtendo, na maioria das vezes, resultados não exatos.

[2]*Detrended qq-plot*:gráfico quantil-quantil dos z-escores, para maiores detalhes consultar Van Buuren, S. & Fredriks, (2001).

Na análise de diagnósticos, Cook (1986) propos um método com uma abordagem relativamente simples, baseada em uma verossimilhança bem comportada juntamente com alguns conceitos de geometria diferencial. Tal método visa estudar a variação dos resultados sob pequenas modificações na formulação do problema.

A metodologia não exige deleção de observações e permite avaliar a influência conjunta de todos os pontos. Além de permitir que diferentes gráficos de influência sejam desenvolvidos. (Paula, 2004)

Na literatura é possível encontrar muitos artigos abordando o assunto, como em Beckman, Nachtsheim e Cook (1987), Lawrence (1990), Fung e Kwan (1997) que mostram que o afastamento pela verossimilhança é uma medida de influência invariante com mudanças de escalas nos dados, Thomas e Cook (1990) que utilizam a teoria em modelos lineares generalizados, Poon e Poon (1999) que propuseram a curvatura conforme, entre outros.

Seja $L(\boldsymbol{\theta})$ a log-verossimilhança do modelo postulado e $\boldsymbol{\theta}$ o vetor de parâmetros desconhecidos $p \times 1$. Denota-se por $L(\boldsymbol{\theta}|\boldsymbol{w})$ a log-verossimilhança do modelo perturbado, sendo $\boldsymbol{w} = (w_1, ..., w_s)^T$ um vetor de perturbações $n \times 1$, com $\boldsymbol{w} \in \Omega \subset R^n$. Assume-se que o modelo postulado está encaixado ao modelo perturbado, ou seja, existe um vetor $\boldsymbol{w}_0 \in \Omega$, vetor de não perturbação, tal que $L(\boldsymbol{\theta}|\boldsymbol{w}_0) = L(\theta)$ e mais, $L(\boldsymbol{\theta}|\boldsymbol{w})$ é duas vezes diferenciável $(\boldsymbol{\theta}^T, \boldsymbol{w}^T)$.

A partir disso, tem-se que o **afastamento da verossimilhança**, medida de influência utilizada para estudar as estimativas do modelo após as perturbações proposta por Cook (1986) é dada por:

$$LD(\boldsymbol{w}) = 2\left[L(\widehat{\boldsymbol{\theta}}) - L(\widehat{\boldsymbol{\theta}_w})\right], \tag{2.43}$$

sendo $\widehat{\boldsymbol{\theta}}$ e $\widehat{\boldsymbol{\theta}}_w$ os estimadores de máxima verossimilhança dos modelos não perturbado (postulado) e perturbado respectivamente. Note que $LD(\boldsymbol{w}) \geq 0$.

Seja

$$\boldsymbol{\alpha}(\boldsymbol{w}) = (\boldsymbol{w}^T, LD(\boldsymbol{w}))^T, \tag{2.44}$$

o **gráfico de influência** que é uma superfície geométrica (n+1)-dimensional formada pelos valores do vetor $(n+1) \times 1$, com w variando em Ω.

A ideia proposta por Cook (1986) consiste em estudar o comportamento do afastamento da verossimilhança em uma vizinhança de $\boldsymbol{w}_0$, ou seja, como a superfície $\boldsymbol{\alpha}(\boldsymbol{w})$ desvia-se do seu plano tangente. Este estudo pode ser feito a partir das curvaturas das seções normais da superfície $\boldsymbol{\alpha}(\boldsymbol{w})$ em $\boldsymbol{w}_0$ na direção de um vetor unitário $\mathbf{l}$. Tais curvaturas são chamadas de **curvaturas normais**, que são interseções da superfície com planos contendo o vetor normal com seu plano tangente em $\boldsymbol{w}_0$. A interseção do plano tangente com a seção normal é denominada **linha projetada**, que pode ser obtida pelo gráfico de $LD(\boldsymbol{w}_0 + a\boldsymbol{l}) \times a$, com $a \in R$.

A curvatura normal da linha projetada, denotada por $\boldsymbol{C}_l$, é definida como sendo a curvatura de $(a, LD\{w(a)\})$ em $a = 0$, em que $w(a) = \boldsymbol{w}_0 + a\boldsymbol{l}$. Denomina-se $\boldsymbol{C}_l$ curvatura normal da superfície $\alpha(\boldsymbol{w})$ em $\boldsymbol{w}_0$ e na direção unitária $\boldsymbol{l}$. (Paula 2004)

A partir da geometria diferencial, Cook (1986) mostra que a curvatura normal na direção $\boldsymbol{l}$ tem a seguinte forma:

$$C_l(\boldsymbol{\theta}) = 2|\boldsymbol{l}^T \boldsymbol{K}^T \ddot{\boldsymbol{L}}^{-1} \boldsymbol{K}\boldsymbol{l}|, \tag{2.45}$$

sendo $-\ddot{\boldsymbol{L}}$ a matriz de informação observada de Fisher, $\boldsymbol{K} = \partial L(\theta|w)/\partial\theta_i \partial w_j)$ avaliada em $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}$ e $\boldsymbol{w} = \boldsymbol{w}_0$, com $i = 1, ...p$ e $j = 1, ..., s$.

Altos valores de $\boldsymbol{C}_l$ representam sensibilidade para as perturbações introduzidas na direção de $\boldsymbol{l}$. O autovetor normalizado correspondente ao maior autovalor $\boldsymbol{C}_{l_{max}}$ da matriz $\boldsymbol{B} = \boldsymbol{K}^T(-\ddot{\boldsymbol{L}}^{-}1)\boldsymbol{K}$ é a direção que produz maior influencia local, $\boldsymbol{l}_{max}$. Assim, uma sugestão é utilizar o gráfico de $|\boldsymbol{l}_{max}|$ versus as ordens das observações, pois pode indicar observações que sejam mais influentes sob o esquema de perturbação adotado.

2.5.1 Influência Local em Modelos Lineares Generalizados

Considere novamente a fórmula dada em 2.19:

$$f(y; \theta_i, \phi) = \exp[a(\phi)^{-1} \{y\theta_i - b(\theta_i)\} + c(y, \phi)],$$

onde θ e ϕ são parâmetros e a(.), b(.) e c(.) são funções conhecidas. A dependência de y_i sobre x_i é modelada através do parâmetro θ_i, o qual Thomas e Cook (1990) denotam como $\theta_i = k(\boldsymbol{x}_i^T\boldsymbol{\beta})$. Para distinguir da função de ligação usual da média, detona-se k(.) como ligação-θ.

Por simplificação, o parâmetro de dispersão é assumido ϕ conhecido, quando não, é possível estimá-lo, $\widehat{\phi}$, e escreve-se $\widehat{a} = a(\widehat{\phi})$. Com isso, a log verossimilhança para $\boldsymbol{\beta}$ sem os termos independentes de $\boldsymbol{\beta}$ é dada por:

$$L(\boldsymbol{\beta}) = \frac{1}{\widehat{a}} \sum_{i=1}^{n} [y_i k(\boldsymbol{x}_i^T\boldsymbol{\beta}) - bk(\boldsymbol{x}_i^T\boldsymbol{\beta})]. \tag{2.46}$$

Assim, o afastamento da verossimilhança é dado por:

$$LD(\boldsymbol{w}) = L(\widehat{\boldsymbol{\beta}}) - L(\widehat{\boldsymbol{\beta}}_w).$$

E o calculo de $\boldsymbol{l}_{max}$ é feito a partir da:

- Função Escore $U(\boldsymbol{\beta}) = \frac{\partial L(\boldsymbol{\beta}; y)}{\partial \boldsymbol{\beta}^T} = \dot{\boldsymbol{e}}(\boldsymbol{\beta})^T \boldsymbol{X}/\widehat{a}$, onde $\boldsymbol{X}$ é a matriz de variáveis explanatórias e $\dot{e}(\boldsymbol{\beta})$ é um vetor $n \times 1$, com elementos $\dot{e}_i = [y_i - \dot{b}(\theta_i)k(\boldsymbol{x}_i^T\boldsymbol{\beta})]\dot{k}(\boldsymbol{x}_i^T\boldsymbol{\beta})$, onde "."ou ".."sobre a função denota primeira e segunda derivada respectivamente;

- Informação observada de $\boldsymbol{\beta}$: $-\frac{\partial^2 L(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}\boldsymbol{\beta}^T} = \frac{1}{\widehat{a}} \boldsymbol{X}^T \ddot{\boldsymbol{E}}(\boldsymbol{\beta}) \boldsymbol{X}$, onde $\ddot{\boldsymbol{E}}(\boldsymbol{\beta})$ é a matriz diagonal com elementos

$$\ddot{e}_i(\boldsymbol{\beta}) = \ddot{b}(\theta_i)\dot{k}(\boldsymbol{x}_i^T\boldsymbol{\beta})^2 - y_i - \dot{b}(\theta_i)\ddot{k}(\boldsymbol{x}_i^T\boldsymbol{\beta}).$$

A informação esperada de $\boldsymbol{\beta}$ calculada em $\widehat{\boldsymbol{\beta}}$ é dada por $\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X}/\widehat{a}$, onde $\boldsymbol{W}(\boldsymbol{\beta})$ é o valor esperado de $\ddot{\boldsymbol{E}}(\boldsymbol{\beta})$ e $\boldsymbol{W} = W(\widehat{\boldsymbol{\beta}})$, que é uma matriz diagonal de pesos. Para modelos lineares generalizados com ligação- $\boldsymbol{\theta}$ canônica, $k(\boldsymbol{u}) = \boldsymbol{u}$, tem-se que $\ddot{\boldsymbol{E}} = \boldsymbol{W}$

A curvatura normal na direção unitária $\boldsymbol{l}$ é então denotada por (Cook, 1986):

$$\boldsymbol{C}_l = 2\widehat{a}|\boldsymbol{l}^T\boldsymbol{B}^T(\boldsymbol{X}^T\ddot{(\boldsymbol{E})}\boldsymbol{X})^{-1}\boldsymbol{B}|, \tag{2.47}$$

com $\boldsymbol{B} = \partial^2 L(\boldsymbol{\beta}|\boldsymbol{w})/\partial\boldsymbol{\beta}\partial\boldsymbol{w}^T$ calculado em $\boldsymbol{w}_0$ e $\widehat{\boldsymbol{\beta}}$.

De acordo com Thomas e Cook (1990), no diagnóstico de influência, primeiramente se especifica o esquema de perturbação e deriva $\boldsymbol{B}$, então, a direção de maior curvatura normal $\boldsymbol{l}_{max}$ é o autovetor correspondente ao maior autovalor de $\boldsymbol{B}^T(\boldsymbol{X}^T\ddot{\boldsymbol{E}}\boldsymbol{X})^{-1}\boldsymbol{B}$.

Tipos de perturbação

Os tipos de perturbações são os mais diversos, entre eles, os que mais se destacam são:

- Perturbação de casos: $L(\boldsymbol{\theta}|\boldsymbol{w}) = \sum_{i=1}^{n} w_i L_i(\boldsymbol{\theta})$, com $0 \leq w_i \leq 1$;
- Perturbação na resposta: $y_{iw} = y_i + \sigma_{y_i} w_i$, com $w \in R$;
- Perturbação em x_i: $x_{iw} = x_i + \sigma_{x_i} w_i$, com $w \in R$;
- Perturbação na matriz de variância e covariância: $\Sigma_{iw} = w_i^{-1}\Sigma_i$, com $w_i \in R - 0$.

Na perturbação de casos, é utilizado um vetor de pesos $\boldsymbol{w}^T = (w_1, ..., w_n)$ onde as mudanças de pesos são feitas simultaneamente para todos os casos afim de verificar a maior mudança local nos coeficientes de regressão estimados, medida por $LD(\boldsymbol{w}_0)$. O vetor $\boldsymbol{w}_0 = (1, ..., 1)$ não apresenta perturbações, dessa maneira, tem-se que a log verossimilhança é dada por:

$$L(\boldsymbol{\beta}|\boldsymbol{w}) = \sum_{i=1}^{n} w_i[y_i k(\boldsymbol{x}_i^T\boldsymbol{\beta}) - bk(\boldsymbol{x}_i^T\boldsymbol{\beta})]. \tag{2.48}$$

De 2.47, $\boldsymbol{B} = \boldsymbol{X}^T\dot{\boldsymbol{E}}/\widehat{a}$. Então, a direção de máxima curvatura normal $\boldsymbol{l}_{max}$ é o autovetor de $\dot{\boldsymbol{E}}\boldsymbol{X}(\boldsymbol{X}^T\ddot{\boldsymbol{E}}\boldsymbol{X})^{-1}\boldsymbol{X}^T\dot{\boldsymbol{E}}$ associado ao maior autovalor δ_1, e $\boldsymbol{C}_{max} = 2|\delta_1|/\widehat{a}$.

Os demais casos serão omitidos, para maiores informações ver Thomas e Cook, (1990).

Diferença Relativa

Uma maneira de verificar o impacto da retirada da observação considerada influente nas estimativas dos parâmetros é através da diferença relativa. A diferença relativa é dada por

$$RC_{\widehat{\theta}} = \left| \frac{\widehat{\theta} - \widehat{\theta}_{(I)}}{\widehat{\theta}} \right|,$$

onde $\widehat{\theta}_{(I)}$ é a estimativa do parâmetros sem a observação influente. A influência para cada observação deve ser $(1/n) * 100\%$, em que n é o número de observações.

2.6 Considerações Finais

Neste capítulo foram apresentadas as principais metodologias de regressão e suas propriedades para a formação de preços em dados imobiliários. Os modelos lineares, que embora bastante utilizados, apresentam algumas limitações como a homocedasticidade e a normalidade que dificilmente estão presentes em dados de imóveis. Os modelos lineares generalizados (MLGs), os quais são mais flexíveis que os modelos normais, exigindo que a distribuição da variável resposta pertença à família exponencial, e por último o GAMLSS que é uma técnica de regressão mais geral que permite que a variável resposta, contínua ou discreta, não necessarimente pertença à família exponencial, que possa ter uma grande assimetria e curtose, e que todos os parâmetros da variável resposta sejam modelados de forma linear ou não-linear. Também foi mostrada a teoria de influência local proposta por Cook (1986), a qual permite identificar os potenciais pontos influentes no conjunto de dados.

No próximo capítulo, a fim de considerar a diferença existente nos preços dos imóveis, ou seja, o valor do imóvel ofertado e vendido, será abordada a análise de sobrevivência e suas principais características.

Capítulo

3

Análise de Sobrevivência para Dados Imobiliários com Censura à Esquerda

No mercado imobiliário os preços dos imóveis também carregam distorções importantes. Uma delas diz respeito à estratégia de negociações, pois ao se coletar os dados, existem os imóveis efetivamente negociados (vendidos) e outros em negociação (não vendidos). Os modelos comumente estudados para analisar dados imobiliários, como modelos de regressão linear e modelos lineares generalizados não se preocupam em analisar essa diferença podendo sofrer algumas limitações.

Ferraudo (2008) propôs uma metodologia mais flexível que permitiu incorporar no processo da modelagem também os imóveis em negociação. Para isso considerou-se a presença de censura nos dados, ou seja, o preço dos imóveis é dividido em dois tipos: já negociados (vendidos) e anunciados (não vendidos), onde os valores dos imóveis vendidos são observados e os não vendidos são considerados censurados à esquerda. A análise de sobrevivência em geral, considera a variável resposta como o tempo até a ocorrência de um evento de interesse, denominado tempo de falha. Neste caso, a falha é representada pelo valor da venda em reais do imóvel. O que se deseja saber é qual o preço (V) mais provável que o imóvel poderá ser vendido sendo que ele foi ofertado por um valor (Z), sendo que $V \leq Z$ (Ferraudo, 2008). O que se deseja agora é utilizar o *GAMLSS* no processo de estimação.

3.1 Análise de Sobrevivência

A análise de sobrevivência ou de confiabilidade consiste em uma coleção de procedimentos estatísticos para a análise de dados relacionados ao tempo até a ocorrência de um determinado

evento de interesse, a partir de um tempo inicial pré-estabelecido. Ela é caracterizada pela presença de dados incompletos da resposta, chamados de dados censurados.

3.1.1 Tipos de Censura

A censura ocorre quando o tempo até o evento para alguns indivíduos, não é observado (Lawless, 1982), ou quando o estudo, ou experimento, deve ser encerrado e ainda existem itens funcionando, ou indivíduos vivos, cujos tempos de sobrevivência, obviamente, ainda não foram observados.

Colosimo e Giolo (2006) ressaltam o fato de que, toda informação, mesmo as que não foram observadas (censurados), devem ser usados na análise estatística. Duas razões justificam tal procedimento: (i) mesmo sendo incompletas, as observações censuradas fornecem informações sobre a variável resposta; (ii) a omissão das censuras no cálculo das estatísticas de interesse pode acarretar conclusões viciadas.

Existem diversos mecanismos de censura, dentre os quais podem ser citados:

- Censura do Tipo I: o experimento ocorre até um tempo pré-determinado.
- Censura do Tipo II: o experimento ocorre até se obter um número pré-estabelicido de falhas.
- Censura Aleatória: ocorre quando um indivíduo é retirado do experimento sem que tenha acontecido o evento de interesse.

As censuras também podem ser informativas ou não informativas. Suponha-se que em determinado experimento deseja-se testar o efeito de um novo medicamento, no entanto, no decorrer do estudo, alguns pacientes começaram a sofrer complicações de saúde, e precisaram ser retirados do estudo, neste caso, trata-se de uma censura informativa. Já a censura não informativa refere-se ao fato desta ser independente do mecanismo que causa a falha.

Outra característica é que as censuras também podem ser do tipo à direita, à esquerda ou intervalar. No mecanismo de censura à direita, o tempo de ocorrência do evento de interesse está à direita do tempo registrado, ou seja, até o final do estudo a unidade experimental não falhou. A censura intervalar é mais geral e ocorre quando o tempo de falha pertence a um intervalo e não é exatamente conhecido.

Por fim tem-se a censura à esquerda, a qual será considerada neste trabalho. Ela ocorre quando, por exemplo, o tempo registrado é maior do que o tempo de falha. Isto é, o evento de interesse já aconteceu quando o indivíduo foi observado (Colosimo e Giolo, 2006). Por exemplo, considere o estudo do valor dos imóveis no qual o interesse é o valor da venda. Ao iniciar a pesquisa, alguns imóveis já tinham sido vendidos, o que caracteriza a censura à esquerda. (Ferraudo, 2008)

No mercado imobiliário existem situações em que o valor do imóvel coletado pode ser o valor do imóvel já negociado ou em negociação. O valor do imóvel negociado é considerado um valor observado e o valor do imóvel em negociação é um valor caracterizado como censura

à esquerda. Por isso é muito importante diferenciar o valor de um imóvel negociado do ainda em negociação.

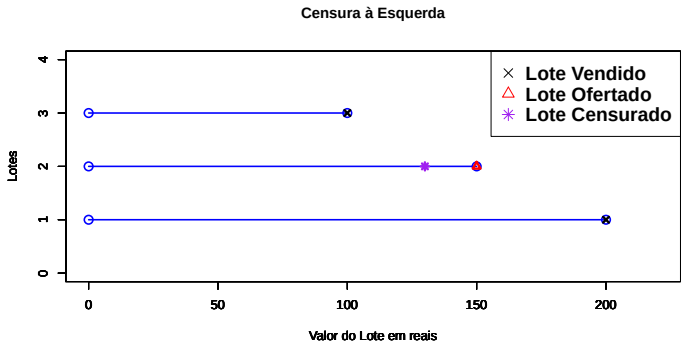


Figura 3.1: Mecanismo de censura à esquerda no valor de lotes urbanos. (Adaptação de Ferraudo (2008))

3.2 Funções de interesse

Considere o esquema de censura como não informativo à esquerda, ou seja, a ocorrência da censura é independente do mecanismo que provoca a falha. Os dados são representados pelo par $(\boldsymbol{V}, \boldsymbol{\delta})$, constituído das n observações dos valores dos imóveis sendo $\boldsymbol{V} = (V_1, ..., V_n)^T$, provenientes da mesma distribuição de probabilidade, e a variável indicadora $\boldsymbol{\delta} = (\delta_1, ..., \delta_n)^T$, com δ_i, $i = 1, ...n$, dada por

$$\delta_i = \begin{cases} 1, \text{se o imóvel for vendido} \\ 0, \text{caso contrário.} \end{cases}$$

A variável V, contínua e não negativa, que representa o tempo de falha é descrita, em geral, pela sua função densidade de probabilidade $f(v)$, pela sua função de sobrevivência $S(v)$, ou pela função de risco (taxa de falha), $h(v)$, as quais são relacionadas.

A função *densidade de probabilidade* é definida como o limite da probabilidade de um imóvel ser vendido no intervalo de valor $[v, v + \Delta v)$ por unidade de valor e é dada por

$$f(v) = \lim_{\Delta v \to \infty} \frac{P(v \leq V, v + \Delta v)}{\Delta v}, \tag{3.1}$$

onde $f(v) \geq 0$ para todo v e $\int_0^\infty f(u)\, du$. A função de distribuição acumulada de V é dada por $F(V) = \int_0^v f(u) du$.

A função de *sobrevivência* é definida como a probabilidade de um imóvel permanecer à venda pelo menos até atingir um valor v pré estabelecido, isto é,

$$S(v) = P(V \geq v) = \int_v^\infty f(u) du = 1 - F(v), \tag{3.2}$$

onde $S(v)$ possuí as seguintes propriedades:

- é não crescente;
- $S(0) = 1$;
- $\lim_{v\to\infty} S(v) = 0$.

A função *de risco ou taxa de falha* definida como o limite da probabilidade de um imóvel ser vendido no intervalo de valor $[v, v + \Delta v)$, dado que o mesmo não foi vendido até o valor v , a qual é expressa por

$$h(v) = \lim_{\Delta v\to 0} \frac{P(v \leq V < v + \Delta v | V \geq v)}{\Delta v}. \tag{3.3}$$

A função de risco descreve a forma como a taxa instantânea de falha muda com o tempo. Graficamente a função de risco pode apresentar comportamento, constante, crescente, decrescente e até mesmo formas não monótonas. Algumas funções são usuais para descrever os tempos de vida, como por exemplo, a distribuição exponencial que apresenta forma de risco constante, a distribuição Weibull que modela as formas crescente, decrescente e constante. (Calsavara, 2011)

O relacionamento entre as três funções para representar o comportamento do valor de permanência à venda pode ser expresso como:

- $f(v) = -\frac{dS(v)}{dv}$;
- $H(v) = \int_0^v h(u)du$;
- $h(v) = \frac{f(v)}{S(v)}$;
- $S(v) = \exp(-H(v))$,

sendo $H(v)$ conhecida como função de *taxa (ou risco) acumulada*.

3.3 Estimador Kaplan-Meier

Seja um conjunto de dados com censura, o estimador não paramétrico de Kaplan-Meier (Kaplan e Meier, 1958), também conhecido como estimador produto-limite é um dos mais utilizados em estudos clínicos e de acordo com Colosimo e Giolo (2006) vem ganhando cada vez mais espaço em estudos de confiabilidade. Utilizado para estimar a função de sobrevivência, ele é uma adaptação da função empírica que considera tantos intervalos de valor quanto forem o número de vendas distintas. Os limites dos intervalos de tempo (valor do imóvel) são os tempos de falha (vendidos) da amostra. Considere:

- $v_1 \leq v_2 \leq ... \leq v_r$ os r valores distintos e ordenados de falha,
- v_r é o maior valor de permanência à venda menor ou igual a v,
- d_i o número de falhas em v_i,

- n_i é o número de imóveis não vendidos até o valor v_i, ou seja, número de imóveis em risco em v_i.

O estimador de Kaplan-Meier (KM) é definido por:

$$\widehat{S}(v) = \prod_{i;v_i \leq v} \left(\frac{n_i - d_i}{n_i} \right) = \prod_{i;v_i \leq v} \left(1 - \frac{d_i}{n_i} \right). \tag{3.4}$$

A expressão 3.4 é uma função escada com degraus nos valores observados de falha de tamanho $1/n$, em que n é o tamanho da amostra. Na ausência de censuras, o estimador de Kaplan-Meier da função de permanência à venda se reduz a,

$$\widehat{S}(v) = \frac{\text{Número de imóveis com valores de permanência à venda} > v}{\text{Número total de imóveis}}. \tag{3.5}$$

O estimador de Kaplan-Meier da função de risco acumulado no intervalo de valor $(0, v]$ é dado por

$$\widehat{H}(v) = -ln \left\{ \widehat{S}(v) \right\}.$$

3.4 Modelos Probabilísticos

As distribuições Exponencial, Weibull, Log-Normal e Log-Logística são as mais utilizadas na modelagem de dados de sobrevivência. Entretanto, várias outras distribuições têm sido consideradas. Entre elas podemos citar as distribuições gama, gama generalizada e a F generalizada. A seguir será apresentada a distribuição Weibull, a qual é utilizada neste trabalho.

3.4.1 Distribuição Weibull

A distribuição Weibull é considerada relativamente simples e apresenta uma variedade de formas, as quais mantém a propriedade básica da função taxa de falha ser monótona. É uma das mais utilizadas tanto em dados biomédicos como industriais e foi proposta originalmente por Weibull (1939). Aqui sua utilização é avaliada para modelar o valor total, em reais, do lote, cuja densidade pode ser escrita na forma

$$f(v) = \frac{\gamma}{\alpha^\gamma} v^{\gamma-1} \text{exp} \left\{ - \left(\frac{v}{\alpha} \right)^\gamma \right\}, \tag{3.6}$$

com $\gamma \geq 0$ e $\alpha \geq 0$, parâmetros de forma e escala respectivamente. Quando $\gamma = 1$, a distribuição Exponencial é obtida como caso particular da Weibull.

As funções de sobrevivência (permanência à venda) e de risco, e os percentis da distribuição Weibull são dados, respectivamente, por

$$S(v) = \text{exp} \left\{ - \left(\frac{v}{\alpha} \right)^\gamma \right\}, \tag{3.7}$$

$$h(v) = \frac{\gamma}{\alpha^{\gamma}} v^{\gamma-1}, \tag{3.8}$$

$$v_p = \alpha \left[-log(1-p)\right]^{1/\gamma}. \tag{3.9}$$

A forma da função de risco é de extrema importância em estudos com dados de sobrevivência. Uma das características importantes da distribuição Weibull na modelagem de tempos de sobrevivência está relacionada à sua flexibilidade em acomodar diferentes formas de funções de risco. Para o parâmetro de forma $\gamma < 1$ as funções de risco são monótonas decrescentes, para $\gamma > 1$ as funções de risco são monótonas crescentes e para $\gamma = 1$ a distribuição exponencial apresenta função de risco constante. (Louzada-Neto et al., 2002).

As expressões para a média e a variância da Weibull incluem o uso da função gama e são expressas por

$$E(V) = \alpha\Gamma\left[1 + (1/\gamma)\right],$$

$$\text{Var}(V) = \alpha^2\left[\Gamma\left[1 + (2/\gamma)\right] - \Gamma\left[1 + (1/\gamma)\right]^2\right],$$

sendo $\Gamma(k) = \int_0^{\infty} x^{k-1}\text{exp}\left\{-x\right\} dx$.

O comportamento das funções de densidade, sobrevivência e risco da Weibull podem ser observados na Figura 3.2.

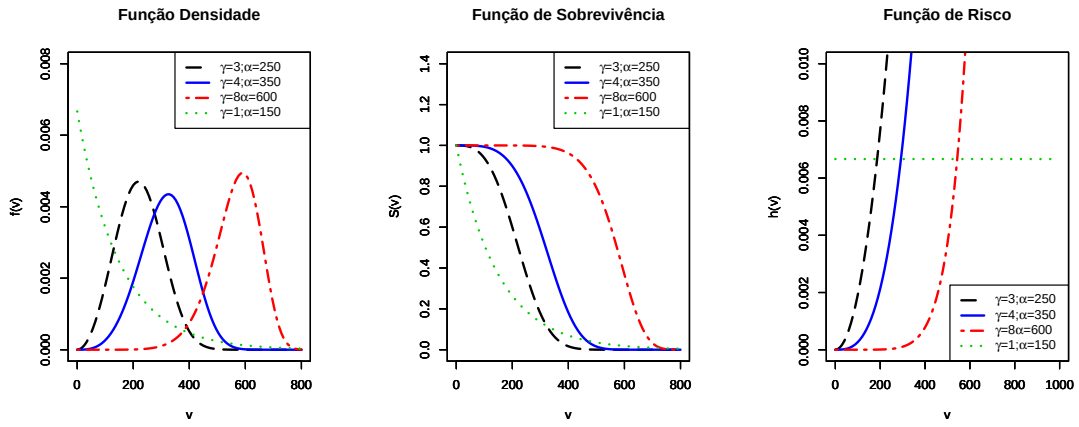


Figura 3.2: (a) Função Densidade de Probabilidade da Weibull, (b) Função de Sobrevivência da Weibull, (c) Função de Risco da Weibull.

3.4.1.1 Linearização no modelo Weibull

De acordo com Colosimo e Giolo (2006), este é um método gráfico que consiste na linearização da função de sobrevivência, ou seja, caso o modelo esteja adequado, o gráfico obtido se aproxima de uma reta.

Considere a função de sobrevivência 3.7 para o modelo Weibull com parâmetros γ e α. Desse modo,

$$-\log\left[S(v)\right] = \left(\frac{v}{\alpha}\right)^{\gamma},$$

e,

$$\log\left[-\log\left[S(v)\right]\right] = -\gamma\log(\alpha) + \gamma\log(v),$$

o que mostra que $\log\left[-\log\left[S(v)\right]\right]$ é uma função linear de $\log(v)$.

Assim, o gráfico de $\log\left[-\log\left[\widehat{S}(v)\right]\right]$ versus $\log(v)$, sendo $\widehat{S}(v)$ o estimador de Kaplan-Meier, deve ser aproximadamente linear, se o modelo Weibull for apropriado. Se além de linear, o gráfico passar pela origem e tiver inclinação igual a 1, é uma indicação a favor do modelo exponencial. (Colosimo e Giolo, 2006)

3.4.2 Modelo de Regressão Weibull

Assim como nos modelos apresentados no Capítulo 2, a análise de sobrevivência muitas vezes envolve covariáveis que estão correlacionadas com o tempo (valor) de sobrevivência. Dessa forma, uma maneira de verificar o efeito dessas covariáveis é utilizar um modelo apropriado para dados censurados o qual considera o tempo de sobrevivência representado pelo Valor Total, em reais do imóvel e a variável indicadora de censura.

Neste contexto, considerando p covariáveis, o modelo abordado será o Weibull que é representado por

$$Y = log(V) = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p + \sigma\nu = x^{'}\boldsymbol{\beta} + \sigma\nu, \tag{3.10}$$

onde $\boldsymbol{x}^{'} = (1, x_1, ..., x_p)$, $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_p)$, σ é o parâmetro de escala e $\nu = log(\epsilon)$, sendo ϵ o erro aleatório com distribuição normal. No caso do modelo Weibull, o erro ν segue uma distribuição do valor extremo padrão ($f(\nu) = \exp\{\nu - exp\{\nu\}\}$), para maiores informações sobre essa distribuição consultar Lawless (1982).

Sabe-se que para $\log(V)$ ter uma distribuição do valor extremo com parâmetro de escala σ, V deve ter uma distribuição Weibull. Assim, a função de sobrevivência para Y condicional a $\boldsymbol{x}$ é dada por

$$S(y|\boldsymbol{x}) = \exp\left\{-\exp\left\{\frac{y - \boldsymbol{x}^{'}\boldsymbol{\beta}}{\sigma}\right\}\right\} \tag{3.11}$$

e, para V condicional a $\boldsymbol{x}$, por:

$$S(v|\boldsymbol{x}) = \exp\left\{-\left(\frac{v}{\exp\{\boldsymbol{x}^{'}\boldsymbol{\beta}\}}\right)^{1/\sigma}\right\}. \tag{3.12}$$

3.4.3 Inferência

Segundo Colosimo & Giolo (2006), existem vários métodos de estimação, no entanto, para modelos de análise de sobrevivência, o método mais utilizado é o de máxima verossimilhança,

por ser capaz de incorporar dados censurados e possuir propriedades ótimas para grandes amostras, como por exemplo a normalidade assintótica dos estimadores de máxima verossimilhança.

A seguir será definida de forma geral a construção do método de estimação de máxima verossimilhançada e em detalhes a função de verossimilhança da distribuição Weibull quando se tem dados censurados à esquerda.

Médodo de Máxima Verossimilhança

Suponha uma amostra de observações $v_1, ..., v_n$ de uma certa população de interesse. Suponha, ainda, que a população é caracterizada pela sua função densidade $f(v)$. Considerando que nenhuma observação tenha sido censurada, a função de verossimilhança para um vetor de parâmetro genérico $\boldsymbol{\theta}$ desta população é, então, expressa por:

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} f(v_i, \boldsymbol{\theta})$$

.

O interesse aqui é encontrar os valores de $\boldsymbol{\theta}$ que maximize a função $L(\boldsymbol{\theta})$. A contribuição de cada observação é a sua função de densidade $f(v)$. Para uma observação com censura à direita a contribuição para $L(\boldsymbol{\theta})$ é a sua função de sobrevivência S(v). Neste trabalho considera-se observações censuradas à esquerda, onde a contribuição para $L(\boldsymbol{\theta})$ é a sua função de distribuição acumulada F(v), ou seja, a contribuição é dada por 1-S(v), onde S(v) é a sua função de sobrevivência.

Portanto, considerando que o indivíduo i, com $i = 1, ..., n$ sob estudo é representado pelo par (v_i, δ_i), sendo v_i o tempo de falha ou de censura do i-ésimo indivíduo e δ_i o indicador de censura, a função de verossimilhança pode ser escrita da seguinte maneira:

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} [f(v_i, \boldsymbol{\theta})]^{\delta_i} [1 - S(v_i, \boldsymbol{\theta})]^{1-\delta_i} = \prod_{i=1}^{n} [f(v_i, \boldsymbol{\theta})]^{\delta_i} [F(v_i, \boldsymbol{\theta})]^{1-\delta_i} \tag{3.13}$$

Considere a distribuição Weibull 3.6, uma amostra aleatória $v_1, ..., v_n$ e a variável indicadora de censura δ_i. Então a expressão da sua log-verossimilhança é

$$\begin{aligned} l(\alpha, \gamma; v_i, \delta_i) = \sum_{i=1}^{n} \{ & \delta_i [(\gamma - 1)\ln(v_i) - \gamma \ln(\alpha) + \ln(\gamma) - (\frac{v_i}{\alpha})^{\gamma}] \\ & + (1 - \delta_i)\ln(1 - \exp\{-(\frac{v_i}{\alpha})^{\gamma}\})\}, \end{aligned} \tag{3.14}$$

com $l(\alpha, \gamma; v_i, \delta_i) = \log [L(\alpha, \gamma; v_i, \delta_i)]$.

Para encontrar os estimadores de máxima verossimilhança de 3.14 basta encontrar valores de $\boldsymbol{\theta}$ que maximizem $l(\boldsymbol{\theta})$. Portanto, pode-se encontrá-los resolvendo o sistema de equações:

$$U(\boldsymbol{\theta}) = \frac{\partial l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}. \tag{3.15}$$

Devido a complexidade da função de verossimilhança, algumas vezes o estimador de máxima verossimilhança não possui forma fechada, sendo necessário usar algum procedimento numérico para estimar os parâmetros, como por exemplo, os implementados no *R*.

Como apresentado no Capítulo 2, o modelo *GAMLSS* é bem flexivel e uma vez conhecida a função de distribuição, a função de risco, e a de sobrevivência na qual deseja-se trabalhar, a utilização dessa classe de modelos em dados censurados é feita de maneira bem simples, principlamente porque o ajuste de muitas distribuições já estão implementados no *R*, e para algumas dessas distribuições é possível até encontrar mais de uma parametrização no pacote, como o caso da Weibull.

Assim, para utilizar o GAMLSS na estimação dos parâmetros, basta considerar a log-verossililhança l da função de verossimilhança penalizada 2.42, como a log-verossimilhança 3.14 do modelo Weibull censurado à esquerda. Feito isso, a seleção do modelo e testes de diagósticos são realizados exatamente como os descritos no Capítulo 2.

A seguir serão apresentados o Teste da Razão de Verossimilhança e outras técnicas de diagnósticos utilizadas para dados censurados.

Teste da Razão de Verossimilhança

A forma mais simples e eficiente de selecionar o modelo mais adequado a ser usado para um conjunto de dados é por meio de técnicas gráficas. Entretanto, testes de hipóteses com modelos encaixados também podem ser utilizados para esta finalidade. (Colosimo e Giolo, 2006)

Para um modelo com um vetor $\boldsymbol{\theta} = (\theta_1, ..., \theta_p)$ de parâmetros, muitas vezes há o interesse em testar hipóteses relacionadas a este vetor ou a um subconjunto dele. Três testes são em geral utilizados para esta finalidade: o de Wald, o Escore e o da Razão de Verossimilhanças (Colosimo e Giolo, 2006). A seguir uma breve descrição do Teste da Razão de Verossimilhanças será apresentada. .

Este teste é baseado na função de verossimilhança e envolve a comparação dos valores do logaritmo da função de verossimilhança maximizada sob as seguintes hipóteses:

$$H_0 : \text{o modelo de interesse é adequado} \times H_a : \text{hipótese alternativa.}$$

A estatística do teste é calculada como

$$\text{TRV} = -2log\left[\frac{L(\widehat{\boldsymbol{\theta}}_0)}{L(\widehat{\boldsymbol{\theta}}_a)}\right] = 2\left[L(\widehat{\boldsymbol{\theta}}_a) - L(\widehat{\boldsymbol{\theta}}_0)\right], \tag{3.16}$$

onde $\widehat{\boldsymbol{\theta}}_0$ é o estimador de máxima verossimilhança do modelo de interesse e $L(\widehat{\boldsymbol{\theta}}_a)$ é o estimador de máxima verossimilhança do modelo generalizado. Sabe-se que a estatística da razão de verossimilhança sob H_0, segue aproximadamente uma distribuição qui-quadrado com graus de liberdade igual a diferença de número de parâmetros ($\widehat{\theta}_0$ e $\widehat{\theta}_a$), a qual denota-se por p. Para amostras grandes,H_0 é rejeitada, a um nível $100b\%$ de significância, se TRV $> \chi^2_{p,1-b}$. (Colosimo e Giolo, 2006).

3.4.4 Adequação do Modelo Ajustado

Após o ajuste de um modelo é necessário analisar se o mesmo é adequado,para isso, uma das formas de verificar essa qualidade é através dos resíduos. Em análise de sobrevivência alguns dos resíduos utilizados são: resíduos de Cox-Snell, resíduos padronizados, resíduos *martingal* e resíduos *deviance*. Os dois primeiros servem para examinar o ajuste global, o terceiro para determinar a forma funcional de uma covariável e o último é utilizado para verificar a acurácia do modelo para cada indivíduo sob estudo. A seguir será descrito apenas o resíduo Cox-Snell, para maiores detalhes, consultar Colosimo e Giolo (2006).

Os resíduos de Cox-Snell são quantidades determinadas por

$$\widehat{e}_i = \widehat{h}(v_i|\boldsymbol{x}_i), \tag{3.17}$$

onde $\widehat{h}(.)$ é a função de risco acumulado obtida do modelo ajustado. Para o modelo Weibull, o resíduo é dado por

$$\widehat{e}_i = \left[v_i \exp\left\{-\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}\right\}\right]^{\widehat{\gamma}}.$$

Os resíduos $\widehat{e}_i$ vêm de uma população homogênea e devem seguir uma distribuição exponencial padrão se o modelo for adequado (Lawless, 1982). O gráfico $\widehat{e}_i$ versus $\widehat{h}(\widehat{e}_i)$ deve ser aproximadamente uma reta. (Colosimo e Giolo 2006)

3.5 Considerações Finais

Em geral, os estudos com dados imobiliário são realizados considerando apenas os valores dos imóveis vendidos. Com o objetivo de analisar os valores de imóveis que ainda não foram de fato comercializados, este capítulo introduziu a análise de sobrevivência a qual é caracterizada pela presença de censura nos dados, a distribuição Weibull e algumas propriedades, como por exemplo a monotonicidade da sua função de falha. Também foi abordada a regressão utilizando dados censurados.

Para entender as propriedades assintóticas dos estimadores, no próximo capítulo serão apresentados os estudos de simulação para os modelos de regressão apresentados neste e no Capítulo 2.

Capítulo

4

Estudo de Simulação

O estudo de simulação é utilizado para analisar o comportamento de modelos e verificar o desempenho dos métodos em estudo e suas propriedades, aprender com os erros que possam vir a ser cometidos em um ambiente simulado, ganhar experiência sem risco, criação de dados quando não se é possível obter uma amostra, entre outros. Seu uso é ilimitado e permite que o procedimento seja feito das mais variadas formas cabendo ao pesquisador a escolha de como será desenvolvida a simulação de acordo com os pressupostos que se tem, os resultados que se busca e o problema em questão. (Paiva, Freire e Cecatti (2008); Terra (2009); Vieira (2003))

Considerando a natureza dos dados imobiliários, que em sua maioria a variável resposta é positiva e assimétrica e nos dados reais em estudo as covariáveis são dadas pela área e pela localização que é dividida em variáveis *dummys*, o estudo de simulação, com o objetivo de verificar as propriedades assintóticas dos estimadores, foi realizado da seguinte maneira:

Conforme Ferraudo (2008), gerou-se 1000 conjuntos de dados com diferentes tamanhos de amostras: 30, 60, 200, 500 e 1000. Para cada amostra, foram simuladas 5 covariáveis: $x_1 \sim bernoulli(0.4)$, $x_2 \sim bernoulli(0.7)$, $x_3 \sim bernoulli(0.5)$ e $x_4 \sim bernoulli(0.3)$ que representam a localização do imóvel e $x_5 \sim Weibull(2, 3)$ que representa a variável área, podendo ser assimétrica.

Devido à utilização das diferentes metodologias no trabalho (regressão linear, modelos lineares generalizados, *GAMLSS* e análise de sobrevivência), o estudo de simulação abordou separadamente cada uma delas. Ou seja:

1. No primeiro caso considerando a metodologia linear, a variável resposta valor do lote (R$), possui distribuição normal, $Y \sim Normal$, com média igual ao preditor linear ($\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5$). Como dados imobiliários dificilmente

apresentam variância constante, afim de se estudar o comportamento da mesma, esta foi fixada de três maneiras diferentes: 10^2 (baixa), 100^2 (moderada) e 1000^2 (alta).

2. Para o modelo linear generalizado foi considerada a variável resposta $Y \sim Gama(\nu, \nu/\mu)$ com função de densidade dada por 2.22. Como o ajuste do modelo considerou a função de ligação logarítmica, a média é dada por $\mu = exp(\beta_0+\beta_1 X_1+\beta_2 X_2+\beta_3 X_3+\beta_4 X_4+\beta_5 X_5)$ e para analisar a variância, o parâmetro de forma ν foi fixado de três maneiras diferentes: 10, 100, 1000.

3. No caso do *GAMLSS* o procedimento foi exatamente igual ao item anterior. Foi considerado a distribuição Gama para a variável resposta dada por 2.22, função de ligação logarítmica e ν fixado com valores 10, 100 e 1000.

4. Para a análise de sobrevivência utilizando o *GAMLSS* na estimação dos parâmetros, a característica principal é a inclusão da censura, a qual foi considera como censura à esquerda. Neste estudo foi considerado a distribuição Weibull para variável resposta, $Y \sim Weibull$, com parâmetros de escala, $\alpha = \exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5)$ e de forma $\gamma = 1/2$. Para os diferentes tamanhos de amostras foram consideradas 6 porcentagens de censuras, 0%, 1%, 5%, 15%, 30% e 50%.

Realizou-se o estudo de simulação separadamente para cada variável resposta. Foi obtido o intervalo de confiança de 95% para cada parâmetro baseado na teoria assintótica e verificado se o intervalo de confiança continha o verdadeiro valor do parâmetro com o intuito de se obter a probabilidade de cobertura (PC) dos intervalos de confiança, ou seja, a proporção de vezes que o intervalo de confiança cobriu os 1000 conjuntos de dados para cada um dos parâmetros.

Um teste para a igualdade de proporções pode ser contruído de Casella & Berger, (2002)

$$Z = \frac{\widehat{p} - p^*}{\sqrt{\frac{p^*(1-p^*)}{k}}} \to N(0,1),$$

em que $\widehat{p}$ é a probabilidade de cobertura observada. Supondo um nível de significância de 5%, a região crítica do teste delimitada por η_1 e η_2, é obtida de

$$P(\eta_1 < \widehat{p} < \eta_2 | p^* = 0,95) = 0,95,$$

ou seja,

$$P = \left(\frac{\eta_1 - 0,95}{\sqrt{\frac{0,05\times 0,95}{1000}}} < Z < \frac{\eta_2 - 0,95}{\sqrt{\frac{0,05\times 0,95}{1000}}} \right) = 0,95,$$

onde se obtém $\eta_1 = 0,936$ e $\eta_2 = 0,964$. Assim, rejeita-se a igualdade entre as proporções para as coberturas observadas fora do intervalo (0,936; 0,964).

Para verificar se os estimadores utilizados não apresentam tendenciosidade, foi calculado também o erro quadrático médio (EQM) das estimativas, que é igual à soma entre a variância e o quadrado do viés.

$$EQM(\theta) = \text{Var}(\theta) + B(\theta)^2.$$

O viés é calculado como a diferença entre o valor esperado do estimador e o verdadeiro valor do parâmetro a estimar, denotado por B.

$$B(\theta) = E(\widehat{\theta}) - \theta.$$

Quando $B(\theta) = 0$, ou equivalentemente $E(\widehat{\theta}) = \theta$, o vício do estimador é nulo e neste caso, diz-se que o estimador é não-viesado para θ. Caso $E(\widehat{\theta}) \neq 0$, diz-se que o estimador $\widehat{\theta}$ é viesado, no entanto pode ocorrer que a esperança do estimador se aproxima do verdadeiro valor de θ a medida que aumenta o tamanho da amostra, i.e. $\lim_{n\to\infty} E(\widehat{\theta}) = \theta$. Neste caso, $\widehat{\theta}$ é dito ser uma estimador assintoticamente não viesado para θ.

Considerando todas as situações citadas acima, a seguir serão apresentados as tabelas e os gráficos com os resultados dos estudos de simulação.

4.1 Simulação com ajuste via MRL

Na Tabela 4.1e nas Figuras 4.1 e 4.2 encontram-se as estimativas e a probabilidade de cobertura dos parâmetros β_0, β_1, β_2, β_3, β_4 e β_5 para amostras de tamanho 30,60, 200, 500 e 1000 respectivamente. Os vícios dos parâmetros encontram-se apenas nas Figuras 4.3 e 4.4.

Sabendo que os verdadeiros valores dos parâmetros para o modelo linear foram $\beta_0 = 9.7$, $\beta_1 = -8.1$, $\beta_2 = 3$,$\beta_3 = -2.5$,$\beta_4 = 3$ e $\beta_5 = 4.3$, nota-se que mesmo quando a variância foi moderada (100) e o tamanho da amostra pequeno (n=30), β_5 apresentou um bom desempenho, atingindo o valor 4,47. β_0 atingiu valores razoáveis apenas com variância baixa (10) e variância moderada (100) a partir de amostras médias (n=200). β_1 , β_2, β_3 e β_4 tiveram comportamentos parecidos, ou seja, para variância baixa conseguiram valores próximos dos reais com todos os tamanhos de amostras, no entanto, para variância média (100) os dois primeiros conseguiram resultados satisfatórios apenas a partir de amostras grandes (n=500), enquanto que os dois últimos a partir de amostras médias (n=200).

Por fim, para variância grande todos os parâmetros apresentaram estimativas ruins. Conforme aumenta o tamanho da amostra, as estimativas melhoram e percebe-se também que quando a variância é muito alta, o desempenho do modelo é baixo, o que condiz com a teoria assintótica.

Além da Tabela 4.1, nas Figuras 4.1, 4.2, 4.3 e 4.4 é possível visualizar que, quando se considera uma variância baixa, todos os parâmetros apresentam cobertura dentro dos limites nominais a partir das amostras de tamanho 30. Ao considerar uma variância moderada, β_0 e β_3 apresentam cobertura dentro dos limites nominais a partir das amostras de tamanho 60 e os demais parâmetros a partir de amostras de tamanho 30. Quando a variância é alta β_1, β_2 e β_5 apresenta cobertura dentro dos limites nominais a partir de amostras de tamanho 60 e os demais a partir de amostras de tamanho 30.

Ressalta-se ainda, que foi realizada a probabilidade de cobertura com tamanho de amostra 2000 a fim de verificar a possível tendência dos parâmetros β_0 na presença de baixa variância e para o parâmetro β_3 com variâncias baixa e moderada. Nenhuma delas apresentou valores fora dos limites nominais.

Com relação ao EQM, verificou-se que conforme aumenta a variância (moderada e alta) seus valores também aumentam. Nota-se que ao aumentar o tamanho da amostra, os EQMs decrescem, como o esperado.

Por fim, conclui-se que a performance do estudo de simulação para MRL coincide com a teoria assintótica dos estimadores, pois conforme aumenta a amostra, melhora o desempenho do modelo e também diminui o viés dos estimadores. Lembrando ainda que, por se trabalhar com simulação, os valores utilizados foram escolhidos de tal forma que fossem coerentes para verificar se estavam ou não contidos nos intervalos baseados em valores práticos conforme os últimos capítulos e aplicação em dados reais.

Tabela 4.1: Probabilidade de cobertura (PC) dos intervalos de confiança de 95% e estimativa dos parâmetros do modelo linear para diferentes variâncias e tamanhos de amostras.

		Tamanho da amostra									
		30	60	200	500	1000	30	60	200	500	1000
σ^2	$\beta's$	Estimativa					PC				
10^2	β_0	9,98	9,77	9,71	9,66	9,71	0,94	0,96	0,95	0,96	0,96
	β_1	-7,86	-8,09	-8,17	-8,06	-8,11	0,94	0,95	0,95	0,95	0,95
	β_2	2,73	3,04	2,95	3,01	2,99	0,95	0,94	0,94	0,96	0,94
	β_3	-2,52	-2,45	-2,51	-2,49	-2,50	0,95	0,96	0,95	0,95	0,97
	β_4	2,99	2,90	3,04	2,97	3,00	0,95	0,94	0,95	0,95	0,95
	β_5	4,25	4,27	4,33	4,30	4,30	0,95	0,95	0,95	0,97	0,95
100^2	β_0	8,50	8,29	9,30	10,16	9,47	0,93	0,95	0,94	0,96	0,94
	β_1	-7,65	-7,53	-7,64	-8,38	-8,35	0,95	0,96	0,94	0,96	0,95
	β_2	4,01	4,21	3,29	2,81	3,41	0,94	0,95	0,95	0,96	0,96
	β_3	-3,49	-1,62	-2,36	-2,07	-2,44	0,93	0,95	0,94	0,94	0,95
	β_4	1,18	0,73	3,07	2,93	3,08	0,94	0,95	0,95	0,95	0,94
	β_5	4,47	4,51	4,31	4,18	4,29	0,94	0,94	0,95	0,96	0,95
1000^2	β_0	5,73	6,55	10,77	9,31	3,60	0,94	0,96	0,94	0,95	0,95
	β_1	-25,52	-3,38	-9,13	-10,06	-7,00	0,93	0,94	0,96	0,95	0,95
	β_2	9,25	-0,68	7,28	8,42	6,59	0,93	0,95	0,95	0,95	0,95
	β_3	-6,81	-17,18	-12,61	0,00	-3,50	0,94	0,94	0,96	0,95	0,96
	β_4	6,94	5,02	3,88	2,36	-2,06	0,96	0,94	0,96	0,95	0,96
	β_5	5,51	8,17	5,32	3,64	6,45	0,93	0,95	0,95	0,94	0,95

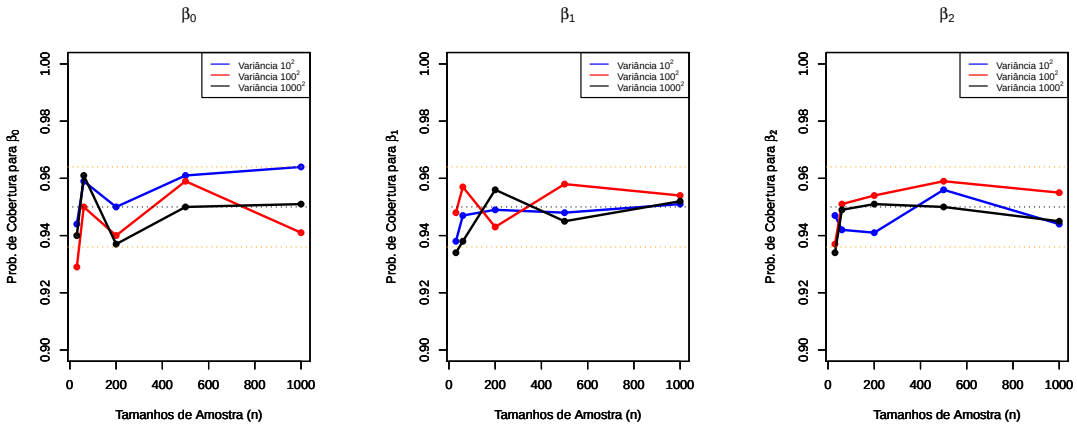


Figura 4.1: Probabilidade de Cobertura dos intervalos de 95% versus o tamanho da amostra dos parâmetros β_0, β_1 e β_2 para o modelo normal com diferentes variâncias.

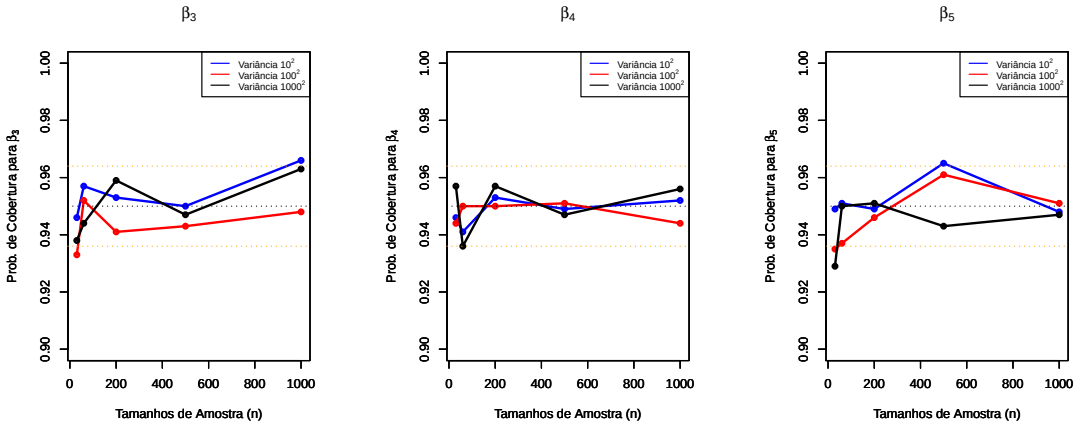


Figura 4.2: Probabilidade de Cobertura dos intervalos de 95% versus o tamanho da amostra dos parâmetros β_3, β_4 e β_5 para o modelo normal com diferentes variâncias.

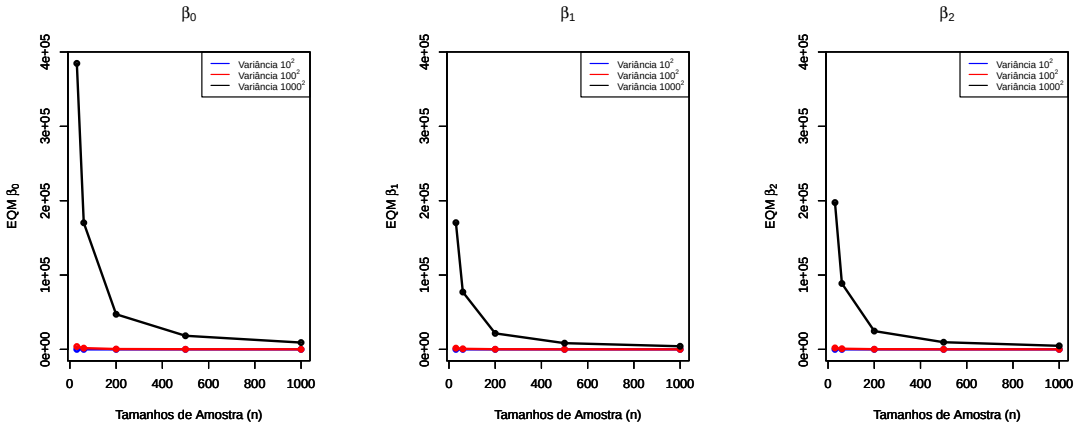


Figura 4.3: EQM das estimativas dos parâmetros β_0, β_1 e β_2 para o modelo normal com diferentes variâncias.

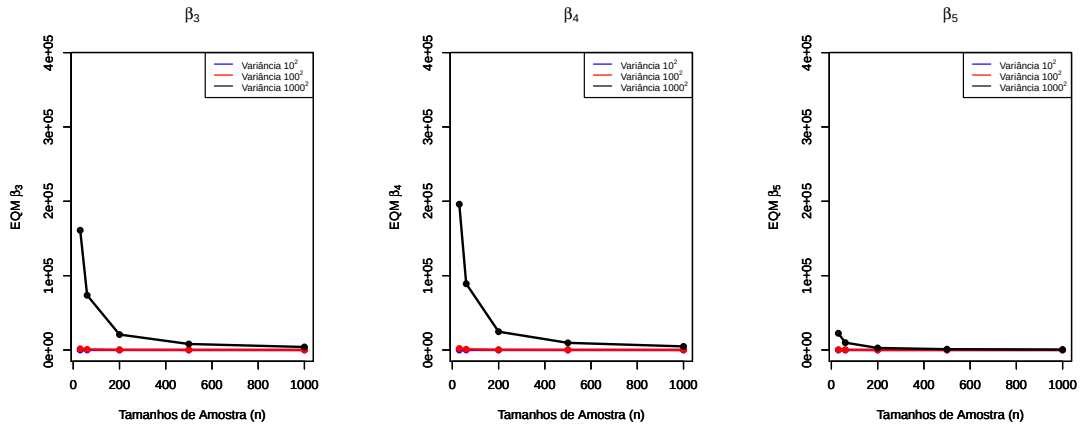


Figura 4.4: EQM das estimativas dos parâmetros β_3, β_4 e β_5 para o modelo normal com diferentes variâncias.

4.2 Simulação com ajuste via MLG's

Uma forma de verificar a adequabilidade do estudo de simulação é analisar o decaimento da variância das estimativas conforme o aumento da amostra, o qual deve ser da ordem $n^{-1/2}$. Para isso, pode-se ajustar um modelo de regressão linear $log(var(\beta)) = a + blog(n)$, onde as estimativas de b devem ser de aproximadamente -1. (Cremasco, 2005)

Na Tabela 4.2 encontram-se as estimativas de a e b e seus respectivos desvio padrão entre parênteses. As probabilidades de cobertura dos intervalos de confiança de 95% e as estimativas dos parâmetros do modelo são apresentados na Tabela 4.3 e nas Figuras 4.5 e 4.6.

Sabendo que os verdadeiros valores dos parâmetros para o modelo foram $\beta_0 = 2$, $\beta_1 = 0.1$, $\beta_2 = -0.2$, $\beta_3 = 0.6$, $\beta_4 = 0.1$ e $\beta_5 = 0.2$, nota-se que independente do tamanho da amostra ou do valor do parâmetro de forma, todos os $\beta^{'}s$ apresentaram estimativas próximas dos valores reais dos parâmetros. Observa-se também que a partir de amostras de tamanho 60, todos os parâmetros apresentam cobertura dentro dos limites nominais independente do tamanho do parâmetro de forma ν.

Com relação aos EQMs, Figuras 4.7 e 4.8, verificou-se que conforme diminui o parâmetro de forma, aumentam os valores dos EQMs. Nota-se que ao aumentar o tamanho da amostra, o EQMs decrescem, como o esperado. Pela Tabela 4.2 verifica-se que as estimativas de b do modelo $log(var(\beta)) = a + blog(n)$ estão próximas de -1, o que significa que:

- Confirma que o decaimento da variância das estimativas dos coeficientes é da ordem de $n^{-1/2}$;
- A simulação está correta. Isso é esperado, particularmente em se tratando de uma estrutura direcionada por uma aproximação de 2ª ordem;
- As probabilidades de cobertura estão adequadas, resguardando os casos em que o n é muito pequeno. Entretanto, a sub cobertura é na ordem de até (aproximadamente) $2,5\%$ em relação a probabilidade de cobertura nominal de 95%.

Tabela 4.2: Estimativa de a e b do modelo linear $log(var(\beta)) = a + b\log(n)$ para diferentes valores do parâmetro de forma ν para o estudo do modelo linear generalizado.

	$l(var(\beta_0))$	$l(var(\beta_1))$	$l(var(\beta_2))$	$l(var(\beta_3))$	$l(var(\beta_4))$	$l(var(\beta_5))$	ν
a	0,2856	-0,5578	-0,3851	-0,6137	-0,3562	-2,5190	
	(0.08112)	(0.06401)	(0.07333)	(0.05828)	(0.08638)	(0.09100)	
b	-1,0592	-1,0492	-1,0553	-1,0468	-1,0600	-1,0686	10
	(0.01518)	(0.01198)	(0.01372)	(0.01091)	(0.01617)	(0.01703)	
a	-2,0282	-2,8570	-2,6836	-2,8987	-2,6834	-4,8124	
	(0.08264)	(0.07127)	(0.07922)	(0.07062)	(0.07748)	(0.09537)	
b	-1,0570	-1,0496	-1,0556	-1,0495	-1,0557	-1,0690	100
	0.01547	(0.01334)	(0.01483)	(0.01322)	(0.01450)	(0.01785)	
a	-4,2905	-5,1197	-4,9390	-5,1638	-4,9350	-7,1199	
	0.08600)	(0.07527)	(0.09297)	(0.08060)	(0.08886)	(0.08441)	
b	-1,0639	-1,0560	-1,0635	-1,0556	-1,0640	-1,0687	1000
	0.01609)	(0.01409)	(0.01740)	(0.01508)	(0.01663)	(0.01580)	

Tabela 4.3: Probabilidade de cobertura (PC) dos intervalos de confiança de 95% e estimativa dos parâmetros do MLG considerando a distribuição Gama para diferentes valores de ν e tamanhos de amostras.

		Tamanho da amostra									
		30	60	200	500	1000	30	60	200	500	1000
ν	$\beta's$	Estimativa					PC				
	β_0	2,00	2,00	2,00	2,00	2,00	0,94	0,96	0,95	0,96	0,95
	β_1	0,11	0,10	0,10	0,10	0,10	0,93	0,94	0,96	0,95	0,95
10	β_2	-0,20	-0,20	-0,20	-0,20	-0,20	0,91	0,95	0,95	0,96	0,95
	β_3	0,60	0,60	0,60	0,60	0,60	0,93	0,95	0,95	0,96	0,95
	β_4	0,10	0,09	0,10	0,10	0,10	0,94	0,95	0,96	0,96	0,95
	β_5	0,20	0,20	0,20	0,20	0,20	0,96	0,94	0,96	0,95	0,95
	β_0	2,00	2,00	2,00	2,00	2,00	0,94	0,95	0,96	0,94	0,95
	β_1	0,10	0,10	0,10	0,10	0,10	0,93	0,95	0,96	0,95	0,94
100	β_2	-0,20	-0,20	-0,20	-0,20	-0,20	0,94	0,95	0,95	0,95	0,96
	β_3	0,60	0,60	0,60	0,60	0,60	0,94	0,94	0,95	0,94	0,94
	β_4	0,10	0,10	0,10	0,10	0,10	0,93	0,94	0,94	0,95	0,96
	β_5	0,20	0,20	0,20	0,20	0,20	0,95	0,95	0,94	0,96	0,93
	β_0	2,00	2,00	2,00	2,00	2,00	0,93	0,94	0,95	0,95	0,94
	β_1	0,10	0,10	0,10	0,10	0,10	0,94	0,96	0,95	0,95	0,95
1000	β_2	-0,20	-0,20	-0,20	-0,20	-0,20	0,94	0,95	0,96	0,95	0,95
	β_3	0,60	0,60	0,60	0,60	0,60	0,94	0,93	0,94	0,96	0,95
	β_4	0,10	0,10	0,10	0,10	0,10	0,93	0,95	0,95	0,96	0,96
	β_5	0,20	0,20	0,20	0,20	0,20	0,93	0,94	0,96	0,94	0,96

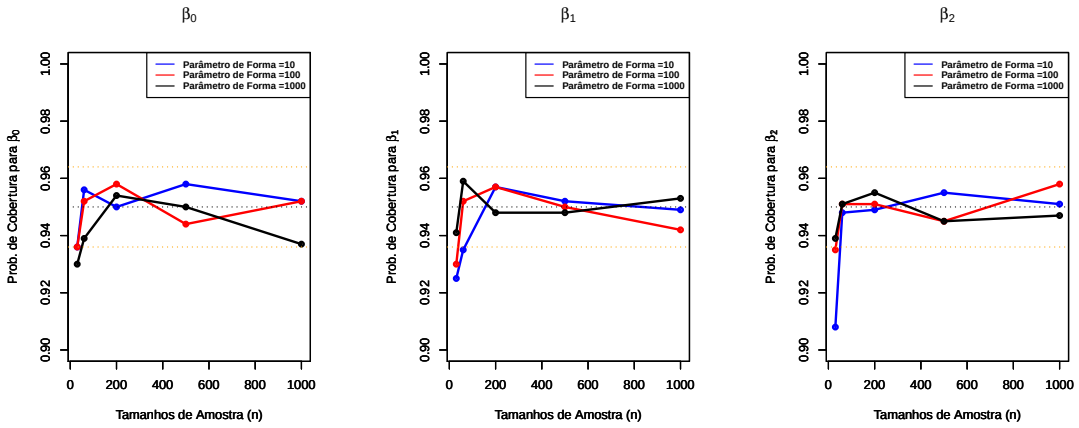


Figura 4.5: Probabilidade de Cobertura dos intervalos de 95% versus o tamanho da amostra dos parâmetros β_0, β_1 e β_2 do modelo gama.

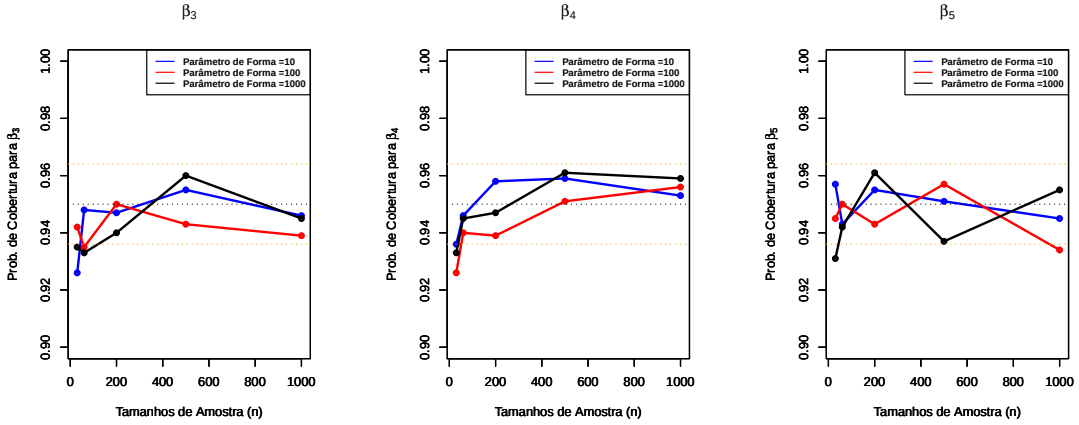


Figura 4.6: Probabilidade de Cobertura dos intervalos de 95% versus o tamanho da amostra dos parâmetros β_3, β_4 e β_5 do modelo gama.

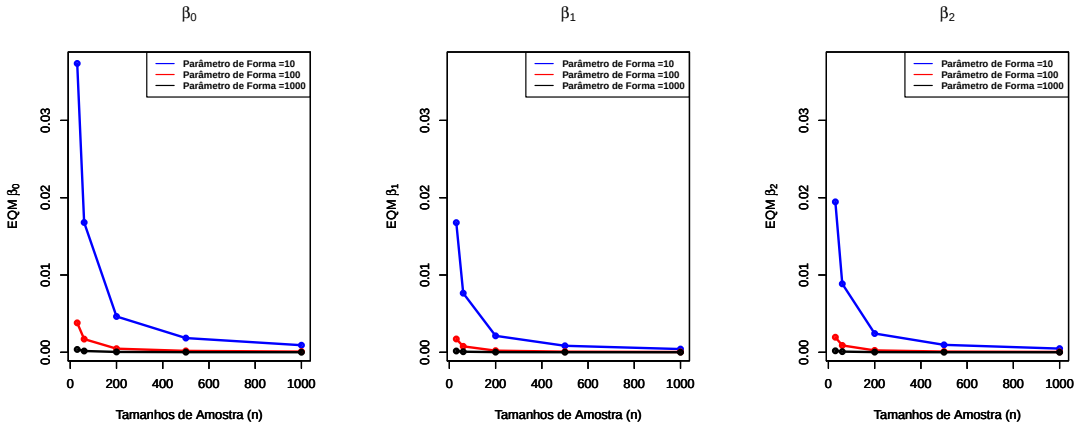


Figura 4.7: EQM das estimativas dos parâmetros β_0, β_1 e β_2 do modelo gama.

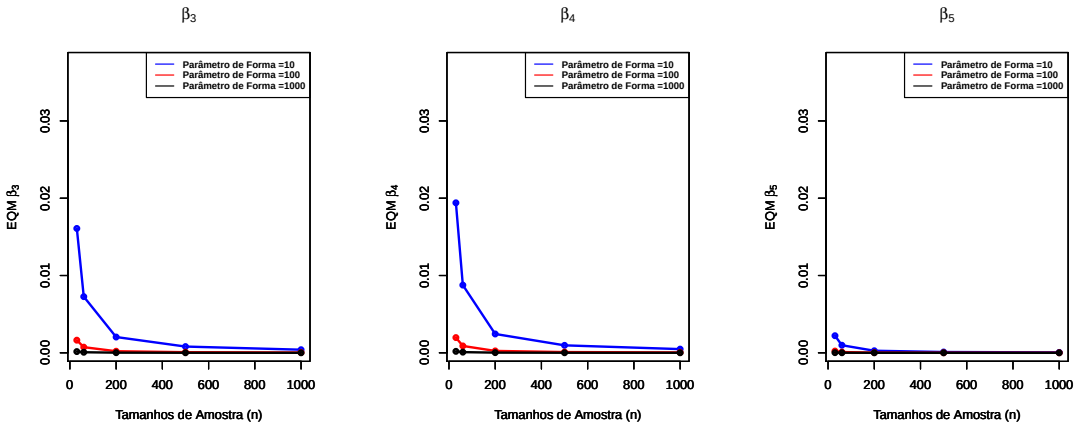


Figura 4.8: EQM das estimativas dos parâmetros β_3, β_4 e β_5 do modelo gama.

4.3 Simulação com ajuste via GAMLSS

Na Tabela 4.4 encontram-se as estimativas do modelo linear $log(var(\beta)) = a + blog(n)$. Observa-se que os valores de b se aproximam de -1 confirmando que o decaimento da variância é da ordem de $n^{-1/2}$.

As probabilidades de cobertura dos intervalos de confiança de 95% e as estimativas dos parâmetros do modelo são apresentados na Tabela 4.5 e nas Figuras 4.9, 4.10 e 4.11. Os EQMs encontram-se nas Figuras 4.12, 4.13 e 4.14.

Verifica-se que o comportamento da classe de modelos *GAMLSS* é parecido com a da classe dos modelos lineares generalizados, apresentando probabilidade de cobertura dentro dos limites nominais para a maioria dos casos, não sendo indicada para tamanhos pequenos de amostras. Nota-se ainda que conforme o tamanho da amostra aumenta, as estimativas dos parâmetros melhoram e os EQMs diminuem.

Uma característica do modelo *GAMLSS* é que ele modela todos os parâmetros da distribuição, e portanto, apresenta um parâmetro a mais referente ao intercepto do modelo de ν da distribuição gama, o qual foi denominado σ_0. Dessa forma, ao mudar o valor de ν que foi fixado como 10,100 e 1000, os valores inicias de σ_0 também foi alterados. Os valores iniciais para σ_0 foram: $-1,1$ quando $\nu = 10$, $-2,3$ quando $\nu = 100$ e -3.5 quando $\nu = 1000$. Concluí-se que o comportamento de σ_0 e dos demais parâmetros são coerentes com a teoria assintótica,

Tabela 4.4: Estimativa de a e b do modelo linear $log(var(\beta)) = a + b\log(n)$ para diferentes valores do parâmetro de forma ν para o estudo do modelo *GAMLSS*.

	$l(var(\beta_0))$	$l(var(\beta_1))$	$l(var(\beta_2))$	$l(var(\beta_3))$	$l(var(\beta_4))$	$l(var(\beta_5))$	$l(var(\sigma_0))$	ν
a	-0,0319	-0,8762	-0,7035	-0,9302	-0,6753	-2,8265	-0,7149	
b	-1,0101	-1,0000	-1,0061	-0,9979	-1,0107	-1,0210	-1,0017	10
a	-2,3955	-3,2256	-3,0522	-3,2672	-3,0519	-5,1792	-0,6953	
b	-1,0004	-0,9929	-0,9988	-0,9927	-0,9990	-1,0125	-1,0002	100
a	-4,6655	-5,4946	-5,3139	-5,5387	-5,3096	-7,4950	-0,6934	
b	-1,0061	-0,9982	-1,0057	-0,9978	-1,0063	-1,0108	-1,0000	1000

Tabela 4.5: Probabilidade de cobertura dos intervalos de confiança de 95% e estimativa dos parâmetros do modelo *GAMLSS* para diferentes valores de ν e tamanhos de amostras.

		Tamanho da amostra									
		30	60	200	500	1000	30	60	200	500	1000
ν	$\beta^{'}s$	Estimativa					PC				
10	β_0	1,98	1,99	2,00	2,00	2,00	0,92	0,94	0,95	0,96	0,95
	β_1	0,10	0,10	0,10	0,10	0,10	0,92	0,96	0,95	0,95	0,95
	β_2	-0,19	-0,20	-0,20	-0,20	-0,20	0,92	0,95	0,95	0,95	0,95
	β_3	0,60	0,60	0,60	0,60	0,60	0,93	0,94	0,95	0,96	0,95
	β_4	0,10	0,10	0,10	0,10	0,10	0,91	0,95	0,96	0,96	0,96
	β_5	0,20	0,20	0,20	0,20	0,20	0,92	0,94	0,96	0,95	0,95
	σ_0	-1,28	-1,21	-1,17	-1,16	-1,15	0,81	0,89	0,93	0,95	0,94
100	β_0	2,00	2,00	2,00	2,00	2,00	0,92	0,92	0,94	0,94	0,95
	β_1	0,10	0,10	0,10	0,10	0,10	0,94	0,95	0,94	0,96	0,94
	β_2	-0,20	-0,20	-0,20	-0,20	-0,20	0,92	0,93	0,95	0,95	0,96
	β_3	0,60	0,60	0,60	0,60	0,60	0,93	0,95	0,94	0,94	0,95
	β_4	0,10	0,10	0,10	0,10	0,10	0,93	0,94	0,93	0,95	0,96
	β_5	0,20	0,20	0,20	0,20	0,20	0,92	0,95	0,95	0,94	0,95
	σ_0	-2,43	-2,36	-2,32	-2,31	-2,31	0,81	0,87	0,93	0,94	0,93
1000	β_0	2,00	2,00	2,00	2,00	2,00	0,91	0,93	0,95	0,95	0,95
	β_1	0,10	0,10	0,10	0,10	0,10	0,93	0,96	0,95	0,95	0,95
	β_2	-0,20	-0,20	-0,20	-0,20	-0,20	0,93	0,96	0,94	0,95	0,94
	β_3	0,60	0,60	0,60	0,60	0,60	0,94	0,94	0,96	0,95	0,95
	β_4	0,10	0,10	0,10	0,10	0,10	0,93	0,94	0,95	0,95	0,96
	β_5	0,20	0,20	0,20	0,20	0,20	0,92	0,93	0,95	0,94	0,96
	σ_0	-3,59	-3,52	-3,47	-3,46	-3,46	0,81	0,85	0,92	0,94	0,94

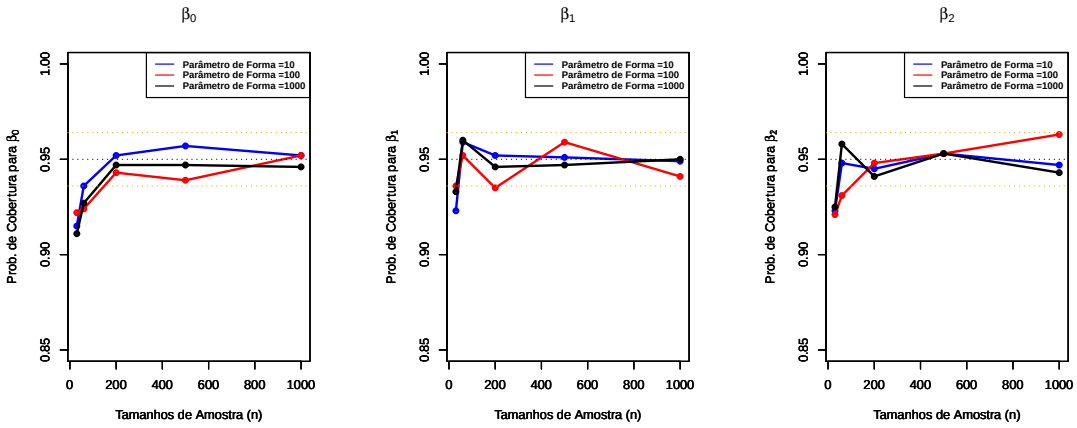


Figura 4.9: Probabilidade de Cobertura dos intervalos de 95% versus o tamanho da amostra do parâmetro β_0, β_1 e β_2 do modelo *GAMLSS*.

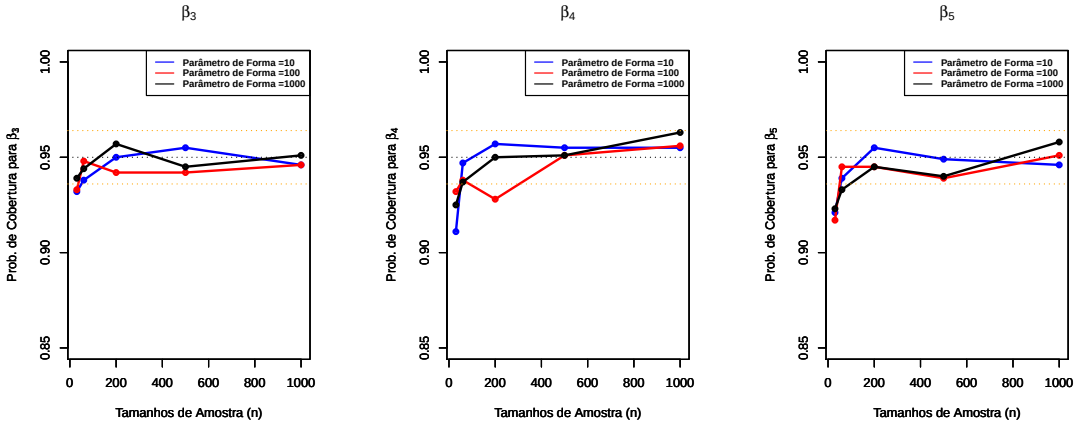


Figura 4.10: Probabilidade de Cobertura dos intervalos de 95% versus o tamanho da amostra do parâmetro β_3, β_4 e β_5 do modelo *GAMLSS*.

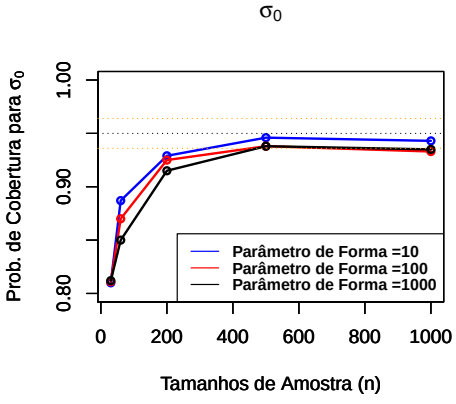


Figura 4.11: Probabilidade de Cobertura dos intervalos de 95% versus o tamanho da amostra do parâmetro σ_0 do modelo *GAMLSS*.

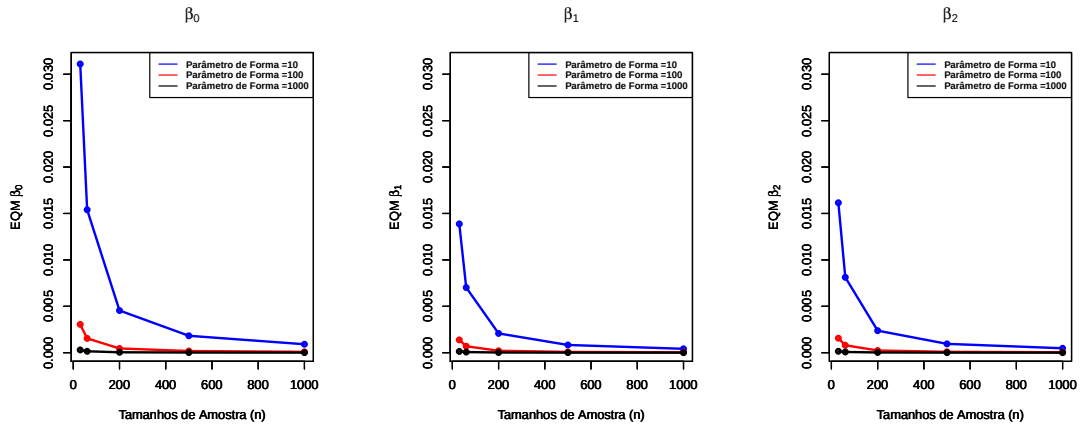


Figura 4.12: EQM versus o tamanho da amostra dos parâmetros β_0, β_1 e β_2 do modelo *GAMLSS*.

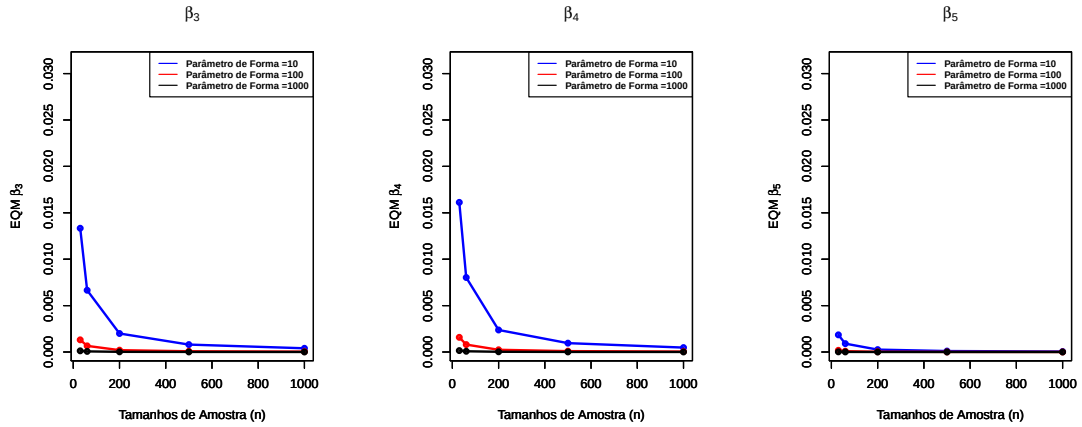


Figura 4.13: EQM versus o tamanho da amostra dos parâmetros β_3, β_4 e β_5 do modelo *GAMLSS*.

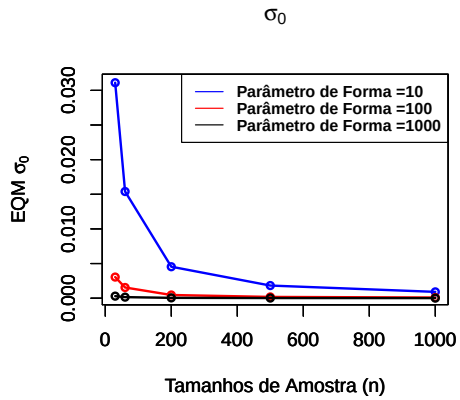


Figura 4.14: EQM versus o tamanho da amostra do parâmetro σ_0 do modelo *GAMLSS*.

4.4 Simulação com ajuste via GAMLSS considerando censura à esquerda

Nesta seção foi utilizado o pacote *GAMLSS* para a estimação dos parâmetros e o pacote *Survival* para definir o tipo de censura (à esquerda). O estudo de simulação considerou diferentes porcentagens de censura nos dados.

Nas Tabelas 4.18 e 4.19 encontram-se as estimativas do modelo linear $log(var(\beta)) = a + blog(n)$ para 0%, 1%, 5%, 15%, 30% e 50% de censura. Observa-se que os valores de b se aproximam de -1 confirmando que o decaimento da variância é da ordem de $n^{-1/2}$.

Sabendo que os verdadeiros valores dos parâmetros para o modelo foram $\beta_0 = 1.50$, $\beta_1 = 0.6$, $\beta_2 = 0.1$, $\beta_3 = 0.8$, $\beta_4 = -0.5$ e $\beta_5 = 0.2$, nota-se que ao aumentar o tamanho da amostra, as estimativas tendem a melhorar, no entanto, ao aumentar a censura, essas estimativas em geral não são próximas do verdadeiro valor do parâmetro. No caso de β_0 as estimativas se aproximam do verdadeiro valor do parâmetro para até 5% de censura. Para β_1 e β_4 apenas para 0% e 1%. Já β_2 β_3 e β_5 apresentaram boas estimativas independente da porcentagem de censura. No caso de σ_0 ao mudar a porcentagem de censura, seu valor inicial também foi modificado e todos se aproximaram do verdadeiro valor do parâmetro.

As probabilidades de cobertura dos intervalos de confiança de 95%, as estimativas dos parâmetros do modelo encontram-se na Tabela 4.20 e nas Figuras 4.15, 4.16, 4.17, 4.18, 4.19 e 4.20. Ao considerar 0% e 1% de censura, todos os parâmetros apresentaram cobertura dentro dos limites nomiais a partir de amostras de tamanho 200. Para as demais porcentagens de censura, a partir de tamanhos de amostra 200, β_1, β_2, β_3, β_4 e β_5 permaneceram com as probabilidades de cobertura dentro dos limites nomiais. A partir de 5%, o aumento de censura direciona a uma diminuição da probabilidade de cobertura dos parâmetros β_0 e σ_0. No caso de σ_0 essa diminuição não passou de 15%, já para β_0 a probabilidade de cobertura chegou a atingir 0% considerando 50% de censura.

Os EQM estão apresentados nas Tabelas 4.21, 4.22, 4.23, 4.24, 4.25 e 4.26. De modo geral, verifica-se que ao aumentar o tamanho da amostra, o EQM tende a diminuir. O aumento de censura, a partir de 15%, também causa um aumento do EQM, principalmente para β_0 e σ_0.

Tabela 4.6: Estimativa de a e b do modelo linear $log(var(\beta)) = a + b\log(n)$ dos parâmetros β_0, β_1,β_2 e β_3 do Modelo Weibull considerando diferentes porcentagens de censuras nos dados.

	$l(V(\beta_0))$	$l(V(\beta_1))$	$l(V(\beta_2))$	$l(V(\beta_3))$	$l(V(\beta_4))$	$l(V(\beta_5))$	$l(V(\sigma_0))$	Censura
a	3,45	2,61	2,79	2,56	2,79	0,64	-0,72	0%
b	-0,97	-0,96	-0,97	-0,96	-0,97	-0,98	-0,97	
a	2,59	2,59	2,78	2,53	2,75	0,61	-0,72	1%
b	-0,96	-0,96	-0,96	-0,95	-0,96	-0,97	-0,97	
a	2,63	2,63	2,81	2,57	2,79	0,65	-0,72	05%
b	-0,96	-0,96	-0,96	-0,95	-0,96	-0,97	-0,97	
a	2,83	2,83	3,04	2,77	2,99	0,87	-0,68	15%
b	-0,98	-0,97	-0,98	-0,96	-0,97	-0,99	-0,98	
a	3,03	3,03	3,23	2,98	3,19	1,07	-0,60	30%
b	-0,98	-0,97	-0,98	-0,97	-0,98	-0,99	-1,00	
a	3,23	3,23	3,40	3,17	3,41	1,24	-0,72	50%
b	-0,95	-0,95	-0,95	-0,94	-0,95	-0,96	-0,98	

Tabela 4.7: Probabilidade de cobertura dos intervalos de confiança de 95% e estimativa dos parâmetros do modelo Weibull para diferentes porcentagens de censura.

		Tamanho da amostra									
		30	60	200	500	1000	30	60	200	500	1000
Censura	$\beta's$	Estimativa					PC				
0%	β_0	1,41	1,48	1,48	1,49	1,50	0.883	0.901	0.951	0.947	0.962
	β_1	0,47	0,53	0,55	0,55	0,60	0.871	0.918	0.950	0.948	0.963
	β_2	0,07	0,08	0,08	0,09	0,10	0.872	0.911	0.950	0.956	0.953
	β_3	0,71	0,76	0,74	0,78	0,80	0.870	0.917	0.948	0.935	0.955
	β_4	-0,43	-0,46	-0,48	-0,49	-0,50	0.887	0.923	0.953	0.945	0.948
	β_5	0,19	0,20	0,20	0,20	0,20	0.887	0.909	0.946	0.948	0.936
	σ	-0,54	-0,57	-0,59	-0,59	-0,69	0.781	0.872	0.919	0.946	0.934
1%	β_0	1,37	1,45	1,46	1,47	1,47	0.868	0.922	0.936	0.954	0.946
	β_1	0,56	0,55	0,54	0,53	0,52	0.877	0.908	0.939	0.957	0.952
	β_2	0,13	0,10	0,06	0,07	0,08	0.867	0.917	0.947	0.935	0.948
	β_3	0,72	0,75	0,76	0,74	0,75	0.865	0.916	0.939	0.943	0.954
	β_4	-0,57	-0,55	-0,54	-0,54	-0,54	0.883	0.913	0.932	0.954	0.946
	β_5	0,19	0,19	0,20	0,20	0,20	0.884	0.913	0.933	0.949	0.953
	σ	-0,55	-0,63	-0,68	-0,69	-0,70	0.755	0.857	0.924	0.924	0.935
5%	β_0	1,27	1,38	1,39	1,40	1,40	0.861	0.918	0.938	0.934	0.921
	β_1	0,57	0,54	0,47	0,41	0,41	0.873	0.920	0.942	0.957	0.953
	β_2	0,13	0,10	0,06	0,07	0,08	0.863	0.914	0.945	0.938	0.945
	β_3	0,73	0,75	0,76	0,74	0,75	0.864	0.920	0.935	0.939	0.960
	β_4	-0,51	-0,49	-0,44	-0,43	-0,43	0.875	0.905	0.927	0.952	0.950
	β_5	0,20	0,19	0,20	0,20	0,20	0.878	0.903	0.935	0.947	0.955
	σ	-0,57	-0,65	-0,70	-0,72	-0,72	0.747	0.850	0.917	0.925	0.940
15%	β_0	1,04	1,15	1,19	1,19	1,18	0.845	0.876	0.898	0.834	0.696
	β_1	0,49	0,34	0,37	0,39	0,36	0.839	0.910	0.945	0.936	0.947
	β_2	0,14	0,12	0,06	0,07	0,08	0.848	0.917	0.944	0.957	0.947
	β_3	0,73	0,76	0,77	0,75	0,75	0.859	0.910	0.957	0.956	0.943
	β_4	-0,50	-0,49	-0,42	-0,41	-0,41	0.848	0.908	0.940	0.934	0.946
	β_5	0,20	0,19	0,20	0,20	0,20	0.859	0.920	0.939	0.946	0.954
	σ	-0,62	-0,71	-0,76	-0,78	-0,78	0.774	0.895	0.901	0.912	0.925
30%	β_0	0,72	0,86	0,89	0,90	0,99	0.796	0.856	0.748	0.415	0.160
	β_1	0,51	0,47	0,43	0,41	0,39	0.848	0.904	0.937	0.961	0.948
	β_2	0,15	0,13	0,07	0,07	0,09	0.841	0.903	0.943	0.955	0.943
	β_3	0,73	0,76	0,76	0,75	0,75	0.844	0.902	0.944	0.948	0.952
	β_4	-0,49	-0,37	-0,37	-0,37	-0,35	0.851	0.908	0.931	0.950	0.952
	β_5	0,20	0,20	0,20	0,20	0,19	0.844	0.888	0.935	0.938	0.958
	σ	-0,71	-0,81	-0,87	-0,88	-0,89	0.670	0.781	0.868	0.871	0.901
50%	β_0	-0,35	-0,21	0,37	0,36	0,53	0.691	0.672	0.345	0.039	0.000
	β_1	0,52	0,39	0,34	0,33	0,30	0.798	0.895	0.927	0.956	0.937
	β_2	0,25	0,12	0,07	0,07	0,07	0.814	0.886	0.950	0.950	0.958
	β_3	0,72	0,76	0,78	0,75	0,75	0.816	0.907	0.932	0.932	0.948
	β_4	-0,47	-0,46	-0,43	-0,41	-0,38	0.810	0.889	0.925	0.940	0.948
	β_5	0,16	0,19	0,20	0,20	0,20	0.802	0.891	0.945	0.953	0.944
	σ	-0,85	-0,97	-1,04	-1,06	-1,06	0.564	0.708	0.810	0.818	0.830

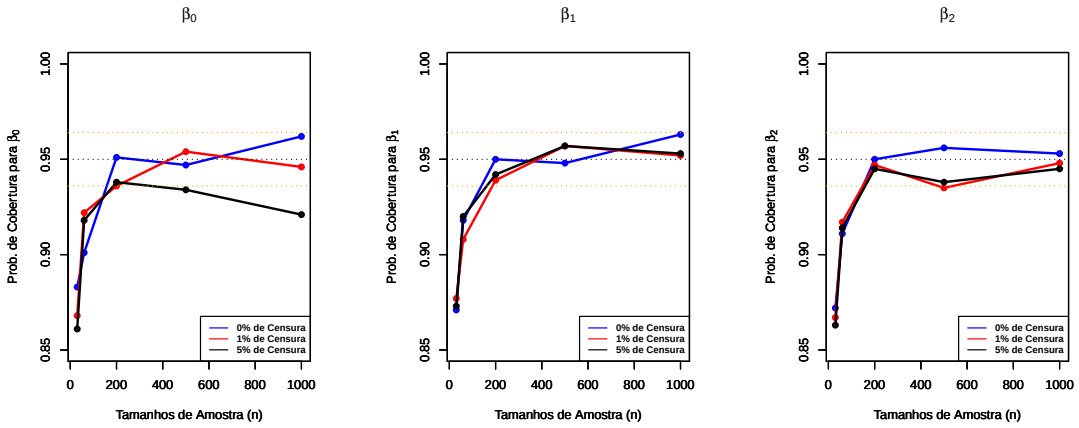


Figura 4.15: Probabilidade de Cobertura dos intervalos de 95% versus o tamanho da amostra dos parâmetros β_0, β_1 e β_2 para 0%, 1% e 5% de censura.

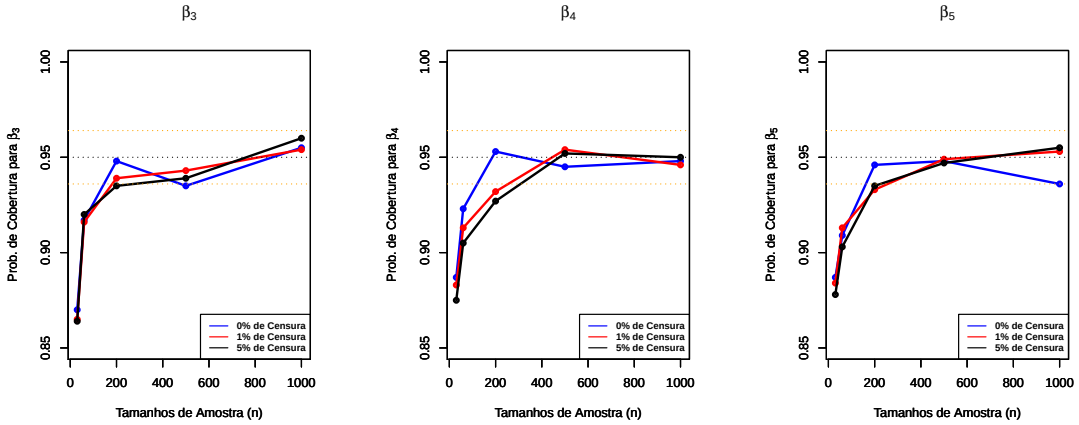


Figura 4.16: Probabilidade de Cobertura dos intervalos de 95% versus o tamanho da amostra dos parâmetros β_3, β_4 e β_5 para 0%, 1% e 5% de censura.

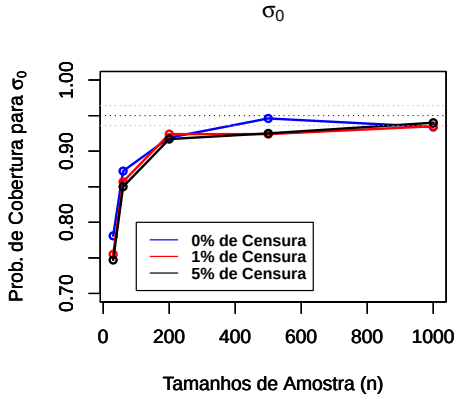


Figura 4.17: Probabilidade de Cobertura dos intervalos de 95% versus o tamanho da amostra do parâmetro σ_0 para 0%, 1% e 5% de censura.

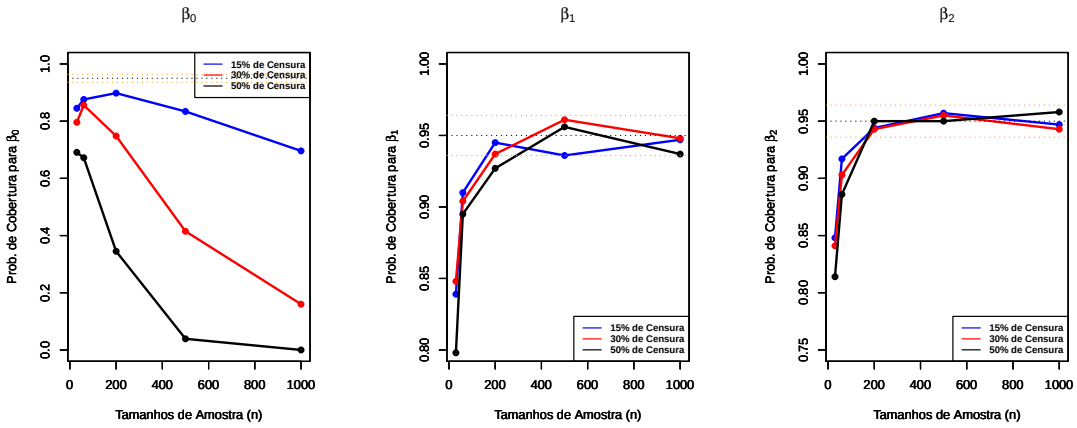


Figura 4.18: Probabilidade de Cobertura dos intervalos de 95% versus o tamanho da amostra dos parâmetros β_0, β_1 e β_2 para 15%, 30% e 50% de censura.

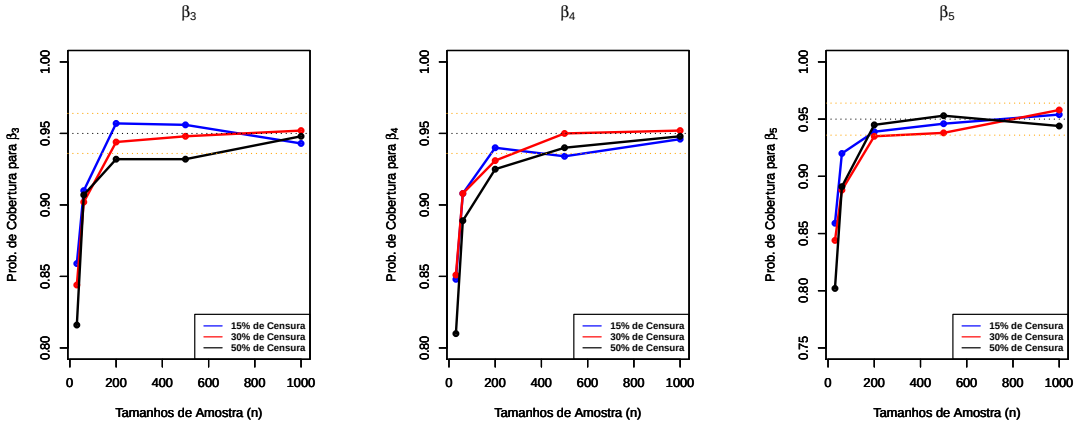


Figura 4.19: Probabilidade de Cobertura dos intervalos de 95% versus o tamanho da amostra dos parâmetros β_3, β_4 e β_5 para 15%, 30% e 50% de censura.

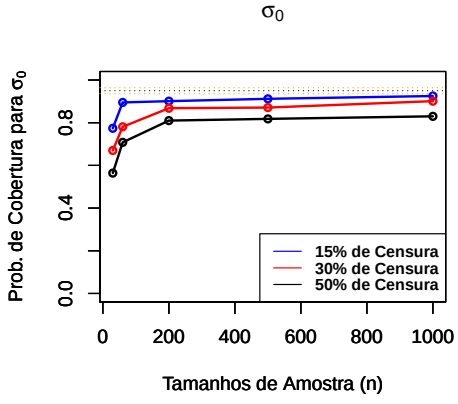


Figura 4.20: Probabilidade de Cobertura dos intervalos de 95% versus o tamanho da amostra do parâmetro σ_0 para 15%, 30% e 50% de censura.

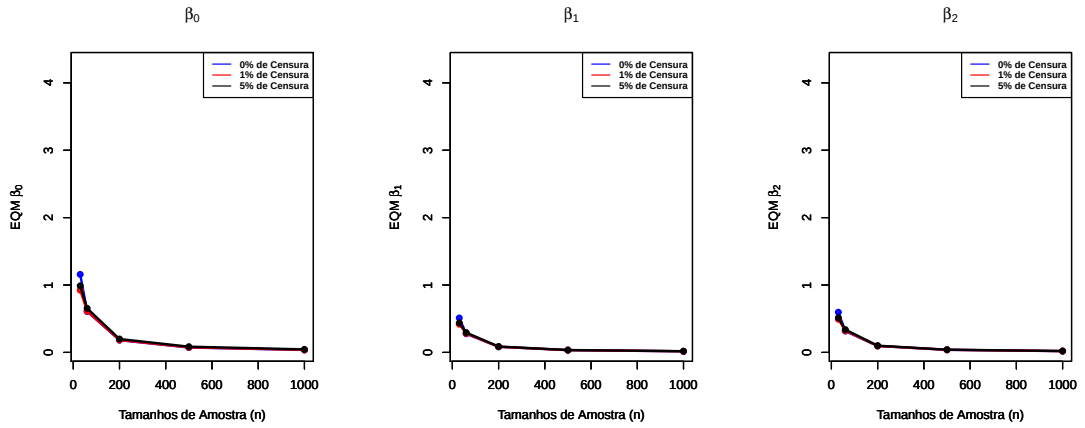


Figura 4.21: EQM versus o tamanho da amostra dos parâmetros β_0, β_1 e β_2 para 0%, 1% e 5% de censura.

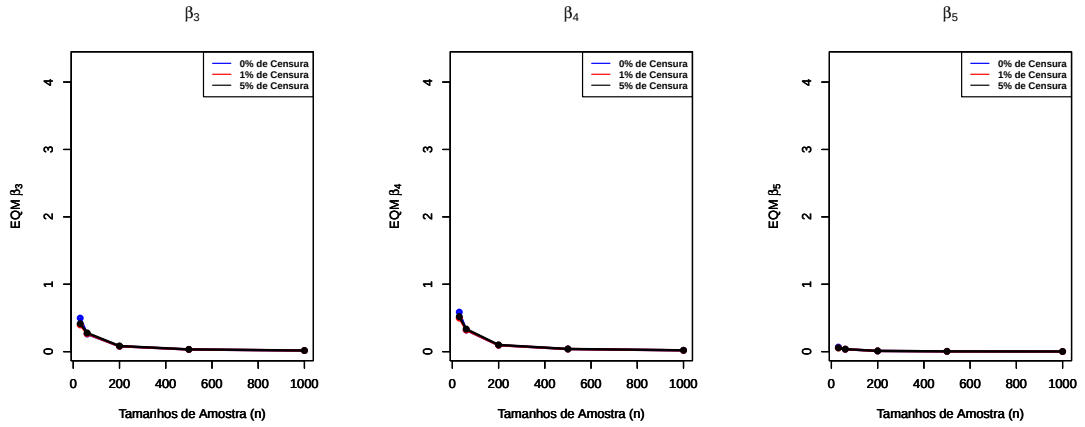


Figura 4.22: EQM versus o tamanho da amostra dos parâmetros β_3, β_4 e β_5 para 0%, 1% e 5% de censura.

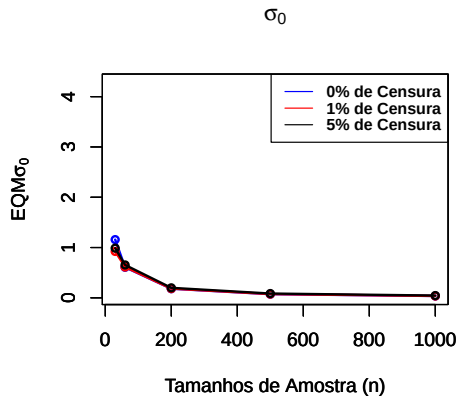


Figura 4.23: EQM versus o tamanho da amostra do parâmetro σ_0 para 0%, 1% e 5% de censura.

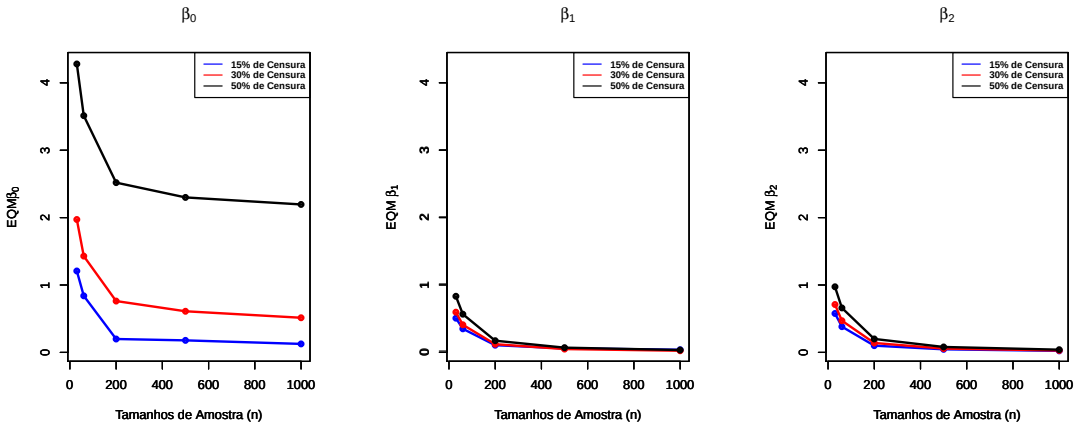


Figura 4.24: EQM versus o tamanho da amostra dos parâmetros β_0, β_1 e β_2 para 15%, 30% e 50% de censura.

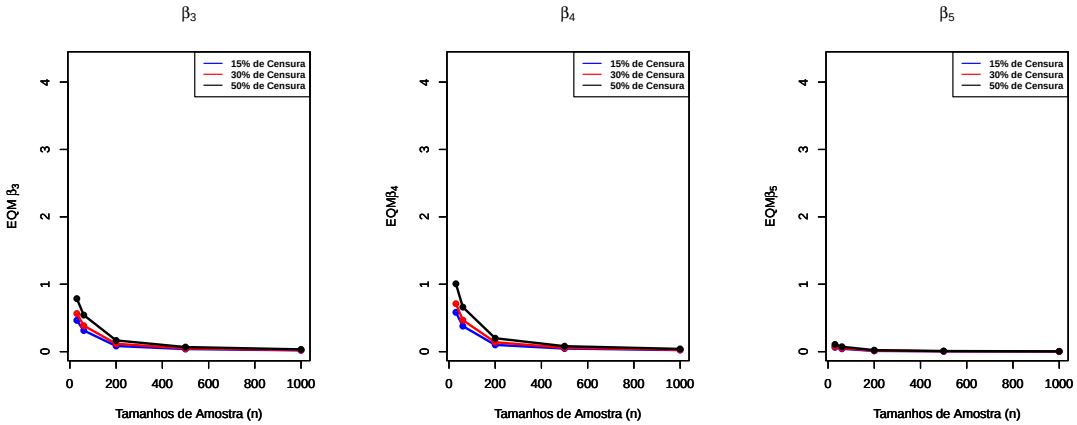


Figura 4.25: EQM versus o tamanho da amostra dos parâmetros β_3, β_4 e β_5 para 15%, 30% e 50% de censura.

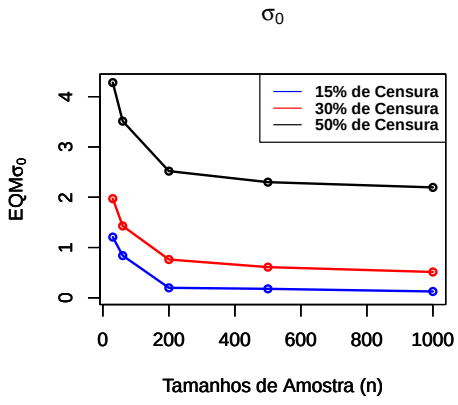


Figura 4.26: EQM versus o tamanho da amostra do parâmetro σ_0 para 0%, 1% e 5% de censura.

4.5 Considerações Finais

Nos estudos de simulação apresentados neste Capítulo, observa-se que ao considerar variâncias baixas e moderadas no modelo linear, os EQM dos estimadores apresentaram valores baixos. No caso dos modelos lineares generalizados e do *GAMLSS*, os EQMS aumentam ao diminuir o valor do parâmetro de forma.

Ao considerar censura nos dados, o aumento desta direciona a uma diminuição da probabilidade de cobertura do parâmentro β_0 e σ_0, sendo menor que 15% para o caso de σ_0. Por fim, observa-se também que para amostras pequenas, a teoria assintótica pode não ser recomendada, pois para certos parâmetros, algumas estimativas podem não estar adequadas.

Com o objetivo de analisar a sensibilidade dos modelos frente às diferentes metodologias descritas, no próximo capítulo elas serão aplicadas a dados reais da Cidade de São Carlos.

Capítulo

5

Estudo de Caso: dados de lotes urbanos de São Carlos 2005

Os dados em estudo são referentes aos lotes urbanos do município de São Carlos do ano de 2005. O espaço amostral possui 311 lotes. Em cada lote da amostra estão especificados o valor do imóvel, as variáveis de localização que serão descritas a seguir, a área referente ao lote e se a venda do mesmo foi ou não realizada. Foram considerados apenas os lotes com área igual ou inferior a 800m^2 totalizando 290 lotes. Desses 290 lotes, 7 não haviam sido vendidos.

Neste contexto as metodologias apresentadas nos Capítulos 2 e 3 serão aplicadas a esse conjunto de dados. Em um primeiro momento, serão abordados os modelos lineares, os modelos lineares generalizados e o *GAMLSS* nos quais serão utilizados apenas os 283 lotes comercializados efetivamente em 2005. Após a análise desses modelos, os lotes não vendidos serão incluidos no estudo e será então aplicada a análise de sobrevivência juntamente com o *GAMLSS* na estimação dos parâmetros. O objetivo é encontrar modelos plausíveis para explicar a formação do valor dos lotes em questão.

A complexidade e abstrações envolvidas na avaliação do território, por meio de métodos científicos juntamente com a pertinência da variável localização nos modelos matemáticos, torna a conceituação da mesma indispensável. Para isso, é necessário compreender os processos formadores de valor do solo e reconhecer a estruturação urbana de uma cidade, que segundo Ferreira (2007) é produto dialético de determinantes históricas em uma estrutura que constantemente se transforma.

As barreiras urbanas são relevantes porque desvalorizam certas localizações em relação a outras. Essas barreiras são de ordem natural, como a conformação do relevo e a hidrografia e sob influência destes, frequentemente são traçadas as ferrovias e as rodovias. (Ferreira 2007)

Ferreira (2007) seleciona as barreiras urbanas, que têm efeitos gerais sobre conjuntos de localizações, como ponto de partida para a modelagem da componente localização, e a partir disso, cria unidades de localização básicas para a formulação de um modelo estatístico que exprima a valorização territorial urbana. Neste contexto, São Carlos apresenta quatro grandes barreiras intra-urbanas: a delimitação da Planície Central, a ferrovia, a rodovia Washington Luiz e a encosta sul.

Como o presente estudo explora os dados utilizados por Ferraudo (2008), que por sua vez adota a proposta de Ferreira (2007), as variáveis utilizadas serão as mesmas dos dois trabalhos. Ressalta-se que Ferreira (2007) considera estas variáveis básicas para a formulação de um modelo matemático que exprima de forma potencial o fenômeno de valorização territorial urbana de São Carlos no contexto de 2005. Com relação as discussões das hipóteses das variáveis é possível encontrar maiores informações em Ferreira (2007), a seguir elas serão apresentadas resumidamente.

As 8 variáveis são dicotômicas gerando zonas homogêneas, ou seja, se o imóvel em avaliação pertence ou não a determinada região atribui-se o valor 0 ou 1, sendo que 1 normalmente representa um fator de valorização enquanto que 0 um fator de desvalorização, dependendo das características específicas de cada uma.

Planície Central: corresponde a um fator de valorização. Pontos internos a planície apresentam um potencial de acessibilidade privilegiado com relação aos pontos externos. Espera-se que a variabilidade de valores encontrados nesta área seja devida à proximidade com os principais eixos viários, Av. São Carlos no eixo norte-sul e no sentido leste-oeste a Av. Carlos Botelho e a Rua 15 de Novembro. Atribui-se o valor 1 às áreas que estão contidas no interior desta área e 0 às áreas externas.

Ferrovia: a ferrovia juntamente com o córrego Gregório formam uma barreira dupla fazendo com que a cidade seja dividida em dois lados, o lado de cá (antes da ferrovia) onde fica o centro tradicional e o lado de lá (além da ferrovia) ocupado na sua maioria pelas camadas mais pobres com um expressivo sub-centro, a Vila Prado. Também expõe os moradores da região a diversos inconvenientes como acidentes com risco de morte, altos ruídos, pragas urbanas e a própria dificuldade de acessibilidade intra-urbana tornando esta, uma barreira de desvalorização. As áreas de acessibilidade não afetadas pela ferrovia são mais valorizadas, sendo assim, atribui-se o valor 1 às regiões que não são afetadas pela barreira ferroviária e 0 às afetadas.

Rodovia: a pista dupla, separada por valas ou muros de concreto e a alta velocidade permitida oferece risco para transposição constituindo um fator de desvalorização. As exceções ocorrem nas localidades cujo acesso é feito sobre o nível do pavimento da SP-310, na continuação da Av. São Carlos e Av. Getúlio Vargas. As áreas de acessibilidade não afetadas pela rodovia Washington Luiz (Sp-310) são mais valorizadas, sendo assim, atribui-se o valor 1 às regiões onde a acessibilidade ao centro não são afetadas pela barreira rodoviária e 0 às regiões fetadas.

Encosta Sul: fator de desvalorização pois possui áreas com os piores indicadores de educação e renda, com baixa oferta de comércios e serviços; acentuado distanciamento dos

principais polos de empregos e estabelecimentos de consumo. Atribui-se 1 às regiões cuja acessibilidade não é afetada pela barreira e 0 às afetadas.

Fechado: os parcelamentos fechados localizam-se com maior frequência no interior do perímetro urbano, porém localizam-se também nas áreas de expansão urbana. É referido popularmente como "condomínio"pelo aspecto murado, no entanto, define-se condomínio por um conjunto de características além do muro. Uma das hipóteses é que seja um fator de valorização, mas isoladamente tal característica não é suficiente para diferenciar os graus de fechamento destes parcelamentos. Atribui-se 1 às regiões que são fechadas por muros e 0 às que não são.

Condomínio: possui áreas privativas e comuns, estabelecendo-se frações ideais da participação. A manutenção das infraestruturas de saneamento, pavimentação, drenagem, iluminação, segurança e outros, normalmente são vinculadas à taxa de condomínio. A hipótese correspondente a esta variável é considerar que a instituição de condomínio é um fator de valorização. Atribui-se o valor 1 às áreas que estão contidas no interior de condomínios fechados registrados e o valor 0 às áreas externas.

Uso Residencial: com base na amostragem preliminar e nas análises estatísticas sobre os dados censitários Ferreira (2007) observou que existe uma forte correlação entre ocupação residencial de maior renda e os loteamentos com restrições a usos não residenciais. As restrições apresentam uma expressiva variedade, podendo ser total, extensiva a toda área do loteamento, vias específicas para usos comerciais e de serviços, excluindo por completo alguns usos e tipologias. O controle através de restrições urbanísticas tende a aumentar nos parcelamentos fechados, estritamente residenciais ou nos condomínios com edifícios. A hipótese ligada à inclusão desta variável no modelo consiste em considerar que a característica do loteamento em ser de uso estritamente residencial ou com alguma restrição neste sentido contribui para valorizar os lotes. Esta variável é de natureza jurídica. Atribui-se o valor 1 às áreas pertencentes a estes loteamentos e o valor 0 às áreas não pertencentes.

Núcleo Sede: localizações relativamente remotas com relação à sede do município, separados por várias bacias hidrográficas. Padrão comum do lote tende a ser maior, normalmente baixa oferta de redes gerais de água e esgoto, pavimentação, energia elétrica e iluminação pública. Atribui-se o valor 1 quando a localidade encontra-se contígua à aglomeração da sede e recebe o valor 0 quando a localidade (ou o parcelamento) encontra-se isolada desta aglomeração.

Na Tabela 5.1 encontra-se o resumo das variáveis de localização.

Tabela 5.1: Variáveis dicotômicas indicadoras da localização.

Variáveis	Descrição
NUC.PRINC:	1 = Lote localiza-se contíguo à aglomeração da Sede Municipal; 0 = Lote localiza-se em parcelamentos rurais.
PLN.CENTRAL:	1 = Lote localiza-se no interior da Planície Central; 0 = Fora da Planície Central.
FERROVIA:	1 = A acessibilidade ao centro não é prejudicada pela ferrovia; 0 = o inverso.
RODOVIA:	1 = A acessibilidade ao centro não é prejudicada pela rodovia SP-310 (Rod. Washington Luís); 0 = o inverso.
ENCOSTA:	1 = A acessibilidade ao centro não é prejudicada pela encosta sul; 0 = o inverso.
CONDO:	1 = Lote localiza-se em condomínio urbanístico; 0 = Lote não se localiza em condomínio urbanístico.
FECHADO:	1 = Lote localiza-se em bairro fechado por muros; 0 = Lote localiza-se em bairro aberto.
ESTRIT.RESID:	1 = O parcelamento a que pertence o lote é estritamente residencial; 0 = O parcelamento tem uso misto.

5.1 Modelagem de dados Completos

Nesta primeira etapa, para os ajustes dos modelos com dados completos, o banco de dados foi separado em duas partes. Utilizou-se 70% dos dados para o ajuste do modelo e 30% para a validação.

5.1.1 Modelo Linear

Esta abordagem nesse conjunto de dados foi proposta inicialmente por Ferraudo (2008), assim, o intuíto é apenas reproduzí-la para compará-la aos outros modelos. O modelo inicial é constituído da variável área, das variáveis descritas na Tabela 5.1 juntamente com a iteração entre cada uma delas com a variável área, assim, sua forma funcional é dada por:

$$\begin{aligned} V_i =& \beta_0 + \beta_1 AREA_i + \beta_2 NUCPRINC_i + \beta_3 PLNCENTRAL_i + \beta_4 FERROVIA_i + \\ & \beta_5 RODOVIAWL_i + \beta_6 ENCOSTA_i + \beta_7 CONDO_i + \beta_8 FECHADO_i + \\ & \beta_9 ESTRITRESID_i + \beta_{10}(NUCPRINC_i * AREA_i) + \\ & \beta_{11}(PLNCENTRAL_i * AREA_i) + \beta_{12}(FERROVIA_i * AREA_i) + \\ & \beta_{13}(RODOVIAWL_i * AREA_i) + \beta_{14}(ENCOSTA_i * AREA_i) + \\ & \beta_{15}(CONDO_i * AREA_i) + \beta_{16}(FECHADO_i * AREA_i) + \\ & \beta_{17}(ESTRITRESID_i * AREA_i) + \epsilon. \end{aligned} \tag{5.1}$$

O primeiro ajuste apresentou um AIC=4377.3 e algumas variáveis não significativas. Foi aplicado os testes de Shapiro-Wilk (p-valor=0,000) e Kolmogorov-Smirnov (p-valor=0,000), re-

jeitando a normalidade dos resíduos. Dessa maneira, para satisfazer os pressupostos exigidos na regressão linear (normalidade e homocedasticidade dos resíduos), utilizou-se a transformação BOX-COX na variável resposta, nos 70% dos dados utilizados para o ajuste e optou-se pela transformação raiz quadrada.

Foi realizado o ajuste com a transformação raiz quadrada, o qual reduziu o AIC (1.966,2). Fez-se a análise de diagnóstico dos resíduos deste ajuste e verificaram-se alguns pontos outliers, os resultados foram omitidos.

Essas observações outliers foram retiradas e o modelo foi novamente ajustado. A retirada das observações melhorou significativamente o ajuste do modelo, reduzindo o erro padrão e o p-valor relacionado ao parâmetro de cada variável.

Para a seleção do modelo mais adequado utilizou-se o método STEPWISE o qual escolheu o seguinte modelo:

$$\begin{aligned} V_i^{1/2} =& \beta_0 + \beta_1 AREA_i + \beta_2 NUCPRINC_i + \beta_3 PLNCENTRAL_i + \beta_4 FERROVIA_i + \\ & \beta_5 RODOVIAWL_i + \beta_6 CONDO_i + \beta_7 FECHADO_i + \beta_8 ESTRITRESID_i + \\ & \beta_9 (NUCPRINCi * AREA_i) + \beta_{10} (PLNCENTRALi * AREA_i) + \\ & \beta_{11} (CONDO_i * AREA_i) + \beta_{12} (FECHADO_i * AREA_i) + \\ & \beta_{13} (ESTRITRESID_i * AREA_i) + \epsilon. \end{aligned} \tag{5.2}$$

No entanto a variável ÁREA e PLNCENTRAL não foram significativas para o modelo. Nesta etapa, também deve ser levada em consideração a avaliação do pesquisador, o qual deve ser responsável por combinar a significância estatística com o objetivo prático. Assim, retirou-se a variável ÁREA, enquanto que a variável PLNCENTRAL permaneceu no modelo devido a sua questão prática .

Ajustou-se o modelo final que diminuiu ainda mais o AIC (1901.7). A análise de resíduos foi realizada e encontra-se nas Figuras 5.1 e 5.2, as quais apresentaram $h_{ii} < 1$, distância de Cook $< 0,2$ e DFFITS < 1 indicando que não existe observação influente. O gráfico dos resíduos versus os valores ajustados configurou-se de forma aleatória, indicando assim, que os resíduos são não correlacionados, ou seja, as hipóteses de independência e variância constante para os resíduos foram aceitas. Os testes de Shapiro (0.182) e Kolmogorov (0.96) não rejeita a normalidade ao nível de 10%. As estimativas encontram-se na Tabela 5.2.

Tabela 5.2: Estimativa dos parâmetros, limite superior e inferior do intervalo de confiança de 95%, erro padrão e p-valor.

Variável	Estimativa	IC(95%)	Erro Padrão	p-valor
Intercepto	103.18	[70.44 ; 135.91]	16.702	0.0000
NUCPRINC	-83.99	[-115.44 ; -52.56]	16.040	0.0000
PLNCENTRAL	23.59	[-16.70 ; 63.88]	20.555	0.2526
FERROVIA	17.56	[3.00 ; 32.12]	7.429	0.0191
RODOVIAWL	34.87	[16.22 ; 53.52]	9.516	0.0000
CONDO	127.29	[62.84 ; 191.74]	32.882	0.0000
FECHADO	-153.33	[-203.29 ;-103.37]	25.490	0.0000
ESTRITRESID	41.15	[4.18 ; 78.14]	18.867	0.0304
NUCPRINC*AREA	0.34	[0.26 ; 0.42]	0.038	0.0000
PLNCENTRAL*AREA	0.045	[-0.06 ; 0.15]	0.053	0.4091
CONDO*AREA	-0.34	[-0.50 ; -0.18]	0.081	0.0000
FECHADO*AREA	0.53	[0.38 ; 0.68]	0.077	0.0000
ESTRITRESID*AREA	-0.13	[-0.23 ; -0.03]	0.053	0.0150

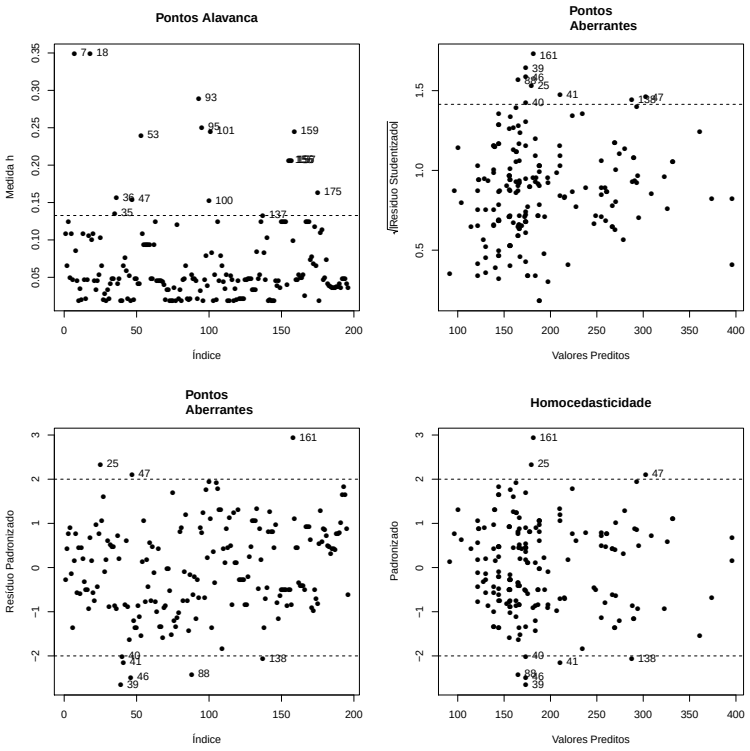


Figura 5.1: Gráficos de Pontos de Alavanca, Pontos Aberrantes e Homocedasticidade do modelo normal com transformação raiz quadrada para a variável resposta.

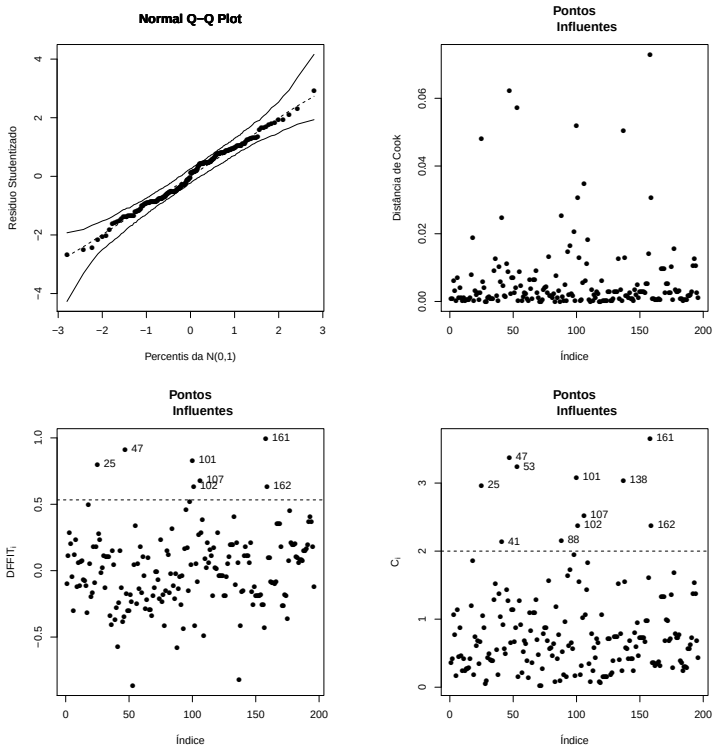


Figura 5.2: Gráficos de diagnóstico para o modelo normal com transformação raiz quadrada para a variável resposta.

A próxima etapa é a validação do modelo proposto, ou seja, aplica-se o modelo encontrado aos dados separados para a validação. Como foi aplicada a transformação raiz quadrada na variável resposta é necessário aplicar a transformação inversa para obter os valores preditos na escala da variável original. Para isso, basta elevar o preditor linear ao quadrado. A diferença relativa média para o modelo é de 32%, ou seja, a taxa de aceitação total é de 68%. A Figura 5.3 mostra os valores observados versus os valores preditos elevados ao quadrado devido à transformação aplicada aos dados. Como era de se esperar os pontos estão em torno de uma reta indicando a adequabilidade do modelo proposto aos dados.

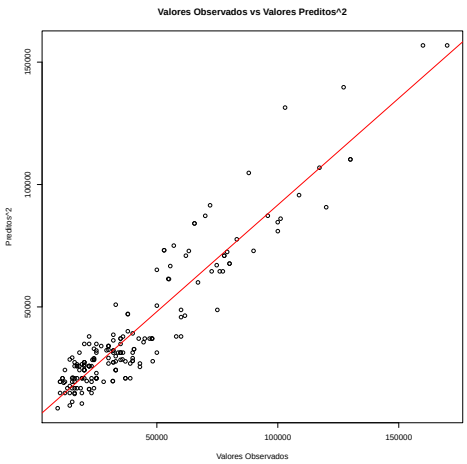


Figura 5.3: Valores observados versus valores preditos^2.

Antes de iniciar o ajuste pelo MLG, vale ressaltar que o ajuste obtido por Ferraudo(2008) não considerou a variável CONDO*AREA e teve uma taxa de aceitação de 75%, a qual pode ser explicada devido a separação da amostra original, que foi feita aleatoriamente além da utilização de outra versão do programa R e outras configurações dos computadores usados no processo de modelagem.

5.1.2 Modelos Lineares Generalizados

Considerando que ao se transformar os dados muda-se a escala dos mesmos, uma solução é buscar alternativas para se manter ao máximo o conjunto inicial. Dessa maneira, será abordada a metodologia dos MLGs. Sabe-se que a variável dependente é contínua, o intervalo de variação é positivo e há uma grande concentração de pontos a esquerda da distribuição (assimetria, ver histograma **??**). Com essas caracteristicas pode-se antever uma inadequação do modelo normal facilmente verificada com um teste de normalidade (Kolmogorov). Neste caso, potenciais candidatos para modelar os dados em causa são os modelos gama ou normal inversa.

Foram testadas algumas distribuições para o modelo inicial 5.1 com diferentes funções de ligação. Os resultados encontram-se na Tabela 5.3. Escolheu-se então a distribuição Gama com função logarítmica por apresentar o menor AIC.

Tabela 5.3: AIC de alguns modelos e suas respectivas funções de ligação.

	Função de Ligação		
Distribuição	Identidade	Logarítimica	Recíproca
Normal	4377.3	———	———
Gama	4311.2	4301.4	4304
Normal Inversa	4346.2	4341.4	4353.5

Após a escolha da função de ligação, ajustaram-se os dados pelo modelo gama com a função logarítmica dado inicialmente por 5.3. Utilizou-se a metodologia de STEPWISE para selecionar as covariáveis do modelo, em seguida verificou-se a significância estatística das variáveis e a análise de diagnósticos, finalizando com a validação.

$$\begin{aligned}\mu_i = exp(&\beta_0 + \beta_1 AREA_i + \beta_2 NUCPRINC_i + \beta_3 PLNCENTRAL_i + \beta_4 FERROVIA_i + \\ &\beta_5 RODOVIAWL_i + \beta_6 ENCOSTA_i + \beta_7 CONDO_i + \beta_8 FECHADO_i + \\ &\beta_9 ESTRITRESID_i + \beta_{10}(NUCPRINCi * AREA_i) + \\ &\beta_{11}(PLNCENTRALi * AREA_i) + \beta_{12}(FERROVIA_i * AREA_i) + \\ &\beta_{13}(RODOVIAWL_i * AREA_i) + \beta_{14}(ENCOSTAi * AREA_i) + \\ &\beta_{15}(CONDO_i * AREA_i) + \beta_{16}(FECHADO_i * AREA_i) + \\ &\beta_{17}(ESTRITRESID_i * AREA_i)). \end{aligned} \quad (5.3)$$

Selecionou-se o modelo final 5.4. A análise de diagnóstico do modelo mostrou algumas observações influentes que foram retiradas e o modelo foi novamente ajustado. A retirada das observações melhorou significativamente o ajuste do modelo, reduzindo o erro padrão e o p-valor dos testes de significância das variáveis.

$$\begin{aligned}\mu_i = exp(\beta_0 + +\beta_1 NUCPRINC_i + \beta_2 PLNCENTRAL_i + \beta_3 FERROVIA_i + \\ \beta_4 RODOVIAWL_i + \beta_5 CONDO_i + \beta_6 FECHADO_i + \beta_7 ESTRITRESID_i + \\ \beta_8 (NUCPRINCi * AREA_i) + \beta_9 (PLNCENTRALi * AREA_i) + \\ \beta_{10} (CONDO_i * AREA_i) + \beta_{11} (FECHADO_i * AREA_i) + \\ \beta_{12} (ESTRITRESID_i * AREA_i)). \end{aligned} \quad (5.4)$$

Na Figura 5.4 encontra-se a análise de resíduos, a qual apresentou $h_{ii} < 1$, e embora alguns pontos indicam distância de $COOK > 0.2$ optou-se por manter tais pontos pois a retirada dessas observações não causou grandes mudanças na variação percentual das estimativas dos parâmetros. A Figura 5.5(a) apresenta o gráfico de envelope indicando que o modelo é razoável para ajustar os dados e a Figura 5.5(b) exibe os valores observados versus os valores preditos. Como era de se esperar os pontos estão em torno de uma reta indicando a adequabilidade do modelo proposto. As estimativas dos parâmetros encontram-se na Tabela 5.4.

Tabela 5.4: Estimativa dos parâmetros,limite inferior, limite superior erro padrão e p-valor utilizando o modelo gama com ligação logarítmica.

Variável	Estimativa	IC(95%)	Erro Padrão	p-valor
Intercepto	9.28	[8.96 ; 9.61]	0.167	0.0000
NUCPRINC	-0.85	[-1.16 ; -0.54]	0.159	0.0000
PLNCENTRAL	0.12	[-0.31 ; 0.55]	0.219	0.5938
FERROVIA	0.26	[0.12 ; 0.41]	0.074	0.0005
RODOVIAWL	0.56	[0.37 ; 0.75]	0.095	0.0000
CONDO	1.03	[0.43 ; 1.63]	0.308	0.0009
FECHADO	-1.16	[-1.66 ; -0.66]	0.255	0.0000
ESTRITRESID	0.45	[0.08 ; 0.82]	0.190	0.0187
NUCPRINC*AREA	0.003	[0.00 ; 0.00]	0.000	0.0000
PLNCENTRAL*AREA	0.0002	[0.00 ; 0.00]	0.000	0.6755
CONDO*AREA	-0.002	[0.00 ; 0.00]	0.001	0.0032
FECHADO*AREA	0.003	[0.00 ; 0.01]	0.001	0.0000
ESTRITRESID*AREA	-0.001	[0.00 ; 0.00]	0.000	0.0115

Na validação do modelo, a diferença relativa média foi de 28%, ou seja, a taxa de aceitação total é de 72%, indicando uma melhora considerável no ajuste ao se comparar com a regressão linear normal.

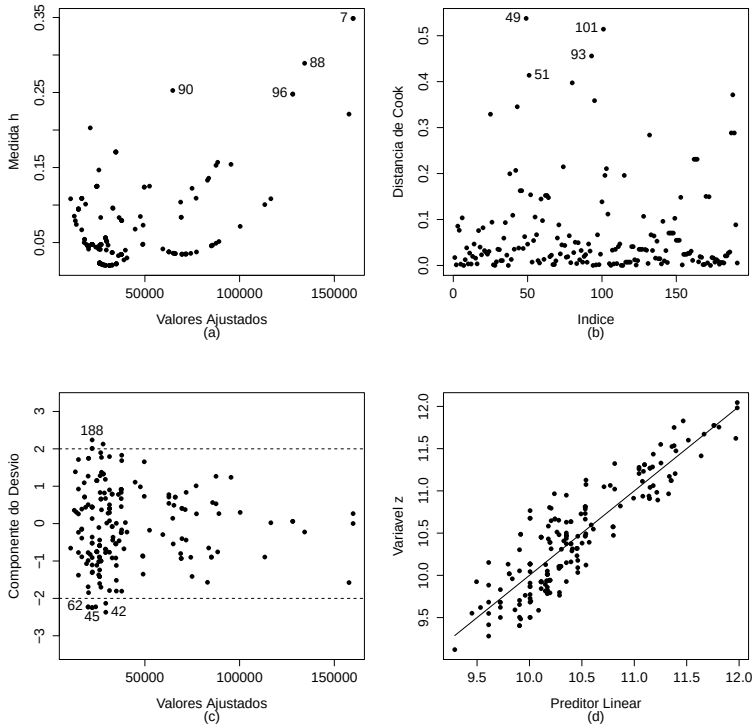
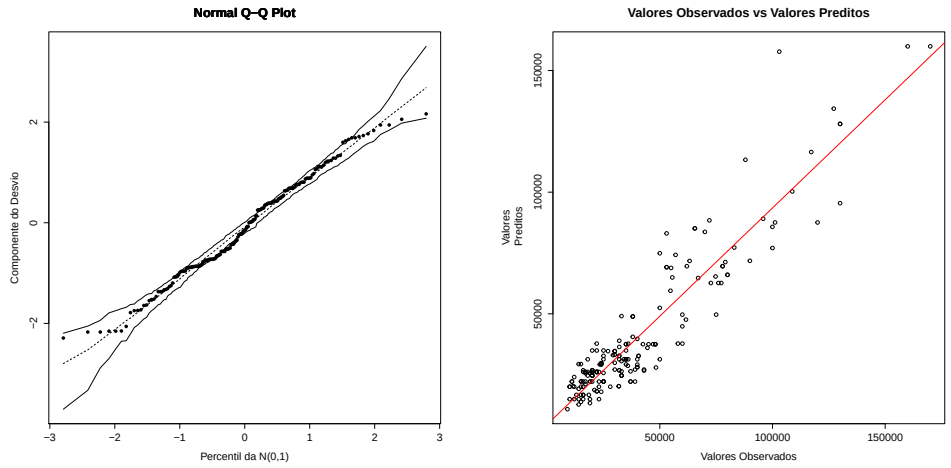


Figura 5.4: Gráficos de diagnóstico para o modelo gama com função de ligação log.



(a) . (b) .

Figura 5.5: Gráfico de envelope para a componente do desvio e Valores observados versus valores preditos para o MLG.(a) e (b)

5.1.3 GAMLSS

Como apresentado no Capítulo 2, os modelos aditivos generalizados para posição, escala e forma além de aumentarem as possibilidades das distribuições da variável resposta, permitem que esta possua uma grande assimetria e curtose, modelam não apenas a média como os demais parâmetros por meio de funções paramétricas ou não paramétricas das covariáveis e efeitos aleatórios, garantindo uma estrutura mais flexível ao modelo.

De acordo com Florencio (2010), o processo de construção e seleção de um modelo GAMLSS consiste em comparar diversos modelos concorrentes em que diferentes combinações dos componentes são utilizadas. Sua construção consiste basicamente das seguintes etapas: (i) identificação das distribuições plausíveis para a variável resposta; (ii) escolha da função de ligação para modelar o parâmetro de posição (μ); (iii) aplicação da técnica stepwise de seleção de covariáveis para modelar μ; (iv) inclusão de termos aditivos não-paramétricos, a exemplo de splines; (v) escolha da função de ligação para modelar o parâmetro de escala (σ); (vi) aplicação da técnica stepwise de seleção de covariáveis para modelar σ.

5.1.3.1 Modelagem do parâmetro (μ)

Com o intuito de ajustar os dados utilizando o modelo *GAMLSS*, inicialmente foi construído um gráfico para verificar quais distribuições se ajustam bem à variável resposta (valor do lote (R$)). De acordo com as características dos dados, as quais já foram bastante comentadas no decorrer do trabalho, é plausível utilizar as distribuições Gama, Weibull, Inversa Gaussiana e Log-Normal. Na Figura 5.7 encontram-se os gráficos com cada uma dessas distribuições, obtendo os seguintes AICs: GA (4553.284), WEI (4571.384), IG (4525.917) e LOGNO (4530.690). Observa-se que as distribuições gama, lognormal e gaussiana inversa, traçadas na cor vermelha, parecem estar mais próximas da função densidade de probabilidade estimada não-parametricamente (traçada na cor azul), indicando que estas distribuições apresentam uma maior aderência aos dados.

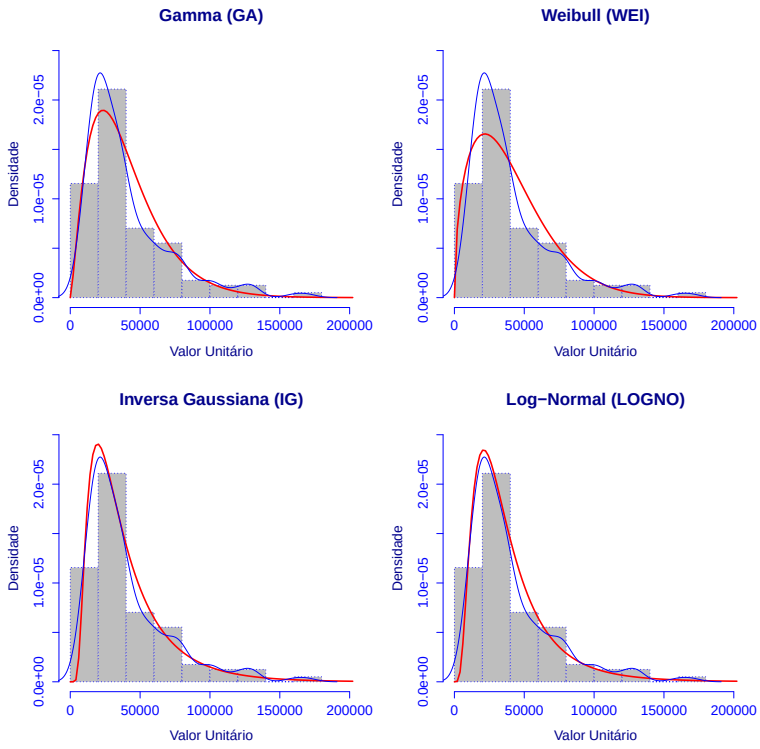


Figura 5.6: Ajuste das distribuições GA, WEI, IG e LOGNO à variável resposta.

Na Tabela 5.6 encontram-se os resultados obtidos na comparação dos modelos Gama, Weibull, Inversa Gaussiana e Log-Normal considerando o modelo 5.1, ou seja, considerando a variável AREA, as variáveis de localização descritas na Tabela 5.1 e a interação de cada uma delas com a AREA. Observa-se que a distribuição Gama apresenta os melhores valores em todos os critérios usados para a comparação.

Tabela 5.5: Resultados de AIC, SBC, GD e $pseudo - R^2$ considerando as distribuições Gama, Weibull, Inversa Gaussiana e Log-Normal.

Distribuição	AIC	SBC	GD	$pseudo - R^2$
Gama	4301.345	4357.332	4267.345	0.8111753
Normal Inversa	4341.373	4397.359	4307.373	0.7715297
Log Normal	4306.576	4362.562	4272.576	0.7385614
Weibull	4309.258	4365.244	4275.258	0.8037892

Após o ajuste do primeiro modelo para as diferentes distribuições para a variável resposta, realizou-se a seleção das covariáveis com o auxilio das ferramentas stepGAIC()[1], stepGAIC.CH()[2], addterm() e dropterm()[3]. Em cada modelo testado verificou-se a significância das variáveis, a inclusão de suavizadores (splines cúbicos) e critérios de comparação de modelos, como GD, AIC, SBC e Pseudo- R^2. Dentre os principais modelos testados, optou-se pelo modelo Gama com função de ligação log por apresentar os melhores resultados nos critérios AIC, SBC e $pseudoR^2$. Ressalta-se ainda que outras combinações foram utilizadas, mas não apresentaram resultados superiores.

A fim de melhorar o ajuste, utilizou-se a função *find.hyper* do pacote gamlss do R para encontrar o número de graus de liberdade ótimo para os suavizadores. Para a verificação do modelo adequado, utilizou-se gráficos de resíduos, gráfico de probabilidade normal e *wormplots*.

A Tabela 5.7 sumariza as estimativas dos parâmetros para o modelo Gama com função de ligação logarítmica para o parâmetro μ. Pela Figura 5.8, o gráfico dos "Valores ajustados x Quantis dos resíduos"distribuidos aleatoriamente, indicam que a variância dos resíduos é constante. O gráfico "Ordem x Quantis dos resíduos"indica a indepêndencia dos resíduos, enquanto que o gráfico "Quantis dos resíduos x Densidade"fornece a idéia da distribuição dos resíduos. Por último, o gráfico "Quantil teórico x Quantil amostral"que indica normalidade dos resíduos. Por esses quatro gráficos de resíduos quantílicos normalizados verifica-se que o pressuposto da normalidade dos resíduos é satisfeito, ou seja, o modelo se ajusta bem aos dados, pois seus resíduos apresentam distribuição próxima de uma normal padrão (média zero e variância 1).

[1]A função stepGAIC() realiza a seleção de modelos (stepwise) usando o critério de informação de AKAIKE generalizado, a qual pode solicitar outras duas funções, stepGAIC.VR() e stepGAIC.CH(), dependendo do argumento aditivo.

[2]A função stepGAIC.CH é baseada na função step.gam () do S (ver Chambers e Hastie (1991)) e é mais adequada para modelos com suavização dos termos aditivos. É usada para construir modelos para os parâmetros individuais da distribuição da variável resposta. (help do *R*)

[3]Funções da biblioteca MASS do *R*, vide Venables & Ripley, 2002.

Na Figura 5.9 (a) apresenta-se o *Worm-plot*[4], o qual indica que o modelo proposto está bem ajustado aos dados, pois os pontos estão situados no interior da região de "aceitação"(entre as duas curvas elípticas). E como último gráfico, tem-se na Figura 5.9 (b) que os pontos estão em torno de uma reta indicando a adequabilidade do modelo proposto aos dados.

Tabela 5.6: Estimativa dos parâmetros, erro padrão e p-valor do Modelo *GAMLSS*.

Coenficiente de μ			
Variável	Estimativa	Erro Padrão	p-valor
Intercepto	8.510	0.132	0.000
PLN.CENTRAL	0.085	0.201	0.671
FERROVIA	0.259	0.069	0.000
RODOVIAWL	0.534	0.087	0.000
CONDO	0.789	0.285	0.006
FECHADO	-1.078	0.236	0.000
ESTRITRESID	0.737	0.175	0.000
cs(AREA, df=3)	0.001	0.000	0.000
cs(AREA:NUCPRINC, df = 5)	0.002	0.000	0.000
AREA:PLN.CENTRAL	0.0003	0.000	0.499
AREA:CONDO	-0.002	0.001	0.006
cs(AREA:FECHADO, df = 5)	0.003	0.001	0.000
cs(AREA:ESTRITRESID, df = 4)	-0.002	0.000	0.000

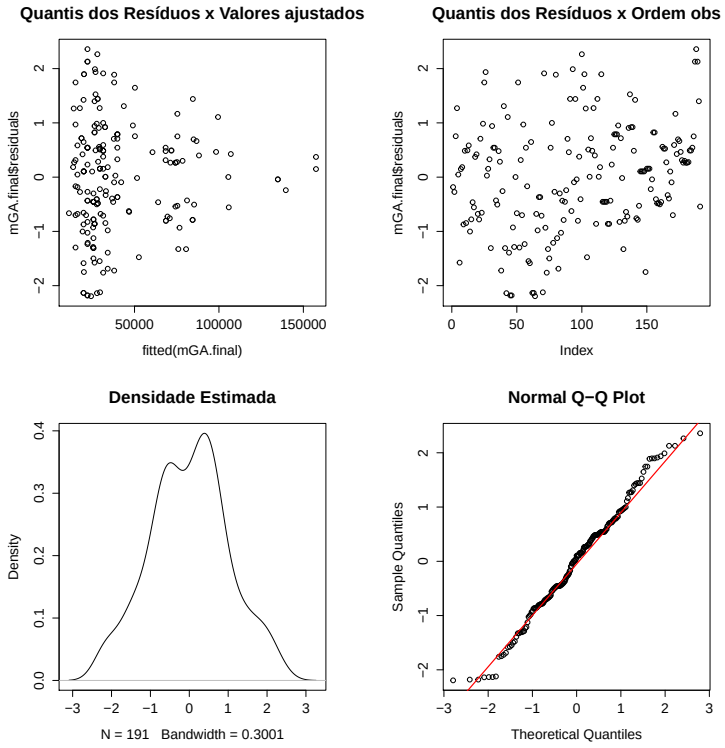
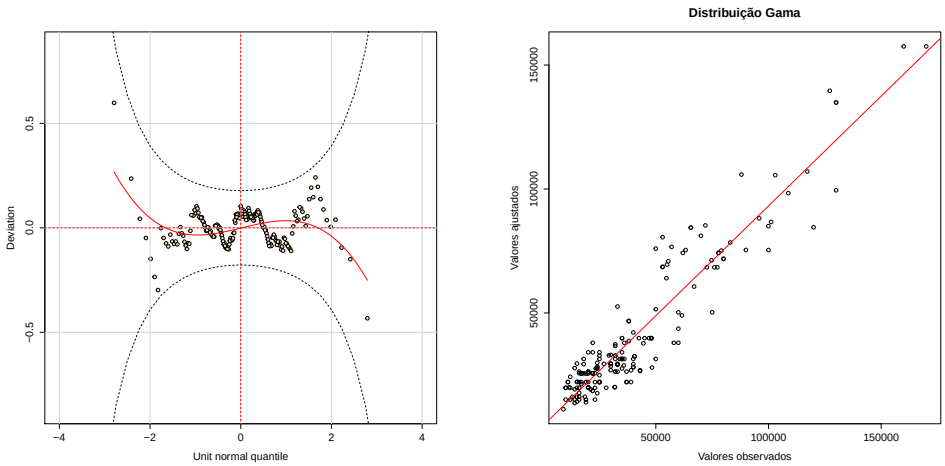


Figura 5.7: Gráficos de diagnósticos com relação ao modelo ajustado pela distribuição Gama com função de ligação log.

[4]Seção 2.4.4 e para maiores detalhes consultar Buuren & Fredriks.



(a) (b)

Figura 5.8: Gráfico Worm-plot e Gráfico dos Valores observados x Ajustados do modelo ajustado pela distribuição Gama com função de ligação log. (a) e (b)

No *GAMLSS* não é possível fazer uma interpretação direta dos coeficientes estimados, no entanto, pode-se examinar os sinais destes coeficientes e fazer algumas considerações a respeito do mercado imobiliário em estudo, como por exemplo que CONDO E ESTRITRESID são as variáveis que exercem maior peso na valorização do imóvel, enquanto que a variável FECHADO tem um peso negativo, o qual pode ser explicado devido a falta de critério para diferenciar os graus de fechamento desses parcelamentos. O sinal positivo do coeficiente AREA indica que o preço unitário médio dos terrenos aumenta à medida que a área dos lotes também aumenta (lembrando que foram considerados apenas lotes com área $\leq 800\ m^2$). O que é bastante coerente, pois ao aumentar o tamanho de um terreno e fixando as demais características, é natural que quanto maior o terreno, maior seja seu valor.

Nos três modelos considerados, tanto a variável PLNCENTRAL como sua iteração com a área, AREA:PLNCENTRAL, embora não tenham sido significativas para o modelo, foram deixadas para seguir a proposta de Ferreira (2007) e Ferraudo (2008), os quais consideraram estas variáveis fundamentais nos modelos conforme se fez a construção territorial de São Carlos.

5.1.3.2 Modelagem do parâmetro de dispersão (σ)

Após encontrar um modelo para o parâmetro μ, as mesmas etapas são necessárias para modelar o parâmetro σ. Assim, fez-se a seleção de variáveis, verificou a significância das mesmas e a inclusão de suavizadores. Para verificar a adequação do modelo, utilizou-se gráficos de resíduos *worm-plots*.

Na Tabela 5.7 encontram-se as estimativas dos parâmetros para o modelo Gama com função de ligação logarítmica para o parâmetros μ e σ. Nota-se que as estimativas dos parâmetros do modelo de μ, após incluir a modelagem de σ, são próximas das estimativas somente com o modelo de μ. Na Figura 5.9 encontra-se os gráficos de resíduos quantílicos normalizados os quais indicam que os resíduos apresentam distribuição próxima de uma normal padrão. Na Figura 5.10(a) o gráfico *worm-plot* apresenta os pontos situados dentro da região de aceitação

e por fim na Figura 5.10(b) encontra-se o gráfico dos valores observados versus os valores ajustados, nos quais os pontos situam-se próximos de uma reta. Os gráficos de diagnóstico sugerem que o modelo proposto é adequado aos dados.

O modelo de σ é estritamente paramétrico, não apresentando termos aditivos na sua composição, sendo consideradas apenas três variáveis significativas. O sinal negativo do coeficiente estimado de CONDO e AREA:FERROVIA indicam que a variabilidade do preço dos imóveis nessas regiões são menores, enquanto que o sinal positivo do coeficiente estimado de AREA:CONDO indica que a medida que a área de lotes situados em condomínios crescem, a variação dos preços é maior.

Tabela 5.7: Estimativa dos parâmetros, erro padrão e p-valor.

Coeficiente de μ			
Variável	Estimativa	Erro Padrão	p-valor
Intercepto	8.572	0.146	0.000
PLN.CENTRAL	0.079	0.232	0,734
FERROVIA	0.255	0.086	0,003
RODOVIA	0.534	0.094	0.000
CONDO	0.778	0.169	0.000
FECHADO	-1.183	0.244	0.000
ESTRITRESID	0.457	0.180	0.012
cs(AREA, df = 3)	0.001	0.000	0.000
cs(AREA:NUCPRINC, df = 5)	0.002	0.000	0.000
AREA:PLNCENTRAL	0.0003	0.001	0,546
CONDO	-0.002	0.000	0.000
cs(AREA:FECHADO, df = 5)	0.004	0.001	0.000
cs(AREA:ESTRITRESID, df = 4)	-0.001	0.000	0.013
Coeficiente de σ			
Variável	Estimativa	Erro Padrão	p-valor
Intercepto	-1.017	0.116	0.000
CONDO	-2.646	0.576	0.000
AREA:FERROVIA	-0.001	0.000	0.006
AREA:CONDO	0.004	0.001	0.004

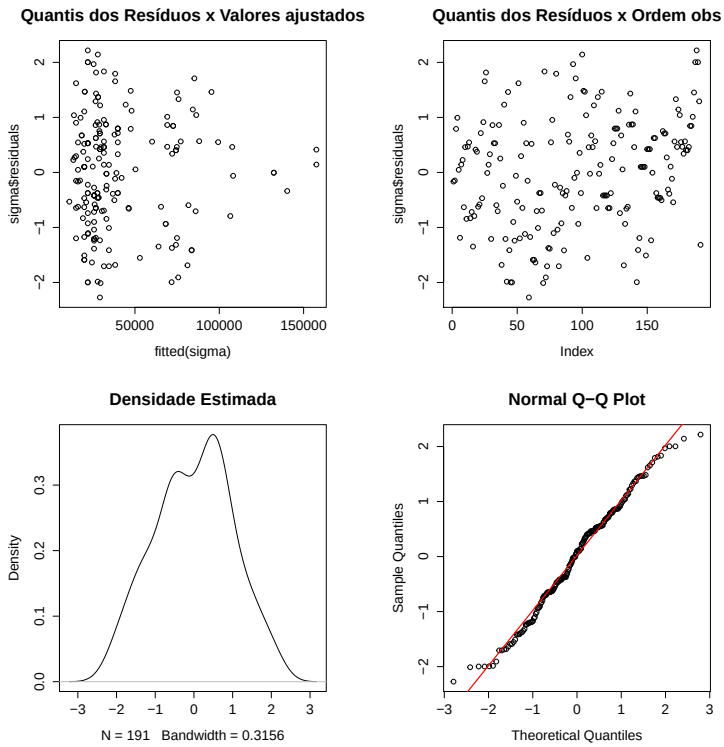
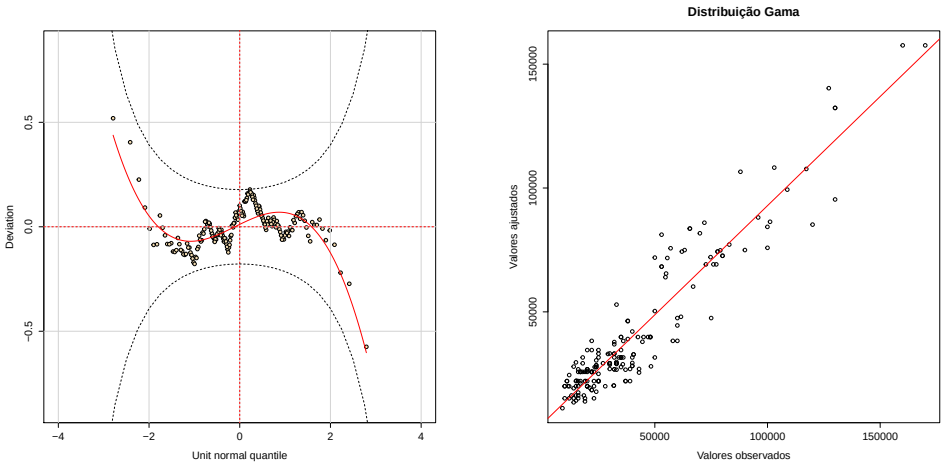


Figura 5.9: Gráficos de diagnósticos com relação ao modelo ajustado pela distribuição Gama com função de ligação log nos parâmetros μ e σ.



(a) (b)

Figura 5.10: Gráfico Worm-plot e Gráfico dos Valores observados x Ajustados do modelo ajustado pela distribuição Gama com função de ligação log nos parâmetros μ e σ. (a) e (b)

O modelo final é dado por:

$$
\begin{aligned}
log(\mu) =& 8.57 + 0.08 * PLNCENTRAL_i + 0.25 * FERROVIA_i + \\
& 0.53 * RODOVIAWL_i + 0.78 * CONDO_i - 1.18 * FECHADO_i + \\
& 0.46 * ESTRITRESID_i + 0.001 * AREA_i + \\
& 0.002 * (NUCPRINC : AREA)_i + 0.0004(PLNCENTRAL : AREA)_i - \\
& 0.002 * (CONDO : AREA)_i + 0.004 * (FECHADO : AREA)_i - \\
& 0.001 * (ESTRITRESID : AREA)_i,
\end{aligned}
$$

e

$$
\begin{aligned}
log(\sigma) =& -1.02 - 2.65 * CONDO_i - 0.0008(AREA : FERROVIA)_i + \\
& 0.004(AREA : CONDO),
\end{aligned}
$$

em que a variável resposta (valor do lote) segue uma distribuição gama (GA) com parâmetro de posição (μ) e de escala (σ).

5.1.4 Comparação dos modelos

Na Tabela 5.8 é apresentado um resumo comparativo entre os modelos estimados via regressão linear, modelo linear generalizado (MLG) e modelo aditivo generalizado para posição escala e forma (GAMLSS) para o parâmetro μ. Os critérios utilizados serão o Pseudo-R^2, capacidade preditiva do modelo e taxa de aceitação.

Como se pode observar, o modelo linear apresentou os piores resultados em todos os critérios. No que diz respeito a taxa de aceitação, o MLG apresentou a melhor taxa, com 72%, enquanto que o GAMLSS mostrou o melhor resultado no valor do Pseudo-R^2 e melhor capacidade preditiva.

Tabela 5.8: Resumo comparativo dos modelos Linear, GLM e GAMLSS.

Classe	$pseudo - R^2$	Capacidade Preditiva	Taxa de aceitação
Normal Linear	0.8534	$69,8\%$	68%
GLM (Gama)	0.8611	$73,7\%$	72%
GAMLSS(Gama)	0.8863	$75,6\%$	$70,5\%$

5.2 Estudo de Influência Local

Embora a parte de diagnósticos já tenha sido realizada nos modelos propostos anteriormente, ao assumir o modelo mais adequado, de acordo com os critérios utilizados, o que deseja-se agora é estudar sua robustez sob pequenas perturbações. Outra observação que deve ser ressaltada é que o estudo do efeito dessas perturbações é feito em 100% dos dados, sem que haja a separação ocorrida nos ajustes anteriores para a validação. Este estudo será realizado abordando o modelo linear generalizado. Para isso, foi considerando o modelo 5.4.

Na Tabela 5.9 encontram-se as estimativas , o erro padrão e o p-valor do ajuste considerando a variável resposta pertencente à família gama com função de ligação log. Nas Figuras 5.11 e 5.12 encontram-se a análise dos resíduos e o gráfico de envelope, os quais mostram que o ajuste apresenta alguns pontos fora dos "limites aceitáveis". No gráfico de envelope, os pontos que estão fora da banda de confiança são 268 e 156.

Com isso, adotando a metodologia de Cook (1986) com o esquema de perturbação de casos, tem-se $C_{l_{max}} = 2.51$ calculada de 2.47. Segundo Cook(1986), uma referência para a curvatura é o valor 2, ou seja, se $C_{l_{max}}$ apresenta valor superior a 2, existe o indício de que as observações são globalmente influentes, isto é, existe ao menos uma observação influente. Na Figura 5.13(a) é apresentado o gráfico do autovetor correspondente a $C_{l_{max}}$ e na Figura 5.13(b) a influência local total. Observe que as observações 201 e 156 destacam-se das demais. No entanto, este método detecta apenas pontos pontencialmente atípicos, sendo necessário realizar novamente os ajustes sem estas observações para verificar se ocorre uma mudança significativa na inferência dos coeficientes do modelo. Embora a metodologia proposta por Cook(1986) seja considerada inovadora principalmente por não necessitar da deleção de pontos, a retirada dos mesmos ainda é bastante utilizada para a "confirmação "de sua influência.

Na reanálise dos dados retiraram-se os pontos que se mostraram mais discrepantes e novamente realizou as técnicas de diagnósticos e o método de Cook (1886), os quais melhoraram significativamente obtendo um $C_{l_{max}} = 2.06$. Como o intuito desta seção é apenas mostrar a eficiência do método e encontrar os possíveis pontos influentes localmente, serão apresentados novamente apenas o gráfico de envelope na Figura 5.14, o qual encontra-se dentro dos limites.

Tabela 5.9: Estimativa dos parâmetros,erro padrão e p-valor utilizando o modelo gama com ligação logarítmica com todas as observações.

Variável	Estimativa	Erro Padrão	p-valor
Intercepto	9.444	0.165	0.0000
NUCPRINC	-0.882	0.162	0.0000
PLNCENTRAL	0.608	0.221	0.0064
FERROVIA	0.098	0.073	0.1824
RODOVIAWL	0.550	0.096	0.0000
CONDO	0.801	0.307	0.0094
FECHADO	-1.004	0.280	0.0004
ESTRITRESID	0.501	0.207	0.0164
NUCPRINC*AREA	0.003	0.000	0.0000
PLNCENTRAL*AREA	-0.0004	0.000	0.4602
CONDO*AREA	-0.001	0.000	0.0705
FECHADO*AREA	0.003	0.000	0.0004
ESTRITRESID*AREA	-0.001	0.000	0.0114

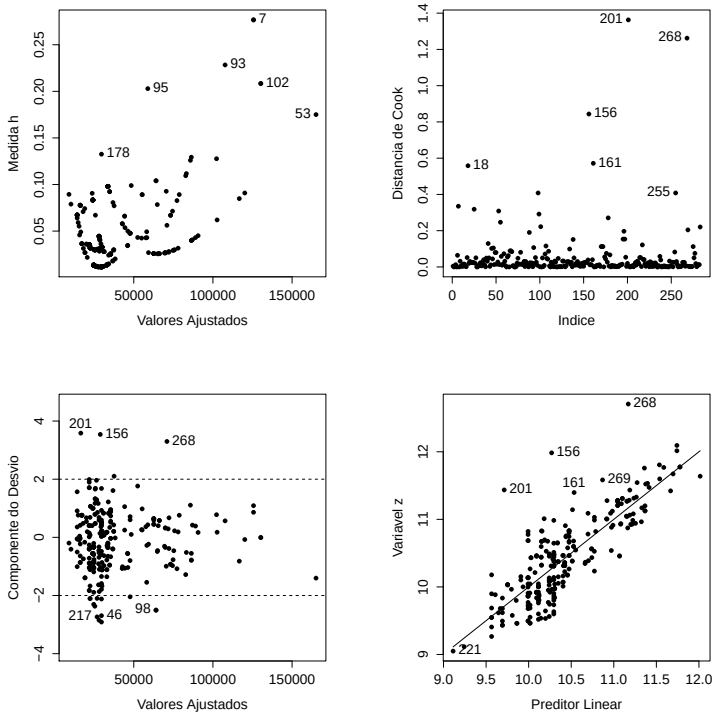


Figura 5.11: Gráfico de diagnóstico para o modelo gama com função de ligação log com todas as observações.

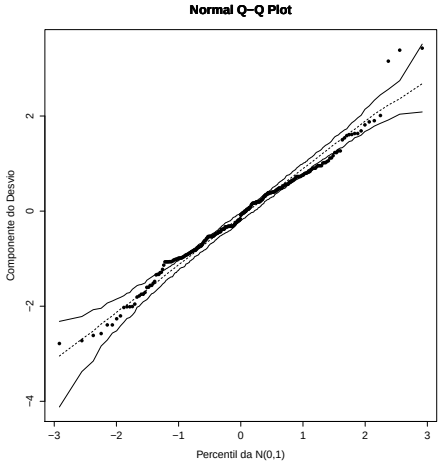
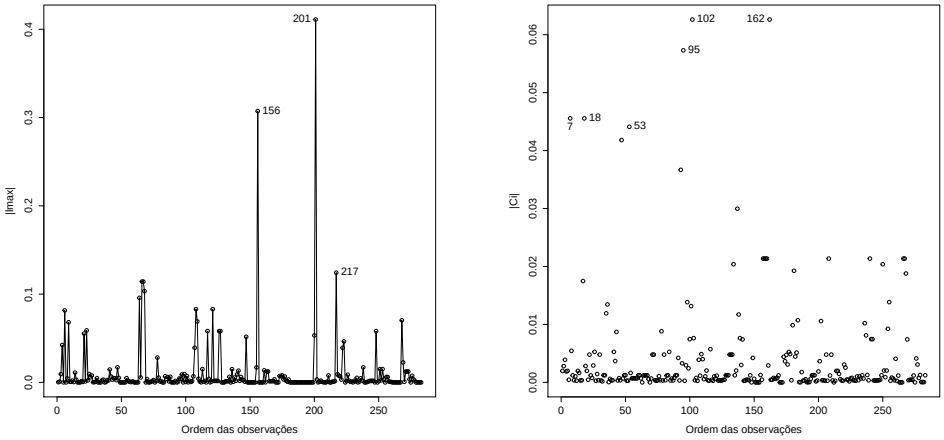


Figura 5.12: Gráfico de envelope para a componente desvio com todas as observações.



(a) Gráfico de influência - ponderação de casos.

(b) Gráfico de influência local.

Figura 5.13: Gráficos de influência local. (a) e (b)

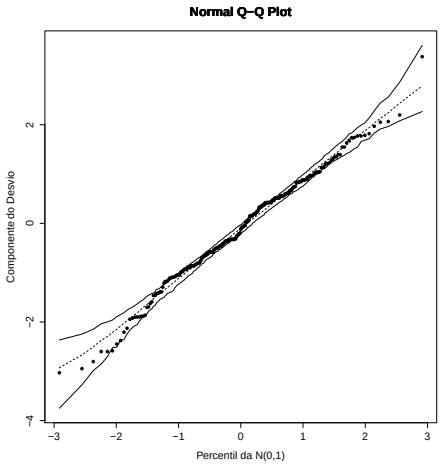


Figura 5.14: Gráfico de envelope após a retirada das observações 156 e 201.

Na Tabela 5.10 encontram-se as mudanças relativas em porcentagem de cada variável do modelo gama. Observa-se que para quase todas as variáveis, a mudança relativa foi muito maior que 0.003, ou seja, as observações 156 e 201 são realmente influentes e causam um efeito desproporcional nas estimativas do modelo.

Outra característica é a mudança inferencial, pois antes de retirar cada uma das observações, a variável FERROVIA não era significativa para o modelo, no entanto, conforme retirou-se essas observações, ela passou a ser significativa no modelo.

Tabela 5.10: Mudanças Relativas em porcentagem de cada variável do modelo gama com função logarítmica.

	Casos Removidos		
Variável	156	201	156, 201
Intercepto	0,62	0,51	1,18
NUCPRINC	2,92	9,06	6,48
PLNCENTRAL	2,82	8,69	6,22
FERROVIA	57,01	57,26	119,78
RODOVIAWL	0,49	1,53	1,09
CONDO	0,00	0,00	0,00
FECHADO	1,28	3,99	2,85
ESTRITRESID	2,57	7,92	5,67
NUCPRINC*AREA	2,47	7,69	5,50
PLNCENTRAL*AREA	13,35	41,34	29,58
CONDO*AREA	0,00	0,00	0,00
FECHADO*AREA	1,36	4,25	3,03
ESTRITRESID*AREA	3,23	9,99	7,15

5.3 Análise de Sobrevivência

Nesta seção, o que se deseja é incluir os lotes urbanos não vendidos no estudo e assim, verificar a importância de se considerar essa diferença (imóveis vendidos e também os não vendidos) na modelagem dos dados.

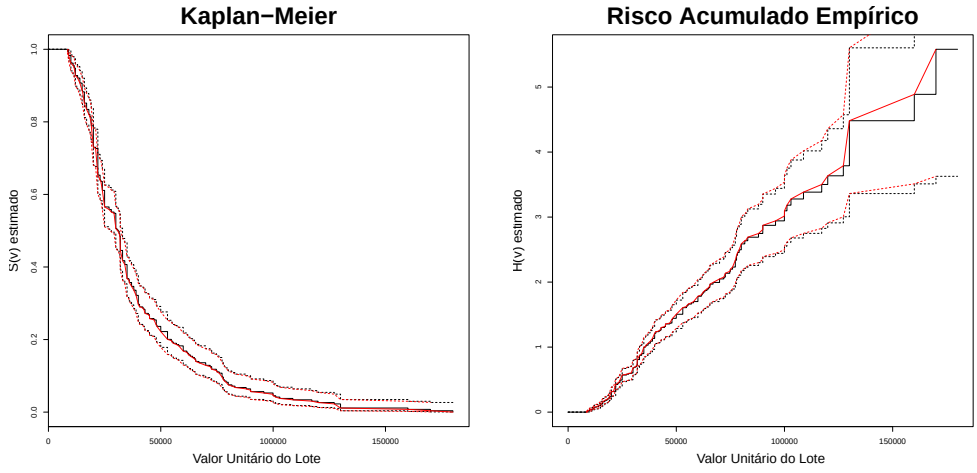
Nesta fase, será abordada a análise de sobrevivência ressaltando que quem introduziu essa abordagem em dados imobiliário foi Ferraudo (2008). A diferença entre os dois trabalhos está na modelagem dos parâmetros do modelo. Enquanto Ferraudo (2008) utilizou o pacote *Survreg*, neste trabalho será usado o pacote *GAMLSS*, que como descrito no Capítulo 2 é mais flexível, pois permite a modelagem de todos os parâmetros da distribuição e a inclusão de termos aditivos, podendo ser modelados através de suavizadores, como os *splines cúbicos*.

Serão utilizados apenas os lotes com área menor ou igual a $800m^2$, efetivamente vendidos e não vendidos (que serão considerados como censurados) em 2005. A variável resposta será o valor do lote e as variáveis de interesse são as mesmas descritas na Tabela 5.1, ou seja, 8 variáveis de localização dummy, a variável área do lote e mais a iteração de cada variável de

localização com a variável área. Por fim, será considerada a distribuição Weibull para a variável resposta e a censura do tipo censura à esquerda.

Análise descritiva dos dados

Em análise de sobrevivência, as análises descritivas das variáveis em estudo consiste em utilizar os métodos não-paramétricos, como o de Kaplan-Meier apresentado em 3.3. Na Figura 5.16(a) encontra-se a Função de permanência à venda estimada através do método de Kaplan-Meier e na Figura 5.16(b) apresenta-se a Função de Risco acumulado empírico. A medida que o valor do lote aumenta, o risco de vendê-lo também aumenta, ou seja, conforme aumenta o valor do lote, diminui a chance de não vendê-lo.



(a) . (b) .

Figura 5.15: (a) Curva de permanência à venda estimada pelo método de Kaplan-Meier; (b) Risco acumulado de venda empírico e os respectivos intervalos 95% de confiança.

Para verificar a escolha do modelo Weibull, na Figura 5.17 encontra-se o gráfico linearizado para o modelo Weibull, o qual não mostra afastamento marcante de uma reta, o que indica que a distribuição escolhida é adequada para ajustar os dados.

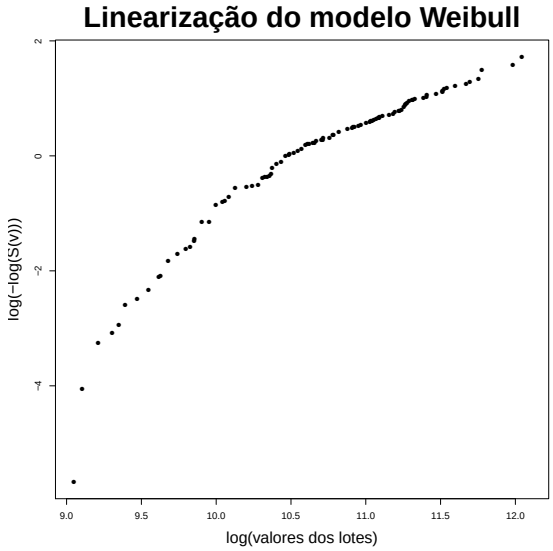


Figura 5.16: Gráfico de $-log(v)$ versus $log(-log(\widehat{S}(v)))$.

5.3.1 Modelagem de dados censurados à esquerda

Após a análise descritiva, uma vez que a distribuição da variável resposta já foi escolhida, o próximo passo é escolher as covariáveis do modelo. Para isso, o ajuste será feito utilizando as funções *Surv* do pacote *Survival* o qual permite especificar o tipo de censura e o pacote *GAMLSS*.

Os procedimentos para ajustar o modelo através do *GAMLSS* são exatamente como no caso sem censura explicados do Capítulo 2, sendo necessário apenas identificar que se está trabalhando com dados censurados e especificar a distribuição, neste caso a distribuição Weibull, e o tipo da censura (à esquerda).

Para a seleção das covariáveis do modelo foram utilizadas as ferramentas stepGAIC(), stepGAIC.CH(), addterm() e dropterm(). Em cada modelo testado verificou-se a significância das variáveis, a inclusão de suavizadores (splines cúbicos) e critérios de comparação de modelos, como GD, AIC, SBC e Pseudo- R^2.

As estimativas dos parâmetros, erro padrão e p-valor do modelo Weibull considerando censura à esquerda encontram-se na Tabela 5.11. Os coeficientes estimados estão expressos na escala logarítmica dos valores, isto é, para $Y = \log(V) = \boldsymbol{x}'\boldsymbol{\beta}$.

Para a verificação do modelo adequado, utilizou-se gráficos de resíduos, gráfico de probabilidade normal e *worm-plots*, os quais mostraram que o ajuste do modelo foi satisfatório.

Tabela 5.11: Estimativa dos parâmetros, erro padrão e p-valor do modelo Weibull.

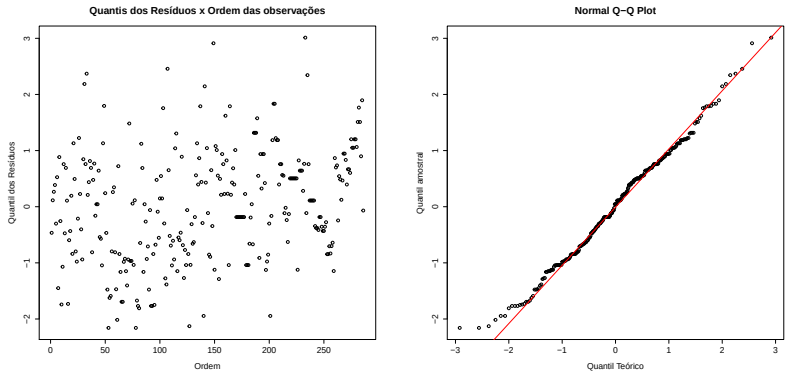
Coenficiente de μ			
Variável	Estimativa	Erro Padrão	p-valor
Intercepto	9,214	0,07	0,00
PLNCENTRAL	0,505	0,06	0,00
FECHADO	-0,823	0,18	0,00
ESTRITRESID	0,538	0,16	0,00
NUCPRINC*AREA	0,002	0,00	0,00
RODOVIAWL*AREA	0,002	0,00	0,00
FECHADO*AREA	0,003	0,00	0,00
ESTRITRESID*AREA	-0,002	0,00	0,00
Coeficiente de σ			
Variável	Estimativa	Erro Padrão	p-valor
Intercepto	1,212	0,04	0,00

Na Figura 5.17(a) encontra-se o gráfico Ordem $\times$ Quantis dos resíduos, os quais estão distribuídos aleatoriamente, indicando a independência dos resíduos. Na Figura 5.17(b) o gráfico mostra a relação entre os valores observados e os ajustados, os quais encontram-se próximos de uma reta.

Ao gerar os gráficos de diagnóstico, o *R* também exibe um sumário de medidas resumo da distribuição dos resíduos. Neste sumário tem-se média $= 0,01$, variância $= 0,96$, Coef. de assimetria $= 0,16$ e Coef. de curtose $= 2,8$ o qual, assim como os gráficos, indica que os resíduos possuem distribuição aproximada de uma normal padrão sugerindo adequabilidade do modelo proposto.

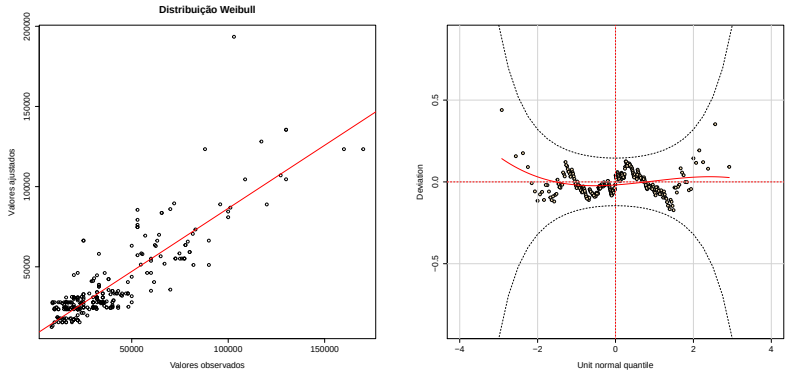
O gráfico de Valores Observados $\times$ Valores Ajustados apresentado na Figura 5.18(a) mostra que os pontos estão em torno de uma reta, como o desejado. Por fim, o gráfico *Worm-plot* da Figura 5.18(b) exibe os pontos dentro da região de aceitação, sugerindo que o modelo é apropriado.

Para avaliar o ajuste do modelo de regressão Weibull aos dados, foram utilizados os resíduos de Cox-Snell, os quais devem ser vistos como provenientes de uma amostra aleatória da distribuição exponencial padrão. Na Figura 5.19(a) encontra-se o gráfico dos pares de pontos $(\widehat{S}(\widehat{e}_i), \widehat{S}(\widehat{e}_i)_{Exp})$ os quais se aproximam de uma reta. Na Figura 5.19(b) as estimativas das curvas de sobrevivência desses resíduos obtidas por Kaplan-Meier $(\widehat{S}(\widehat{e}_i))$ e pelo modelos exponencial padrão estão próximas. Assim, conclui-se que o modelo ajustado pode ser considerado satisfatório.



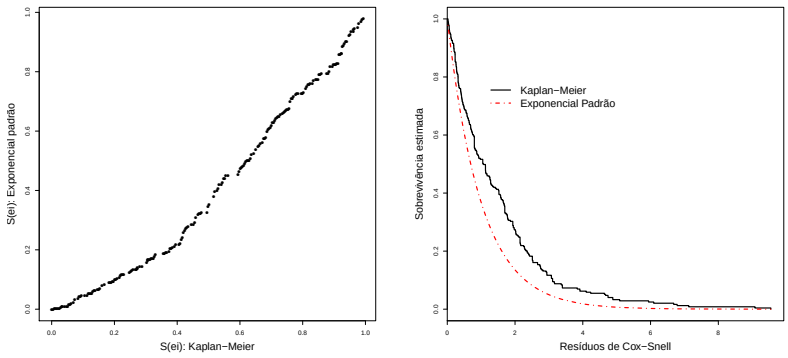
(a) . (b) .

Figura 5.17: Gráficos de diagnósticos com relação ao modelo ajustado pela distribuição Weilbull com funçãode ligação log para os parâmetros μ e σ.



(a) . (b) .

Figura 5.18: Gráfico Worm-plot e Gráfico dos Valores observados x Ajustados do modelo ajustado pela distribuição Weibull.



(a) . (b) .

Figura 5.19: Análise gráfica dos resíduos de Cox-Snell do modelo de regressão Weibull.

O modelo final apresentou um $pseudo - R^2$ de 70% e uma capacidade preditiva de 67%. Ressalta-se ainda que foi realizada a modelagem utilizando o pacote *Survreg*, o qual apresentou

um $pseudo - R^2$ de 62% e capacidade preditiva de $63,5\%$, no entanto os resultados foram omitidos.

A função de sobrevivência obtida pelo modelo de regressão Weibull ajustado para os dados de lotes urbanos da cidade de São Carlos do ano de 2005 é, portanto, expressa por

$$\widehat{S}(v|\boldsymbol{x}) = \exp\left\{-\left(\frac{v}{\exp\{\boldsymbol{x}'\boldsymbol{\beta}\}}\right)^{1/\sigma}\right\}, \tag{5.5}$$

com $\boldsymbol{x}'\boldsymbol{\beta} = 9,214 + 0,505 * x_1 - 0,823 * x_2 + 0,538 * x_3 + 0,002 * x_4 + 0,002 * x_5 + 0,003 * x_6 - 0,002 * x_7$ e $\sigma = 0.297$, o qual é obtido pelo comando no *R*, sigma=1/exp(coef(modelo,"sigma")).

Na Figura 5.20 encontram-se os gráficos do ajuste da curva de Kaplan-Meier e da sobrevivência estimada pelo modelo de regressão Weibull, os quais estão próximos indicando que o modelo ajustado está coerente aos dados.

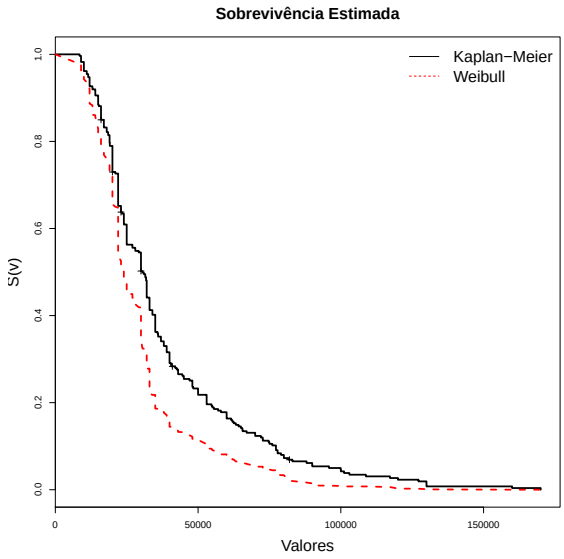


Figura 5.20: Estimativa de Kaplan-Meier e função de sobrevivência estimada do modelo Weibull.

Os coeficientes negativos das variáveis de localização, como por exemplo Fechado, implica que os imóveis localizados nessa região apresentam probabilidade de sobrevivência (permanência à venda) estimada menor do que os imóveis localizados fora da mesma. No caso das variáveis iteradas com a Área com sinais positivos indicam que quanto maior o valor de x, maior a probabilidade de permanência à venda.

5.4 Considerações Finais

Neste capítulo foi aplicado a um conjunto de dados de lotes urbanos da cidade de São Carlos, do ano de 2005, as diferentes metodologias apresentadas no trabalho. Inicialmente foi considerado apenas os lotes efetivamente vendidos e ao abordar a análise de sobrevivência, os lotes em negociação também foram incorporados aos dados.

Ao encontrar as equações representativas da formação do valor de mercado de lotes foi possível perceber as restrições mencionadas na teoria. Embora se consiga um modelo que seja adequado aos dados, algumas vezes é necessário perder algumas informações, como foi o caso do modelo linear onde as observações do valor do imóvel foram transformadas. No modelo linear generalizado e no GAMLSS, por utilizar a mesma distribuição e a função de ligação logarítmica para o parâmetro μ, a restrição do MLG de pertencer à família exponencial não teve uma diferença significativa. No entanto, como o GAMLSS possibilita a modelagem dos demais parâmetros da distribuição da variável resposta e a utilização de suavizadores, do ponto de vista prático, proporcionou o melhor ajuste aos dados de lotes urbanos.

A vantagem da metodologia de influência local (Cook,1986) reside na sua capacidade de avaliar mudanças nos resultados da análise quando pequenas perturbações são incorporadas no modelo e/ou nos dados e com isso indicar possíveis pontos influentes e inadequações do modelo.

Por fim, o modelo de sobrevivência indicou quais regiões estão mais propicias a permanecerem à venda e a utilização do *GAMLSS* na modelagem dos parâmetros se mostrou eficaz.

Capítulo

6

Conclusão e Propostas Futuras

Este trabalho teve como objetivo principal, analisar dados de lotes urbanos da cidade de São Carlos no ano de 2005 abordando diferentes métodos estatísticos para a formação de equações de regressão representativas do valor de mercado desses imóveis.

No Capítulo 2 foram apresentados os principais tópicos das teorias abordadas, suas características e propriedades, juntamente com a influência local, que embora seja mais uma técnica de diagnóstico, permite que a influência dos dados seja estudada sem a deleção de pontos.

No Capítulo 3 foi apresentada a metodologia de análise de sobrevivência e algumas de suas propriedades, a qual é mais flexível, pois permite incluir no processo de modelagem os imóveis efetivamente negociados (não censurado) e os em negociação (censurado).

Nos estudos de simulação realizados com o objetivo de verificar as propriedades assintóticas dos estimadores observa-se que os resultados dos intervalos de confiança de 95% são adequados para tamanhos de amostras moderadas. Verifica-se ainda que no modelo linear, o aumento da variância pode ocasionar o aumento dos EQMS nos estimadores, o mesmo vale ao considerar baixos valores para o parâmetro de forma ν no modelo linear generalizado e no modelo *GAMLSS*.

Para o caso quando há censura, observa-se que os estimadores satisfazem as propriedades assintóticas para amostras maiores que 60 e para uma porcentagem baixa de censuras. Pois a partir de 15%, o aumento de censura implica em uma diminuição da probabilidade de cobertura do parâmetro β_0 conforme o aumento da amostra.

Na estimação empírica da equação de preços hedônicos para os lotes urbanos, apresentada no Capítulo 5, as análises mostraram que embora o modelo linear seja adequado após a transformação da variável resposta, apresenta resultados inferiores ao ser comparado com o modelo linear generalizado, e ao modelo GAMLSS, o qual apresentou uma melhor capacidade

preditiva. Outra consideração importante é a capacidade do GAMLSS em modelar os demais parâmetros da distribuição e não apenas a média, podendo fornecer uma idéia sobre o comportamento da variabilidade dos preços dos lotes.

O estudo de influência local detectou os pontos potencialmente influentes, e após o cálculo da diferença relativa, foi possível observar que tais pontos estavam realmente causando um efeito desproporcional nas estimativas do modelo.

Na aplicação de dados censurados, foi considerada a distribuição Weibul para a variável resposta e a utilização do modelo *GAMLSS* apresentou resultados satisfatórios, identificando as regiões com maior probabilidade de permanencia à venda.

Para título de ilustração, na Tabela 6.1 encontram-se as estimativas referentes aos valores de dois lotes utilizando as diferentes metodologias utilizadas ao longo do trabalho, o valore real do lote, a característica de cada um (se foi ou não vendido) e a porcentagem do aumento do valor estimado em relação ao valor real.

Tabela 6.1: Valor estimado considerando dois lotes urbanos, um com negociação em andamento e outro com negociação finalizada, com diferentes modelos.

Censura	0	1
Valor do Lote	23000	62000
MRL	20553,23 (-10,64%)	71098,78 (14,67%)
MLG	22151,93 (-3,69%)	69914,22 (12,76%)
GAMLSS	22335,04 (-2,9%)	69914,29 (12,76%)
GAMLSS (censurado)	24209,79 (5,25%)	62466,28 (0,75%)

Observa-se que para o lote em negociação, as metodologias apresentaram valores próximos dos valores reais, sendo o modelo linear o que apresentou a maior diferença. Para os lotes já vendidos, a metodologia de análise de sobrevivência utilizando GAMLSS apresentou uma estimativa mais próxima do valor real.

Ressalta-se ainda que muitas variáveis de importância expressiva não foram consideradas, como formato do imóvel, posição do interior da quadra, proximidade aos centros comerciais e serviços entre outras características que influenciam na formação do preço. Além disso, deve ser levado em conta que por se tratar de uma amostra de tamanho moderado (311 lotes), o poder de predição dos modelos é bastante considerável.

A principal contribuição deste trabalho foi a utilização do *GAMLSS* nos dados, considerando ou não a censura à esquerda. Essa ferramenta apresenta uma metodologia unificada de diversos métodos estatísticos, sendo bastante flexível, e com isso permitindo que seu uso seja abordado nos mais diversos estudos e considerando diferentes situações onde as demais técnicas sofrem limitações.

Como outros acréscimos tem-se o estudo de simulação considerando os modelos lineares generalizados e os modelos *GAMLSS* (com e sem censura) e a utilização da influência local no conjunto de dados.

Embora a metodologia com dados completos tenha sido trabalhada, ela não é a mais adequada uma vez que não é capaz de captar todas as informações dos dados. Cumpre ressaltar que até o trabalho de Ferraudo (2008), mesmo quando havia falta de informação nos dados, como por exemplo o valor do imóvel efetivamente vendido, eles eram tratados como dados completos. Considerando ainda que os únicos trabalhos que abordam a presença de censura nos dados imobiliários (este e o de Ferraudo (2008)) utilizaram o mesmo conjunto de dados, apesar da metodologia ter se mostrado eficiente, ainda é pouco para afirmar sobre a efetiva melhora dos modelos baseados em análise de sobrevivência ao se comparar com os modelos mais tradicionais (modelo linear e modelo linear generalizado).

Como propostas futuras, uma das sugestões é verificar a questão da correlação espacial juntamente com o *GAMLSS* e a presença de censura, pois diversas pesquisas, como Dantas (2003), sugerem a sua existência. Verificar a utilização de censura em outros dados imobiliários e com relação a simulação, pode-se ainda, estudar o acréscimo de covariáves no segundo parâmetro da distribuição para o caso do modelo *GAMLSS*.

Apêndice

A

Apêndice

Código para o ajuste do modelo com dados censurados considerando a classe *GAMLSS*

```
require(gamlss)
require(gamlss.cens)
require(survival)
require(gamlss.dist)
require(RODBC) # carregar dados do Excel

setwd('E:\\Amanda\\Mestrado\\art\\lotesanca')
x <- odbcConnectExcel('censura2.xls')
dados <- sqlFetch(x,'Plan1')

y = dados[,1]

#x1: nucleo principal
x1=dados$x1
#x2: pln_central
x2=dados$x2
#x3: ferrovia
x3=dados$x3
#x4: rodovia
x4=dados$x4
#x5: encosta
```

```
x5=dados$x5
#x6: condom
x6=dados$x6
#x7: fechado
x7=dados$x7
#x8: estrit_resid
x8=dados$x8
#x9: area
x9=dados$x9
#y: valor total do lote
y=dados$y
venda=dados$venda

x10 <-x1*x9
x11<-x2*x9
x12<-x3*x9
x13<-x4*x9
x14<-x5*x9
x15<-x6*x9
x16<-x7*x9
x17<-x8*x9


###Modelo inicial considerando todas as covariáveis e a iteração delas
com a variável área###

mg2<-gamlss(Surv(y, venda, type="left" ) ~ x1+x2
+x3+x4+x5+x6+x7+x8+x9+x10+x11+x12+x13+x14
+x15+x16+x17,data=dados,
            family=cens(WEI),type="left")
summary(mg2)
plot(mg2) #gráfico de resíduos para analisar o ajuste
wp(mg2) #gráfico de worm-plot

###Selecionando coveriáveis por stepGAIC###

dropterm(mg1, test="Chisq")
addterm(mg1, scope=~(x1+x2+x3+x4+x5+x6+x7+x8+x9+x10+x11+x12+x13
                     +x14+x15+x16+x17)^2, test="Chisq")
```

```
mgstep <- stepGAIC(mg1, scope=list(lower=~1,upper=~(x1+x2+x3
+x4+x5+x6+x7+x8+x9+x10+x11+x12+x13+x14+x15+x16+x17)^2))

#verificando a inclusao de termos aditivos, usando o mgstep

gs <- gamlss.scope(model.frame(Surv(y, venda, type = "left") ~ x1
+ x3 + x6  + x7+ x8 + x9 + x10 + x12 + x13 + x15 + x16 + x17))

#as covariáveis de gs foram selecionadas pelo stepGAIC


mg2 <- gamlss(Surv(y, venda, type="left" ) ~ 1,
              data=dados,
              family=cens(WEI, type="left"))

mg3 <- stepGAIC(mg2, gs, additive=TRUE)


#Nesta etapa  para escolher o modelo mais adequado deve ser levado
em consideração o modelo escolhido pelos métodos de seleção,
a significância estatística das variáveis e também a experiência
do pesquisador e o interesse prático que se busca com o modelo.

####Modelo final####

m7.5<- gamlss(formula = Surv(y, venda, type = "left") ~ x2 +
                x7 + x8 +  x10 +  x13 + x16 + x17,
              family = cens(WEI3, type = "left"), data = dados)

summary(m7.5)
plot(m7.5)
#gráfico de wormplot
wp(m7.5, ylim.all=15*sqrt(1/length(fitted(m7.5))))
pseudom7.5 <- cor(dados$y,fitted(m7.5))^2 #pseudom-R^2

#plotar os graficos de diagnósticos
par(mfrow = c(1,1))
plot(m7.5$residuals, main='Quantis dos Resíduos x Ordem
das observações',xlab="Ordem",ylab="Quantil dos Resíduos")
qqnorm(m7.5$residuals, main='Normal Q-Q Plot',xlab="Quantil Teórico",
```

```
ylab="Quantil amostral")
qqline(m7.5$residuals,col="red", lwd=1.5)

#gráfico dos valores observados x valores preditos
op1 <- par(mfrow=c(1,1))
plot(y,fitted(m7.5), main = 'Distribuição Weibull',
     xlab = 'Valores observados', ylab = 'Valores ajustados')
abline(lsfit(y,fitted(m7.5)), col='red')

#####Residuos#####
Endereco = "E:\\"

k=1/exp(coef(m7.5,"sigma"))

xb<- m7.5$mu.coefficients[1]+
  m7.5$mu.coefficients[2]*x2+
  m7.5$mu.coefficients[3]*x7+
  m7.5$mu.coefficients[4]*x8+
  m7.5$mu.coefficients[5]*x10+
  m7.5$mu.coefficients[6]*x13+
  m7.5$mu.coefficients[7]*x16+
  m7.5$mu.coefficients[8]*x17

g<- k
gama<-1/g
gama
tempo<-dados$y
soest<- exp(-(tempo/(xb)))
soest
venda=dados$venda
ei<- (y*exp(-xb))^gama              # resíduos de Cox-Snell
ekmr<-survfit(Surv(ei,venda)~1)
t<-ekmr$time
st<-ekmr$surv
sexp<-exp(-t)
pdf(file = paste(Endereco,"cox3.pdf"), width = 10, height = 10)
plot(st,sexp,xlab="S(ei): Kaplan-Meier",
ylab="S(ei): Exponencial padrão",pch=16,cex.lab=1.7)
dev.off()
pdf(file = paste(Endereco,"expadrao2.pdf"), width = 10, height = 10)
```

```
plot(ekmr,conf.int=F,mark.time=F, xlab="Resíduos de Cox-Snell",
 ylab="Sobrevivência estimada",cex.lab=1.7,lwd=3)
lines(t,sexp,lty=4,col="red",lwd=3)
legend(1.0,0.8,lty=c(1,4),lwd=c(3,3),
c("Kaplan-Meier","Exponencial Padrão"),cex=1.7,bty="n",
col=c("black","red"))
dev.off()

###Sobrevivencia estimada###

k=1/exp(coef(m7.5,"sigma"))

g<- k
gama<-1/g
gama
tempo<-dados$y
venda=dados$venda
my.surv <- Surv(tempo, venda)
ekmf <- survfit(my.surv~1, conf.int=0.95)
time<- ekmf$time
st <- ekmf$surv
stw<- exp(-(tempo/exp(xb))^gama)
ord1 = order(tempo)
ord2 = order(stw, decreasing=TRUE)
pdf(file = paste(Endereco,"soes.pdf"), width = 10, height = 10)
plot(ekmf, conf.int=F, xlab="Valores",main="Sobrevivência Estimada",
ylab="S(v)", cex.lab=1.7, cex.main=1.7,lwd=3)
lines(c(0,tempo[ord1]), c(1,stw[ord2]), lty=2, col="red", lwd=3)
legend("topright", lty=c(1,2), c("Kaplan-Meier","Weibull"),
bty="n",cex=1.7, col=c("black","red"))
dev.off()
```

Referências Bibliográficas

AGUIRRE, A. & MACEDO, P.B.R. Estimativas de Preços Hedônicos para o Mercado Imobiliário de Belo Horizonte. *Anais do XVIII Encontro Brasileiro de Econometria 1*, 1-16. Águas de Lindóia-SP, 1996.

AKAIKE, H. A new look at the statistical model identification. *IEEE Transactions on Automatic Control* 19, 716-723, 1974.

ANGLIN, P. & GENCAY, R. Semiparametric estimation of hedonic price function. *Journal of Applied Econometrics*, vol. 11, p.633-648, 1996.

ANGELO, C. F. de and FÁVERO, L. P. L. Modelo de preços hedônicos para a avaliação de veículos novos, *VI Seminário em Administração*-FEA-USP, São Paulo, 2003.

ASSOCIAÇÃO BRASILEIRA DE NORMAS TÉCNICAS. *Avaliação de imóveis urbanos (NBR 5676 e NBR 502).* Rio de Janeiro, 1989

BARBOSA, E.P. & BIDURIN, C.P. Seleção de modelos de regressão para predição via validação cruzada: uma aplicação na avaliação de imóveis. *Revista Brasileira de Estatística 52*, 105-120, 1991.

BECKMAN, R. J., NACHTSHEIM, C. J. & COOK, R. D. Diagnostics for mixed-model analysis of variance. *Technometrics* 29, 413-26, 1987.
BERGER, J.O. *Statistical Decision Theory and Bayesian Analysis*. New York: Springer, 1985.

BIDERMAN, C. *Forças de atração e expulsão na Grande São Paulo*. São Paulo: Fundação Getúlio Vargas, 2001.

CALSAVARA, V. F.; *M*odelos de Sobrevivência com Fração de Cura usando um Termo de Fragilidade e Tempo de Vida Weibull Modificada Generalizada. 2011. 82 p. Dissertação (Mestrado em Estatística)- Universidade Federal de São Carlos (UFSCar), São Carlos, 2011.

CASELLA, G & BERGER, R. *Statistical Inference*. Pacific Grove:Duxbury Press, second edition, 2002.

CHAMBERS, J. M. and HASTIE, T. J. *Statistical Models in S*, Chapman and Hall, London, 1991.

CLEVELAND, W.S.; GROSSE, E. & SHYU, M.J. *Local regression models. In Statistical Modelling in S*. Eds: Chambers, J.M. and Hastie, T.J., 309-376. New York: Chapman and Hall, 1992.

COLE, T.J. & GREEN, P.J. Smoothing reference centile curves: the LMS method and penalized likelihood. *Statistics in Medicine 11*, 1305-1319, 1992.

COLLETT, D. *Modelling Survival Data in Medical Research*. New York: Chapman and 5 Hall, 1994.

COLOSIMO, E. A.; GIOLO, S. R. *Análise de Sobrevivência Aplicada*. São Paulo, SP: 7 Edgard Blücher, 2006, 369 p.

COOK, R. D. Assessment of local influence (with discussion). *J. R. Statist. Soc. B*, 48, 133-169, 1986.

CORDEIRO, G. M., DEMÉTRIO, C. G. B. Modelos Lineares Generalizados. *Minicurso: 12o SEAGRO e a 52a Reunião Anual da RBRAS*, UFSM, Santa Maria, RS, 2008.

Cremasco, C. P. *M*odelagem de Dados de Sobrevivência via Modelos de Risco Logístico Generalizado. 2005. 80 p. Dissertação (Mestrado em Estatística)- Universidade Federal de São Carlos (UFSCar), São Carlos, 2005.

CUNHA, M.C. *Métodos Numéricos*, 2a ed. São Paulo: Unicamp, 2000.

DANTAS, R. A.; & CORDEIRO, G.M. Uma nova metodologia para avaliação de imóveis utilizando modelos lineares generalizados. *Revista Brasileira de Estatística* , v.49, n.191, p. 27-46, 1988.

DANTAS, R.A. & CORDEIRO G.M. Evaluation of the Brazilian city of Recife's condominium market using generalized linear models. *The Appraisal Journal 69*, 247-257, 2001.

DANTAS,R. A. *Modelos Espaciais Aplicados ao Mercado Habitacional: Um Estudo de Caso Para a Cidade do Recife.* 2003. 114 f.Tese (Doutorado em Economia - Área de concentração: Métodos quantitativos) - Universidade Federal de Pernambuco (UFPE), Recife. 2003.

DANTAS, R. A., *Engenharia de Avaliações: Uma Introdução à Metodologia Científica*. 2ª Ed. São Paulo: PINI. 2005.

DEMÉTRIO, C.G.B. e ZOCCHI, S.S. *Modelos de Regressão na Experimentação Agronômica*. Apostila. Departamento de Ciências Exatas, ESALQ/USP, 2007.

DEMÉTRIO, C.G.B. *Modelos Lineares Generalizados em Experimentação Agronômica*. Apostila. Departamento de Ciências Exatas, ESALQ/USP, 2003.

DUNN KP, SMYTH GK. Randomized quantile residuals. *J Comput Graph Stat* 5(1-10): 236-244, 1996.

FERRAUDO, G. M. *I*nferência do valor de mercado de lotes urbanos. Estudo de caso: Município de São Carlos (SP) . 2008. 104 p. Dissertação (Mestrado em Estatística)- Universidade Federal de São Carlos (UFSCar), São Carlos, 2008.

FERRAUDO, G. M. ; LOUZADA NETO, F. ; FERREIRA, J.F. Determinação do valor de Mercado de lotes urbanos: estudo de um caso - município de São Carlos. *Revista Brasileira de Biometria*, São Paulo, Brasil., v.28, p.52-65, 2010.

FERREIRA, J. F. *P*roposta de Tratamento da Variável Localização em Modelos Inferenciais de Avaliação Imobiliária para Municípios Médios . 2007. 157 p. Dissertação (Mestrado em Engenharia Urbana)- Universidade Federal de São Carlos (UFSCar), São Carlos, 2007.

FLORENCIO, L. A. Modelos Espaciais para Avaliação de Imóveis em Massa com Base em Modelos de Regressão GAMLSS. *12 Conferência Internacional da LARES.* São Paulo, Brasil, 2012.

FLORENCIO, L. A. *E*ngenharia de Avaliações com Base em Modelos GAMLSS. 2010. 125f. Dissertação (Mestrado em Estatística) - Departamento de Estatística, Universidade Federal de Pernambuco, Pernambuco. 2010.

FUNG, W. K. e KwWAN, C. W. A note on local influence based on normal curvature.*Journal of the Royal Statistical Society* B 59, 839- 843, 1997.

GENÇAY, R.; YANG, X. A forecast comparison of residential housing prices by parametric versus semiparametric conditional mean estimators. *Economics Letters*, 52,p 129-135, 1996.

GONZÁLEZ, M. A. S. e FORMOSO, C. T. Análise conceitual das dificuldades na determinação de modelos de formação de preços através de análise de regressão.*Engenharia Civil - UM*,8,65-75, 2000

GONZÁLEZ, M. A. S. *A Engenharia de Avaliações na Visão Inferencial*. São Leopoldo -RS: Editora UNISINOS, 1997. 142p.

HÄRDLE, W. *Applied Nonparametric Regression*. Cambridge: Cambridge University Press,1990.

HASTIE, T.J. & TIBSHIRANI, R.J. *Generalized Additive Models*. London: Chapman & Hall,1990.

HASTIE, T. & TIBSHIRANI, R. Varying-coefficient models (with discussion). *Journal of the Royal Statistical Society* B 55, 757-796, 1993.

KALBFLEISCH, J. D. e PRENTICE, R. L. *The Statistical Analysis of Failure Time Data*. New York : John Wiley and Sons, 1980.

KAPLAN, E. & MEIER, P. Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, 53(282), 457–481, 1958.

Lancaster, K.J. A new approach to consumer theory.*Journal of Political Economy*,74, 132-157, 1966.

LAWLESS, J. F. *Statistical Models and Methods for Lifetime Data*. New York: John Wiley and Sons, 1982.

LAWRENCE, A. J. Local and Deletion Influence. *IMA Preprint*, 731,1067-1077, 1990.

LIMA, L.P; ANDRÉ, C.D.S & SINGER, J.M. Modelos aditivos generalizados: metodologia e prática. *Revista Brasileira de Estatística 62*, 37-69, 2001.

LOUZADA-NETO, F.; MAZUCHELI, J.; ACHCAR, J. A. *Análise de Sobrevivência e Confiabilidade*. Monografias del IMCA, 30, Peru, 2002.

MARTINS-FILHO, C. & BIN, O. Estimation of hedonic price functions via additive nonparametric regression. *Empirical Economics 30*, 93-114, 2003.

NELDER, J. A., WEDDERBURN, R. W. M. Generalized Linear Models. *Journal of the Royal Statistical Society A* , Vol. 135, No. 3, pp. 370-384, 1972.

NETO, A. P. *R*edes Neurais Artificiais Aplicadas às Avaliações em Massa. Estudo de Caso para a Cidade de Belo Horizonte/ MG . 2006. 96 p. Dissertação (Mestrado em Engenharia Elétrica - Área de Concentração: Engenharia de Computação e Telecomunicações)-Universidade Federal de Minas Gerais, Belo Horizonte, 2006.

PAIVA, C. S. M., FREIRE, D. M. C. e CECATTI, J. G. Modelos Aditivos Generalizados para Posição, Escala e Forma (GAMLSS) na Modelagem de Curvas de Referência.*Rev. Brasileira de Ciências da Saúde*, vol. 12,n. 3, p. 289-310, 2008

PAULA, G. A., *Modelos de Regressão com Apoio Computacional*. São Paulo: IME/USP, 2004. 392p.

POON, W-Y and POON, YAT SUN. Conformal normal curvature and assessment of local influence. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61, 51-61, 1999.

R Core Team (2012). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

RIGBY, R. A.; STASINOPOULOS, D. M. Generalized additive models for location, scale and shape. *Appl. Statist*, 54, Part 3, p. 507-554, 2005

RIGBY, R. A.; STASINOPOULOS, D. M. Generalized Additive Models for Location, Scale and Shape (GAMLSS) in R. *Journal of Statistical Software*, v.23, Issue 7, 2007

RIGBY, R. A., STASINOPOULOS, D. M. and AKANTZILIOTOU, C (2008). Instructions on how to use the gamlss package in R -Second Edition. Available from $< http : //www.gamlss.org/wp-content/uploads/2013/01/gamlss-manual.pdf >$ access on 02/08/2013

RIGBY, R. A. and STASINOPOULOS, D. M. (2008). A exible regression approach using GAMLSS in R. Available from $< http : //www.gamlss.org/wp-content/uploads/2013/01/book-2008-27-6-08.pdf >$ acess on 02/08/2013

ROSEN, S. Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition *The Journal of Political Economy*, v.82, n.1,p. 34-55, 1974

ROYSTON, P. & ALTMAN, D.G. Regression using fractional polynomials of continuous covariates: parsimonious parametric modelling (with discussion). *Applied Statistics 43*, 429-467, 1994.

SÃO CARLOS. Site da Prefeitura Municipal. Disponível em $< http : //www.saocarlos.sp.gov.br/ >$. Acesso em 02 de agosto de 2013.

SCHWARTZ, G. Estimating the dimension of a model. *Annals of Statistics* 6, 461-464, 1978.

SEBER, G. A. F., *Linear Regression Analysis*. USA: Editora John Wiley & Sons, 2003. 582p.

TERRA, M. L. C. *M*odelos Lineares Generalizados Simétricos Heterocedásticos. 2009. 60 p. Dissertação (Mestrado em Estatística)- Universidade Federal de Pernambuco, Recife, 2009.

THOMAS, W. and COOK, R. D. Assessing influence on regression coefficients in generalized linear models. *Biometrika*, 76, 4, p. 741-749, 1989

TURKMAN, M. A. A. e SILVA, G. L. *Modelos Lineares Generalizados - da teoria à prática*. LISBOA, 2000. 151p.

VAN BUUREN & S. FREDRIKS, M. Worm plot: a simple diagnostic device for modelling growth reference curves. *Statistics in Medicine*, 20, 1259-1277, 2001.

VENABLES, W.N. & RIPLEY, B.D. *Modern Applied Statistics with S*. 4th ed. Springer, 2002.

VIEIRA, G. E. Uma revisão sobre a aplicação de simulação computacional em processos industriais. *Simpósio de Engenharia de Produção*, XIII, Bauru, Anais, 2006

WANG, S. et al.Local Influence Analysis of Generalized Linear Model. *Applied Mathematics*, 3, 1065-1067, 2012.

WEIBULL, W. A Statistical Theory of the Strength of Materials. *Ingeniors Vetenskaps Akademien Handlingar*, 151: The Phenomenon of Rupture in Solid, 293-297, 1939.