

**UNIVERSIDADE FEDERAL DE SÃO CARLOS  
CENTRO DE CIÊNCIAS EXATAS E TECNOLÓGICAS  
DEPARTAMENTO DE ESTATÍSTICA  
MESTRADO EM ESTATÍSTICA**

**INFERÊNCIA DO VALOR DE MERCADO DE LOTES URBANOS.  
ESTUDO DE CASO: MUNICÍPIO DE SÃO CARLOS (SP).**

**SÃO CARLOS  
Novembro - 2008**

**INFERÊNCIA DO VALOR DE MERCADO DE LOTES URBANOS.  
ESTUDO DE CASO: MUNICÍPIO DE SÃO CARLOS (SP).**

**INFERÊNCIA DO VALOR DE MERCADO DE LOTES URBANOS.  
ESTUDO DE CASO: MUNICÍPIO DE SÃO CARLOS (SP).**

Guilherme Moraes Ferraudó

Orientador: 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.

**SÃO CARLOS  
Novembro - 2008**

**Ficha catalográfica elaborada pelo DePT da  
Biblioteca Comunitária da UFSCar**

F381iv

Ferraudo, Guilherme Moraes.

Inferência do valor de mercado de lotes urbanos. Estudo de caso : município de São Carlos (SP) / Guilherme Moraes Ferraudo. -- São Carlos : UFSCar, 2009.

131 f.

Dissertação (Mestrado) -- Universidade Federal de São Carlos, 2008.

1. Análise de sobrevivência. 2. Modelos lineares (Estatística). 3. Probabilidade de Cobertura. 4. Loteamento. 5. Localização residencial. I. Título.

CDD: 519.5 (20<sup>a</sup>)



UNIVERSIDADE FEDERAL DE SÃO CARLOS

Centro de Ciências Exatas e de Tecnologia

Programa de Pós-Graduação em Estatística

Via Washington Luis, Km 235 - C.P.676 - CGC 45358058/0001-40

FONE: (016) 3351-8292/3351-8241 - FAX: (016) 3351-8243

13565-905 - SÃO CARLOS-SP-BRASIL

[www.ufscar.br/~des](http://www.ufscar.br/~des)

[ppgest@power.ufscar.br](mailto:ppgest@power.ufscar.br)



## DECLARAÇÃO

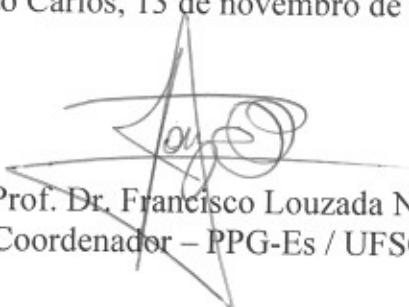
Declaramos, para os devidos fins, que Guilherme Moraes Ferraudo defendeu sua Dissertação de Mestrado no dia 13/11/2008, tendo sido **aprovado**. O aluno deverá apresentar a versão final da dissertação (com as correções e sugestões da Banca, e a ficha catalográfica anexada), e a Certidão Negativa da Biblioteca Comunitária, para formação do processo de homologação e emissão do Diploma do Título.

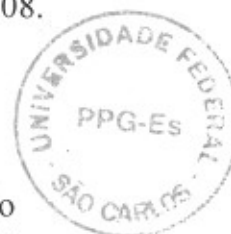
Igualmente, o aluno deverá apresentar a documentação da pesquisa (rotinas, arquivos em LaTeX, resultados complementares etc.) ao seu orientador, visando facilitar a confecção de relatórios técnicos que condensarão os resultados obtidos.

Toda a documentação citada deverá ser entregue em 30 (trinta) dias, a iniciar na data da Defesa.

Essa declaração é válida pelo período de 30 dias.

São Carlos, 13 de novembro de 2008.

  
Prof. Dr. Francisco Louzada Neto  
Coordenador – PPG-Es / UFSCar



*“Àquele que entendendo e respeitando a variabilidade consegue construir uma realidade melhor.”*

## **AGRADECIMENTOS**

Agradeço a Deus pela luz interminável em meu caminho, à minha família pelo incansável apoio material e espiritual e por me mostrar o quanto é essencial, além de importante, ser uma pessoa repleta de ideais e valores. A minha família (Pai, Mãe, Tati e Dani) juntamente com os meus amigos (O RANCA) provam a cada dia que o que precisamos é curtir mais a viagem e não se preocupar tanto com o destino final.

Agradeço à Universidade Federal de São Carlos, em especial, ao Departamento de Estatística por ter me proporcionado um ensino superior de qualidade. Agradeço aos professores e funcionários do Departamento de Estatística pelo respeito e dedicação desde os tempos de graduação. Gostaria de dedicar um carinho especial aos professores com quem realizei trabalhos durante a minha breve jornada no Departamento de Estatística, são eles: Prof. Dr. Benedito G. Benze, Prof<sup>a</sup> Dr<sup>a</sup> Vera L. D. Tomazella, Prof. Dr. Lael A. Oliveira e o grande mentor e orientador do meu mestrado, Prof. Dr. Francisco Louzada Neto.

Neto, a você serei grato eternamente. Jamais esquecerei a sua incansável dedicação desde a minha iniciação científica até o final deste mestrado para me ensinar estatística e assuntos comuns da vida de um homem. Para mim, um simples aprendiz, resta carregar os ensinamentos do mestre e, em algum momento, poder transmiti-los a alguém que necessite, assim como necessito. Neste momento, peço desculpas pelos meus deslizes e sinto que poderia ter me dedicado mais a este trabalho.

Agradeço, em especial, ao José Fabrício Ferreira pela incansável dedicação à nossa parceria que vem acontecendo desde o ano de 2004 e agora produz seus frutos, as nossas dissertações de mestrado. Através do Fabrício agradeço à Prefeitura Municipal de São Carlos, SP.

Agradeço, também, à turma do Mestrado 2007 e do Bacharelado em Estatística 2002 pelo incomensurável apoio nos créditos realizados. Nunca esquecerei os momentos (Rep. Kabanacana, churrascos, futebol, café da tarde, almoço, festas, horas e mais horas estudando, etc.) vividos em São Carlos. Obrigado a todos!



*“...Só feche o seu livro quem já aprendeu... (Taiguara)”*

## RESUMO

Nesta dissertação apresentaremos uma proposta de um modelo de equação de regressão representativa para a formação do valor de mercado dos lotes urbanos do município de São Carlos, SP, ano de 2005, visando à criação de Plantas de Valores Genéricos (PVG) utilizando as técnicas de: Modelos Lineares Usuais (erros normais e variância constante), estes amplamente utilizados, e a Análise de Sobrevivência com censura à esquerda. Após o ajuste, as duas metodologias são comparadas e testadas num estudo de simulação onde examinamos a probabilidade de cobertura de alguns parâmetros envolvidos na regressão.

**Palavras-chave:** Análise de Sobrevivência com Censura à Esquerda; Modelos Lineares Usuais; Probabilidade de Cobertura; Planta de Valores Genéricos; Localização Intra-Urbana.

## ABSTRACT

In this dissertation we present a regression modelling proposal for modelling the market prices of urban batches at São Carlos city (SP), over the year of 2005. Usual regression modelling and survival techniques, with left censoring, are considered. A simulation study examines the coverage probabilities the asymptotic confidence for the parameters of the considered modelling.

**Keywords:** Survival Analysis with Left Censoring; Regression Modelling; Coverage Probability; Generic table of values; Urban intra-localization.

## LISTA DE ILUSTRAÇÕES

Figura 1. Configuração planimétrica da sede do município de São Carlos, SP, com indicação das maiores barreiras intra-urbanas. Fonte: Ferreira (2007).....	28
Figura 2. Mapa da classificação dos setores censitários segundo o valor da distância estatística a um dos pólos sócio-econômicos opostos. Fonte: Ferreira (2007).....	29
Figura 3. Delimitação de regiões homogêneas com base na sobreposição da abrangência territorial das variáveis selecionadas. Fonte: Ferreira (2007). ....	32
Figura 4. Limite inferior nominal, limite superior nominal e valor central nominal (95%) dos intervalos assintóticos normais das probabilidades de cobertura nominais das estimativas dos parâmetros para os 5 tamanhos diferentes de amostra e 3 diferentes variâncias. ....	58
Figura 5. Discrepâncias das estimativas dos parâmetros para os 5 tamanhos diferentes de amostra e 3 diferentes variâncias.....	60
Figura 6. Gráficos de diagnóstico para o modelo normal com transformação raiz quadrada para a variável resposta.....	63
Figura 7. Valores observados versus os valores preditos elevados ao quadrado para o modelo normal usual.....	65
Figura 8. Gráfico mostrando o mecanismo da censura à esquerda no valor do lote urbano.....	67
Figura 9. (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. ....	80
Figura 10. Limite inferior nominal, limite superior nominal e valor central nominal (95%) dos intervalos assintóticos normais das probabilidades de cobertura nominais das estimativas dos parâmetros para os 5 tamanhos diferentes de amostra e 3 diferentes porcentagens de censuras (0%, 1% e 5%).....	85
Figura 11. Limite inferior nominal, limite superior nominal e valor central nominal (95%) dos intervalos assintóticos normais das probabilidades de cobertura nominais das estimativas dos parâmetros para os 5 tamanhos diferentes de amostra e 3 diferentes porcentagens de censuras (15%, 30% e 60%).....	87
Figura 12. Discrepâncias das estimativas dos parâmetros para os 5 tamanhos diferentes de amostra e 3 diferentes porcentagens de censuras (0%, 1% e 5%). ....	89
Figura 13. Discrepâncias das estimativas dos parâmetros para os 5 tamanhos diferentes de amostra e 3 diferentes porcentagens de censuras (15%, 30% e 60%). ....	91
Figura 14. (a) Curva de permanência à venda estimada pelo método de Kaplan-Meier, o p-valor da estatística <i>logrank</i> e o risco relativo para a covariável Ferrovia, (b) Risco acumulado de venda empírico considerando a covariável Ferrovia.....	93
Figura 15. TTT-Plot para os valores unitários dos lotes de São Carlos, SP, no ano de 2005. ....	93

## LISTA DE TABELAS

Tabela 1. Descrição das variáveis dicotômicas indicadoras da localização.....	33
Tabela 2. Tabela de Análise de Variância. ....	42
Tabela 3. Probabilidades de cobertura nominais das estimativas dos parâmetros para os 5 tamanhos diferentes de amostra. ....	55
Tabela 4. Discrepâncias das estimativas dos parâmetros para os 5 tamanhos diferentes de amostra. ....	56
Tabela 5. Estimativa dos parâmetros, respectivos limites: inferior e superior, do intervalo de confiança de 95% e amplitude do intervalo, para cada covariável. ....	63
Tabela 6. Probabilidades de cobertura nominais das estimativas dos parâmetros para os 5 tamanhos diferentes de amostra. ....	81
Tabela 7. Discrepâncias das estimativas dos parâmetros para os 5 tamanhos diferentes de amostra. ....	82
Tabela 8. Seleção de covariáveis usando o modelo Weibull considerando censura à esquerda. ....	94
Tabela 9. Estimativa dos parâmetros, respectivos limites inferior e superior, do intervalo de confiança de 95% e amplitude do intervalo, para cada covariável. ....	95
Tabela 10. Valor estimado, considerando os dois modelos construídos anteriormente, para dois lotes urbanos em condições distintas de negociação. ....	98

## SUMÁRIO

<b>RESUMO .....</b>	<b>10</b>
<b>ABSTRACT .....</b>	<b>11</b>
<b>LISTA DE ILUSTRAÇÕES.....</b>	<b>12</b>
<b>LISTA DE TABELAS .....</b>	<b>13</b>
<b>1. INTRODUÇÃO .....</b>	<b>16</b>
<b>1.1. PROBLEMA.....</b>	<b>17</b>
<b>1.2. OBJETIVOS.....</b>	<b>21</b>
1.2.1. Objetivo geral.....	21
1.2.2. Objetivo específico.....	22
<b>1.3. JUSTIFICATIVA.....</b>	<b>22</b>
<b>1.4. ESTRUTURA DO TRABALHO .....</b>	<b>22</b>
<b>2. REVISÃO BIBLIOGRÁFICA .....</b>	<b>24</b>
<b>2.1. PLANTAS DE VALORES GENÉRICOS .....</b>	<b>24</b>
<b>2.2. A ÁREA DE ESTUDO .....</b>	<b>24</b>
<b>2.3. CONFIGURAÇÃO ESPACIAL DE SÃO CARLOS, SP, E     CONSIDERAÇÕES SOBRE A MODELAGEM DA VARIÁVEL LOCALIZAÇÃO</b>	<b>25</b>
<b>2.4. VARIÁVEIS UTILIZADAS.....</b>	<b>29</b>
<b>3. MODELOS DE REGRESSÃO.....</b>	<b>34</b>
<b>4. MODELOS LINEARES USUAIS.....</b>	<b>35</b>
<b>4.1. ESTIMAÇÃO .....</b>	<b>35</b>
<b>4.2. SOMAS DE QUADRADOS .....</b>	<b>36</b>
<b>4.3. PROPRIEDADES DO EMQ E DOS RESÍDUOS.....</b>	<b>37</b>
<b>4.4. MODELO NORMAL-LINEAR.....</b>	<b>39</b>
<b>4.5. ANÁLISE DE VARIÂNCIA .....</b>	<b>40</b>
<b>4.6. SELEÇÃO DE VARIÁVEIS EXPLICATIVAS.....</b>	<b>42</b>
<b>4.7. INTERVALOS DE CONFIANÇA.....</b>	<b>43</b>
<b>4.8. SELEÇÃO DE MODELOS.....</b>	<b>43</b>
<b>4.9. TÉCNICAS DE DIAGNÓSTICO .....</b>	<b>45</b>
4.9.1. Resíduos.....	47
4.9.2. Influência .....	48
4.9.3. Técnicas gráficas .....	50
<b>4.10. TRANSFORMAÇÃO BOX-COX.....</b>	<b>51</b>
<b>5. MODELAGEM ATRAVÉS DO MODELO LINEAR USUAL .....</b>	<b>53</b>
<b>5.1. INTERVALOS DE CONFIANÇA.....</b>	<b>53</b>
<b>5.2. ESTUDO DE SIMULAÇÃO PARA O MODELO DE REGRESSÃO LINEAR     USUAL .....</b>	<b>54</b>
<b>5.3. APLICAÇÃO DO MODELO LINEAR USUAL .....</b>	<b>61</b>
<b>6. ANÁLISE DE SOBREVIVÊNCIA .....</b>	<b>66</b>
<b>6.1. PARTICULARIDADES DA ANÁLISE DE SOBREVIVÊNCIA .....</b>	<b>67</b>
<b>6.2. DESCRIÇÃO DO COMPORTAMENTO DO VALOR DO IMÓVEL .....</b>	<b>69</b>
6.2.1. A função densidade de probabilidade.....	69
6.2.2. A função de sobrevivência ou função de permanência à venda.....	69
6.2.3. A função de risco.....	70
<b>6.3. A IMPORTÂNCIA DA FUNÇÃO DE RISCO .....</b>	<b>70</b>
<b>6.4. PROCEDIMENTOS NÃO-PARAMÉTRICOS .....</b>	<b>71</b>
6.4.1. O estimador de Kaplan-Meier.....	72
<b>6.5. COMPARAÇÃO DE DUAS FUNÇÕES DE SOBREVIVÊNCIA .....</b>	<b>73</b>
<b>6.6. PROCEDIMENTOS PARAMÉTRICOS.....</b>	<b>74</b>

6.6.1. Determinação empírica da forma da função de risco .....	74
<b>6.7. DISTRIBUIÇÃO WEIBULL .....</b>	<b>75</b>
6.7.1. Regressão Weibull.....	76
6.7.2. Estratégia para a seleção de covariáveis .....	77
<b>6.8. TESTES DE HIPÓTESES .....</b>	<b>78</b>
6.8.1. Teste da razão de verossimilhanças.....	79
<b>7. RESULTADOS DE ANÁLISE DE SOBREVIVÊNCIA .....</b>	<b>80</b>
7.1. ANÁLISE DESCRITIVA E EXPLORATÓRIA.....	80
7.2. ESTUDO DE SIMULAÇÃO PARA O MODELO DE REGRESSÃO WEIBULL .....	80
7.3. AJUSTE DE UM MODELO DE REGRESSÃO PARAMÉTRICO.....	91
<b>8. CONCLUSÃO.....</b>	<b>97</b>
<b>REFERÊNCIAS BIBLIOGRÁFICAS .....</b>	<b>99</b>
<b>APÊNDICE A – Código para a modelagem dos dados .....</b>	<b>104</b>
<b>APÊNDICE B – Código para o estudo de simulação.....</b>	<b>113</b>
<b>APÊNDICE C – Código para gerar TTT-Plot no SAS.....</b>	<b>131</b>

## 1. INTRODUÇÃO

A Localização é um dos principais componentes na formação do valor de mercado do imóvel urbano. Uma das aplicações mais relevantes da apuração destes valores é a Planta de Valores Genéricos (PVG).

A PVG corresponde a generalizações, inferidas através de amostras pontuais de eventos observáveis no mercado imobiliário.

Ao elaborar uma PVG baseada em uma metodologia científica, nesse caso, a metodologia estatística, pretende-se que o gestor (avaliador) se norteie pelo princípio da imparcialidade, princípio este que deve nortear toda a decisão envolvendo os interesses dos cidadãos. A imparcialidade está diretamente associada à fundamentação da metodologia estatística, ou seja, no final, ao questionar a PVG estaremos, indiretamente, questionando a fundamentação da metodologia estatística empregada, desde o plano amostral até a inferência dos valores. Existe o outro lado da imparcialidade, a parte subjetiva, a parte intrínseca ao ser humano e nessa nenhuma metodologia jamais conseguirá influenciar e, portanto, não discutiremos no atual trabalho. Infelizmente, a maioria das decisões de ordem pública é influenciada por essa subjetividade.

A análise de sobrevivência, conforme Colosimo e Giolo (2006) “é uma das áreas da estatística que mais evoluiu nas últimas décadas do século passado”. A análise de sobrevivência, além de permitir a construção de um modelo para prever o valor de mercado do lote urbano, como o modelo de regressão linear usual, permite estimar, para um determinado lote, a probabilidade que ele seja vendido e, também, o risco de venda associado a ele.

O presente trabalho foi desenvolvido em conjunto com a dissertação de Ferreira (2007) defendida em dezembro de 2007 no Mestrado em Engenharia Urbana, Departamento de Engenharia Civil da UFSCar. Ferreira (2007) discute o tratamento da variável “Localização” para modelos estatísticos em avaliação imobiliária enquanto que este trabalho discute as particularidades de tais modelos estatísticos propondo uma metodologia que auxilie nas relações espaço/sociedade, em especial, na cidade de São Carlos, SP.



## 1.1. PROBLEMA

No Brasil, a avaliação em massa de imóveis é de grande importância para os governos municipais. Avaliações em massa são aplicadas para a elaboração de Plantas de Valores Genéricos (PVG), cujo objetivo principal é estabelecer a base de cálculo para a cobrança de tributos municipais, tais como, o Imposto Predial e Territorial Urbano (IPTU), Imposto sobre a Transmissão de Bens Imóveis (ITBI) e as Contribuições de Melhoria. A PVG é a representação do valor unitário do solo, normalmente expresso em Reais por área (R\$/m<sup>2</sup>).

A Norma Brasileira ABNT (2004) distingue duas metodologias básicas no tratamento dos dados: tratamento por fatores de homogeneização e a metodologia científica, com uso de regressão linear. O tratamento científico é, segundo a ABNT (2004), “o tratamento de evidências empíricas pelo uso de metodologia científica que leve à indução de modelo validado para o comportamento do mercado”. A maioria dos modelos obtidos pela metodologia científica utiliza a análise de regressão linear, como previsto em norma, porém “outras ferramentas analíticas para a indução do comportamento do mercado” são permitidas, desde que fundamentadas. Atualmente as redes neurais artificiais, a regressão espacial, a análise envoltória de dados (Charnes *et al.*, 1995) e, neste trabalho, devido à natureza da variável resposta (valores positivos), propõe uma modelagem baseada em Análise de Sobrevivência.

O processo no qual explicitamos a variável localização, ou, mais especificamente, a localização intra-urbana é bastante subjetivo fazendo com que esse processo se torne questionável.

*A questão da amostragem também se torna relevante. Os trabalhos de avaliação em massa de imóveis e de criação da PVG impõem uma duração na qual nem sempre é possível obter uma quantidade e equilíbrio espacial satisfatório na coleta de amostras. Quando isto ocorre, coloca-se ao avaliador o problema sobre como inferir os valores do solo em locais com difícil obtenção de amostras, com base na comparação com localidades semelhantes. Quais seriam os parâmetros de comparação entre duas localizações distintas? Revela-se a necessidade de construir condições de contorno para aplicação das variáveis indicadoras da localização em modelos inferenciais de avaliação imobiliária. Pressupõem-se que estas condições de contorno possam ser aplicáveis ao universo de contextos urbanos brasileiros, especialmente os não-metropolitanos (Ferreira, 2007).*

Visando eliminar esta subjetividade, a presente pesquisa colaborou para desenvolver

um método para o delineamento da variável localização, descrito mais detalhadamente em Ferreira (2007), que reúne os referenciais teóricos na área do Planejamento Urbano. A presente dissertação, por outro lado, avança em relação ao citado estudo, na medida em que o desenvolve do ponto de vista estatístico, utilizando o método da Análise de Sobrevida. Justifica-se, portanto, o elevado número de citações a este trabalho, que são necessárias para a contextualização da abrangência do tema e as motivações de ambos os trabalhos.

Um modelo estatístico pode apresentar crescentes níveis de complexidade, devido ao tipo e quantidade de variáveis envolvidas.

Dentre as dezenas de potenciais variáveis indicadoras de localização, foram selecionadas as de maior abrangência geográfica, em detrimento de variáveis “micro”, de alcance territorial reduzido. Estas variáveis selecionadas são do tipo binário, por representar dicotomias características de contextos urbanos brasileiros. A partir deste conjunto de variáveis é possível avaliar a influência das variáveis “micro”, bem como estabelecer bases de comparação entre municípios distintos, assuntos que ultrapassam os limites do presente trabalho e abordados em Ferreira (2007).

*Os esforços foram direcionados para a obtenção do maior número possível de amostras referentes às transações imobiliárias efetivamente realizadas. No entanto, estes dados são os mais difíceis de serem obtidos. As partes envolvidas no negócio normalmente não revelam o valor real transacionado. Mesmo havendo fidelidade das fontes quanto ao preço, é comum os negócios se realizarem através de financiamentos. As modalidades de juros interferem na observação do valor à vista do bem, o que demandaria a aplicação de índices financeiros e monetários, que representam inconvenientes para obtenção de um preço padronizado. Portanto, a pesquisa se concentra nos preços de oferta dos lotes, obtidos nas empresas imobiliárias locais. A escolha por este valor se justifica por ser fiel, sintético e de fácil obtenção pelo pesquisador. Embora seja de certa forma tendenciosa, por expressar puramente os interesses dos vendedores, é lícito crer que este valor represente o máximo valor possível sobre uma determinada parcela de terra urbana. Associadas a este preço de oferta, seguem as vantagens oferecidas pela aquisição do imóvel: vizinhança, acessibilidade, atrativos naturais, acumulação de infra-estruturas, dentre outras (Ferreira, 2007).*

Os preços de oferta também carregam distorções importantes. Uma delas diz respeito à “estratégia de negociação” do vendedor, enquanto alguns estabelecem o preço real pelo qual desejam vender e mantêm este preço, outros estabelecem um preço acima para dar margem à negociação. Outra distorção diz respeito à “condição do vendedor”, enquanto alguns aceitam

vender por um preço sabidamente baixo, sejam por estarem necessitando do dinheiro ou para encerrar disputas judiciais (entre herdeiros, sócios, casais, etc.) e outros podem impor preços sabidamente exagerados por estarem em condições confortáveis. Ofertas altas e baixas, geralmente, não se anulam, seu comportamento é bastante assimétrico, dado o “efeito manada” que muitas vezes se observa nos mercados em geral e, em especial, no imobiliário.

*Quando nos referimos ao termo “solo urbano” de forma genérica, o que está em questão é a própria localização intra-urbana, ou “localização pura”. A Localização intra-urbana, ou “ponto”, como é popularmente conhecido, é um produto humano por essência. No entanto, embora seja quase intangível, por ser imaterial, o valor da localização relaciona-se diretamente ao seu potencial de uso (Ferreira, 2007).*

*As aglomerações urbanas são compostas por sistemas de localizações. O espaço intra-urbano representa o conjunto dos deslocamentos diários de pessoas em suas atividades diárias, do tipo casa-trabalho, casa-lazer, casa-escola, casa-consumo, dentre outros. Este espaço intra-urbano constitui-se no objeto em estudo. Neste delicado recorte do espaço geográfico, perdem importância os limites tradicionais de perímetro urbano e limites administrativos, uma vez que os trajetos intra-urbanos podem ultrapassar os limites municipais, como é o caso dos municípios vizinhos de São Carlos, Ibaté e Araraquara, bem como as localidades da Represa do Broa e os distritos de Água Vermelha e Santa Eudóxia, distantes da Sede de São Carlos pela ordem dos 20 aos 30 quilômetros (Ferreira, 2007).*

*A criação de loteamentos, condomínios urbanísticos, distritos industriais e outras formas de produção de bens imóveis “urbanos” em áreas de uso tradicionalmente rural é um dos elementos que vêm contribuindo para mudar os paradigmas que estabelecem as fronteiras entre campo e cidade. Deste fenômeno, é importante observar que a expansão territorial da malha urbana induz a variabilidade de características dos lotes, em termos de oferta de infra-estruturas, serviços, porte, segurança, acessibilidade e valor, dentre diversas outras. Segundo os dados do Instituto Brasileiro de Geografia e Estatística – IBGE, entre os anos 2000 e 2007, a população de São Carlos aumentou de 160.000 a 210.000 habitantes, com crescimento populacional superior às médias estadual e federal. É uma cidade de porte intermediário cujos padrões espaciais de valorização imobiliária não são tão simples, como em pequenos municípios, mas delineiam-se fatores cujos efeitos têm expressão máxima no ambiente metropolitano. Um dos principais deles é a segregação e as condições de competição pelo acesso às melhores localizações ocorrem no âmbito do mercado imobiliário (Ferreira, 2007).*

No âmbito deste estudo, o termo lote designa um bem imóvel composto por uma parcela de solo, resultado de loteamento, parcelamento ou fracionamento em condomínio. A rigor, esta conceituação não se restringe a uma definição legal, mas ao *aspecto puramente morfológico do terreno vazio*, assim como conceituado em Ferreira (2007). Como este objeto a ser amostrado constitui-se em terrenos vazios, estimados em 24.975 unidades, praticamente um quarto do universo dos imóveis do município de São Carlos, SP (São Carlos, 2006).

Sobre a escolha dos lotes como objeto de estudo, o autor justifica:

*Justifica-se que neste tipo de imóvel o valor da localização é explícito, ao passo que nos lotes edificados o valor do solo está implícito, juntamente com as características da construção e sua conservação. As interferências nos valores dos lotes vazios, embora sejam muitas, são bem menores que nos lotes edificados (Ferreira, 2007).*

A aplicabilidade do trabalho estende-se aos chamados vazios urbanos, onde costuma incidir a especulação imobiliária. Sobre isto, Ferreira (2007) discorre:

*Do conjunto de terrenos vazios do município, destaca-se o subconjunto das glebas. A avaliação de glebas urbanas envolve a escolha de um dos três métodos previstos em norma específica, a NBR 8591/1985 – Norma Brasileira de Avaliação de Glebas Urbanizáveis. Geralmente, a gleba é avaliada com base no potencial de retorno financeiro em face de determinado tipo de urbanização (parcelamentos, loteamentos, condomínios, entre outros). Pelo seu grande porte, forma, inserção na malha urbana e pelo fato de não terem sido ainda parceladas, as glebas apresentam particularidades que até então as impedem de serem avaliadas da mesma forma que os lotes. Porém, na equação que se propõe cuja modelagem é caracterizada por regiões homogêneas do ponto de vista da localização, a avaliação das glebas ganha uma dimensão contextualizada, na qual são consideradas variáveis comuns aos lotes. Desta forma, é possível quantificar o valor da localização implícito nestas glebas com mais clareza, e assim avaliar de forma mais abrangente aspectos de sub-utilização, especulação imobiliária e realização de sua função social, pois normalmente estas glebas constituem-se em vazios urbanos (Ferreira, 2007).*

As glebas de proprietários particulares também podem constituir áreas verdes inseridas no meio ambiente e, assim, contribuir para a melhoria do ambiente em seu entorno o

que daria à mesma um aspecto diferenciado. Esta contribuição pode estar relacionada ao conforto térmico, visual e de qualidade do ar, podendo até contribuir para a saúde das pessoas que vivem em seu entorno, além de abrigar pequenos animais. Sendo assim, estas glebas devem receber um tratamento diferenciado na PVG. As glebas de responsabilidade do poder público também devem ter um tratamento diferenciado já que propicia os benefícios citados acima para o seu entorno e valoriza os terrenos pertencentes a ela. Neste caso, devem ser tratadas como reservas ambientais e fica a cargo do poder público a sua preservação no que concerne aos aspectos de fauna, flora, segurança, uso, urbanização entre outras.

## 1.2. OBJETIVOS

### 1.2.1. Objetivo geral

Quando se busca a solução de um problema, obviamente, diferentes abordagens podem levar a resultados distintos, algumas vezes semelhantes, mas raramente iguais em todos os aspectos.

O presente trabalho visa dar tratamento científico à avaliação de imóveis urbanos e comparar a abordagem via Análise de Sobrevivência considerando censura à esquerda, desenvolvida neste trabalho, com as abordagens via Modelos Lineares Usuais que é considerada regra nesta área de avaliação de imóveis. Esta pesquisa foi desenvolvida com o auxílio do Software R 2.6.2. O sistema R é um ambiente que incorpora uma implementação da linguagem de programação S, que é poderosa, flexível e possui excelentes facilidades gráficas (R Development Core Team, 2008). R é um projeto de código-aberto (*open-source*) desenvolvido por muitos voluntários por mais de dez anos e está disponível na Internet sobre a *General Public Licence* ([www.gnu.org/copyleft/gpl.html](http://www.gnu.org/copyleft/gpl.html) e [www.fsf.org](http://www.fsf.org)) (Everitt *et al.*, 2006). A fonte principal de informações sobre o sistema R é a internet com a *home page* oficial do projeto R sendo: <http://www.r-project.org>. Todos os recursos estão disponíveis nesta página: o próprio sistema R, uma coleção de pacotes, manuais, documentação, novidades etc. (Everitt *et al.*, 2006).

Será disponibilizada uma metodologia científica para a proposta de uma equação de regressão representativa da formação do valor de mercado do solo urbano do município de São Carlos, SP, utilizando valores de lotes do ano de 2005. Posteriormente, esta metodologia permitirá a criação de uma PVG mais realista, pois futuros estudos partirão de uma divisão da

área urbana em grandes zonas homogêneas, que permitirão avaliar mais detidamente o comportamento de variáveis “micro”.

### **1.2.2. Objetivo específico**

O objetivo geral pode ser atingido desde que se alcancem os objetivos específicos. Será discutido e apresentado um novo caminho para pesquisadores e avaliadores de imóveis. Esse caminho é baseado na modelagem do valor de mercado do solo urbano através dos modelos de Análise de Sobrevivência. Até o presente momento, pesquisadores e avaliadores de imóveis não se enveredaram por esse caminho.

## **1.3. JUSTIFICATIVA**

O enfoque inovador do presente estudo é unir dois campos do conhecimento aparentemente distintos - as metodologias inferenciais de avaliação em massa (baseadas nos modelos lineares usuais e de Análise de Sobrevivência) e os estudos de localização intra-urbana, de natureza sociológica, geográfica e histórica. Ferreira (2007) e o presente trabalho se complementam na discussão do tema potencialmente comum, tal qual é a perspectiva de aplicação no planejamento urbano e tributário visando à imparcialidade das decisões.

Não foram encontradas publicações mencionando a aplicação de modelo estatístico para homogeneizar os valores aferidos na pesquisa de mercado baseado na metodologia de Análise de Sobrevivência.

## **1.4. ESTRUTURA DO TRABALHO**

Este trabalho está estruturado em 8 capítulos. O problema, o objetivo e a justificativa estão neste primeiro capítulo. Os capítulos seguintes estão organizados assim: no capítulo 2

descreve-se uma revisão bibliográfica relacionada à importância da PVG, a cidade de São Carlos, SP, os conceitos de localização, visando seu emprego em modelos estatísticos de avaliação e as variáveis utilizadas; no capítulo 3 ao capítulo 7 são apresentados os conceitos e ajustes dos modelos lineares de regressão usual e de regressão Weibull (Análise de Sobrevivência) para uma amostra de lotes urbanos da cidade de São Carlos, SP e, também, um estudo de simulação da probabilidade de cobertura de alguns parâmetros para ambos os modelos; no capítulo 8 estão as conclusões e perspectivas futuras; todos os códigos dos programas utilizados para a realização das análises estatísticas e gráficos nos *softwares* R e SAS são apresentados nos Apêndices A, B e C. No Apêndice A estão os códigos dos programas em R, versão 2.6.2, referentes à modelagem dos dados para ambos os modelos. No Apêndice B estão os códigos dos programas em R referentes à simulação da probabilidade de cobertura dos parâmetros para ambos os modelos e, por último, no Apêndice C, estão os códigos referentes à implementação do gráfico TTT-Plot no SAS, versão 9.0.

## 2. REVISÃO BIBLIOGRÁFICA

### 2.1. PLANTAS DE VALORES GENÉRICOS

A PVG é a representação do valor unitário do solo, normalmente expresso em Reais por área (R\$/m<sup>2</sup>). A base de cálculo do IPTU e do ITBI é o *valor venal* do imóvel.

*Por princípio, o valor venal deve guardar uma proporcionalidade em relação ao seu valor de mercado. A PVG e conseqüentemente os valores venais são instituídos através de lei municipal. A PVG é um importante instrumento de política tributária para os governos municipais. Um instrumento que, se bem utilizado, promove a justiça fiscal, contribui para o planejamento urbano, interage e influencia o mercado imobiliário, podendo ser utilizada como instrumento de indução de políticas de ocupação do solo urbano, além de aperfeiçoar a arrecadação dos tributos que são de sua competência. Nem sempre bem elaboradas e atualizadas, as PVG comumente se restringem a uma mera base de cálculo de imposto territorial (Ferreira, 2007).*

Neste trabalho, na elaboração da PVG, será utilizado apenas o método científico, devido a sua objetividade. O método científico baseia-se em modelos estatísticos para homogeneizar os valores aferidos na pesquisa de mercado. Atualmente, estes modelos são obtidos através de regressão linear múltipla usual (erros normais com média zero e variância constante).

A análise de regressão é um dos ramos da teoria estatística mais utilizada na pesquisa científica. É a técnica mais adequada quando se deseja estudar o comportamento de uma variável em relação a outras que são de certa forma, responsáveis pela sua formação. O modelo de regressão linear múltipla deve ser adotado quando mais de uma variável é necessária para explicar a variabilidade dos preços praticados no mercado (Dantas, 1998).

### 2.2. A ÁREA DE ESTUDO

A cidade de São Carlos localiza-se no centro geográfico do Estado de São Paulo. O clima ameno, com temperatura média anual de 19,6 °C, somado à altitude média de 856m compreendida entre 520 e 1.000 metros, faz de São Carlos um local muito agradável, com



inúmeras cachoeiras, curiosas formações geológicas e belíssimas paisagens (São Carlos, 2008).

São Carlos conta com a presença de duas universidades públicas, a Universidade de São Paulo (USP), com dois campi na cidade, e a Universidade Federal de São Carlos (UFSCar) além da Embrapa (Empresa Brasileira de Pesquisa Agropecuária) com o Centro de Pesquisa de Pecuária do Sudeste e o Centro Nacional de Pesquisa e Desenvolvimento de Instrumentação Agropecuária. Tais instituições incorporaram à história de São Carlos suas contribuições à ciência e à capacitação profissional de milhares de pessoas além de produzir tecnologia de ponta nas áreas de melhoramento genético bovino e de desenvolvimento de equipamentos agropecuários. Este vigor acadêmico aliado com o potencial tecnológico e industrial conferiu à cidade o título de Capital da Tecnologia (São Carlos, 2008).

Segundo os dados do Instituto Brasileiro de Geografia e Estatística – IBGE, entre os anos 2000 e 2007, a população de São Carlos, SP, aumentou de 160.000 para 210.000 habitantes, com crescimento populacional superior às médias estadual e federal.

*“É uma cidade de porte intermediário cujos padrões espaciais de valorização imobiliária não são tão simples, como em pequenos municípios, mas delineiam-se fatores cujos efeitos têm expressão máxima no ambiente metropolitano. Um dos principais deles é a segregação” (Ferreira, 2007).*

Os limites das coordenadas geográficas do município são 47°30' e 48°30' na longitude oeste 21°30' e 22°30' na latitude sul. São Carlos, SP, apresenta os seguintes municípios vizinhos: Ibaté, Itirapina, Rincão, Santa Lúcia, Analândia, Luís Antônio, Araraquara, Descalvado, Brotas, Américo Brasiliense e Ribeirão Bonito e, tem como distritos, Água Vermelha e Santa Eudóxia (São Carlos, 2008).

### **2.3. CONFIGURAÇÃO ESPACIAL DE SÃO CARLOS, SP, E CONSIDERAÇÕES SOBRE A MODELAGEM DA VARIÁVEL LOCALIZAÇÃO**

O objetivo deste capítulo é conceituar brevemente a localização, visando seu emprego em modelos estatísticos de avaliação. A localização é a principal variável explicativa do valor de mercado dos lotes urbanos, embora seu tratamento em modelos estatísticos seja de certa forma controversa.

*Nos modelos estatísticos, a variável localização pode assumir diversos tipos. Isto cria inúmeras possibilidades de representação, que coloca ao pesquisador o problema sobre qual modelo especificar. São necessárias bases teóricas que dêem sentido à formulação estatística. Há uma quantidade considerável de técnicas matemáticas, estatísticas e computacionais, em paralelo a abordagens relativamente parciais do problema da localização urbana. Formas específicas de avaliação, como as Plantas de Valores, impõem necessariamente abordagens extensivas a todo o universo de lotes urbanos dos municípios. É desejável que os modelos estatísticos e suas ferramentas, ainda que sofisticados por si, tenham sentido em sua aplicação. Este sentido deve seguir um conjunto de pressupostos sobre os mecanismos de valorização do solo urbano (Ferreira, 2007).*

*De fato, a configuração dos valores ao longo da malha urbana é produto dialético de determinantes históricas em uma estrutura que constantemente se transforma. Compreender os processos formadores de valor do solo, sob esta perspectiva genérica, exige o reconhecimento da estruturação urbana de uma cidade (Ferreira, 2007).*

*Os estudos de localização intra-urbana representam uma importante contribuição para a construção destes pressupostos sobre os processos de formação de valor. As aglomerações urbanas são constituídas por diversas localizações, que têm a acessibilidade como um dos principais atributos, com grande influência na valorização do lote urbano. As localizações intra-urbanas que apresentam a característica de minimizar o conjunto dos tempos de deslocamento diário de pessoas, tais como os trajetos casa-trabalho, casa-lazer, casa-escola, casa-consumo, dentre outros, são mais valorizadas. Neste caso, as áreas mais valorizadas são aquelas que se aproximam dos pontos de consumo e trabalho das camadas sociais com melhores condições de educação e renda (Ferreira, 2007).*

*Há duas forças que dispõem de ampla liberdade para escolha de suas localizações. Uma delas é de natureza extra-urbana, regional, ligada à esfera da produção, geralmente industrial. A outra é de natureza local, fortemente associada à esfera do consumo e constitui-se na localização dos bairros residenciais das camadas de maior renda. O território restante é disputado pelos diversos segmentos de menor poder econômico no âmbito do mercado imobiliário, que expressa em preços o valor do solo urbano e estabelece a seleção a partir do poder aquisitivo (Ferreira, 2007).*

*Neste contexto, as barreiras são relevantes porque desvalorizam certas localizações em relação a outras. Estas barreiras são de ordem natural, como a conformação do relevo e a hidrografia e sob a influência destes, freqüentemente, são traçadas as ferrovias e as rodovias (Ferreira, 2007).*

A princípio, Ferreira (2007) identifica quatro grandes barreiras intra-urbanas em São Carlos, SP: a delimitação natural da planície central, a ferrovia, a rodovia SP-310 (Washington Luiz) e a encosta sul. Cada barreira tem sua própria natureza e suas características próprias de transposição, que impõem diferentes graus de dificuldades aos ocupantes. O autor acrescenta:

*Este processo de consolidação de barreiras, reafirmação de trajetos intra e extra-urbanos e consolidação de infra-estruturas vem ocorrendo gradativamente, desde os primórdios da formação do núcleo urbano do município de São Carlos. O momento presente é sempre o resultado de todas estas determinantes. À medida que o tempo transcorre, torna-se mais complexa a leitura e compreensão destes fenômenos ligados à localização intra-urbana (Ferreira, 2007).*

Ferreira (2007) caracterizou a distribuição espacial das moradias dos diversos extratos populacionais da cidade de São Carlos, SP, com base nos dados de educação e renda constantes do censo 2000 (IBGE 2002). O método usado nesta análise consiste na determinação de uma distância estatística entre os 241 setores censitários, com base no método de análise multivariada descrito em Johnson e Wichern (1998), p. 28-32. Ver Figura 2.

Por este método foi possível caracterizar dois setores censitários mais distantes entre si: em um dos extremos, ou pólos, o Parque Faber I, que apresenta as melhores condições de educação e renda dos responsáveis pelos domicílios. No pólo oposto, uma parte do Jardim Gonzaga, chamada “favelinha”, que na época da realização do Censo 2000 ainda não havia sido selecionada para receber as melhorias urbanísticas do Programa Habitar Brasil/BID.

*Além de características geográficas, há mecanismos adicionais que procuram garantir a manutenção dos valores imobiliários das localizações ocupadas pelos bairros residenciais das camadas de maior renda. No contexto de São Carlos, acredita-se que sejam três: a legislação urbanística que torna estes bairros estritamente residenciais; os muros que delimitam alguns bairros, gerando o que se conhece popularmente como condomínio horizontal, embora não o sejam necessariamente; e finalmente o regime condominial, que de fato caracteriza legalmente o condomínio e influi na manutenção das infra-estruturas coletivas, incluindo-se ruas, calçadas, iluminação, arborização, dentre outros. Estas três características remontam basicamente ao problema da segurança, da privacidade e do controle privado de padrões ambientais e urbanísticos (Ferreira, 2007).*

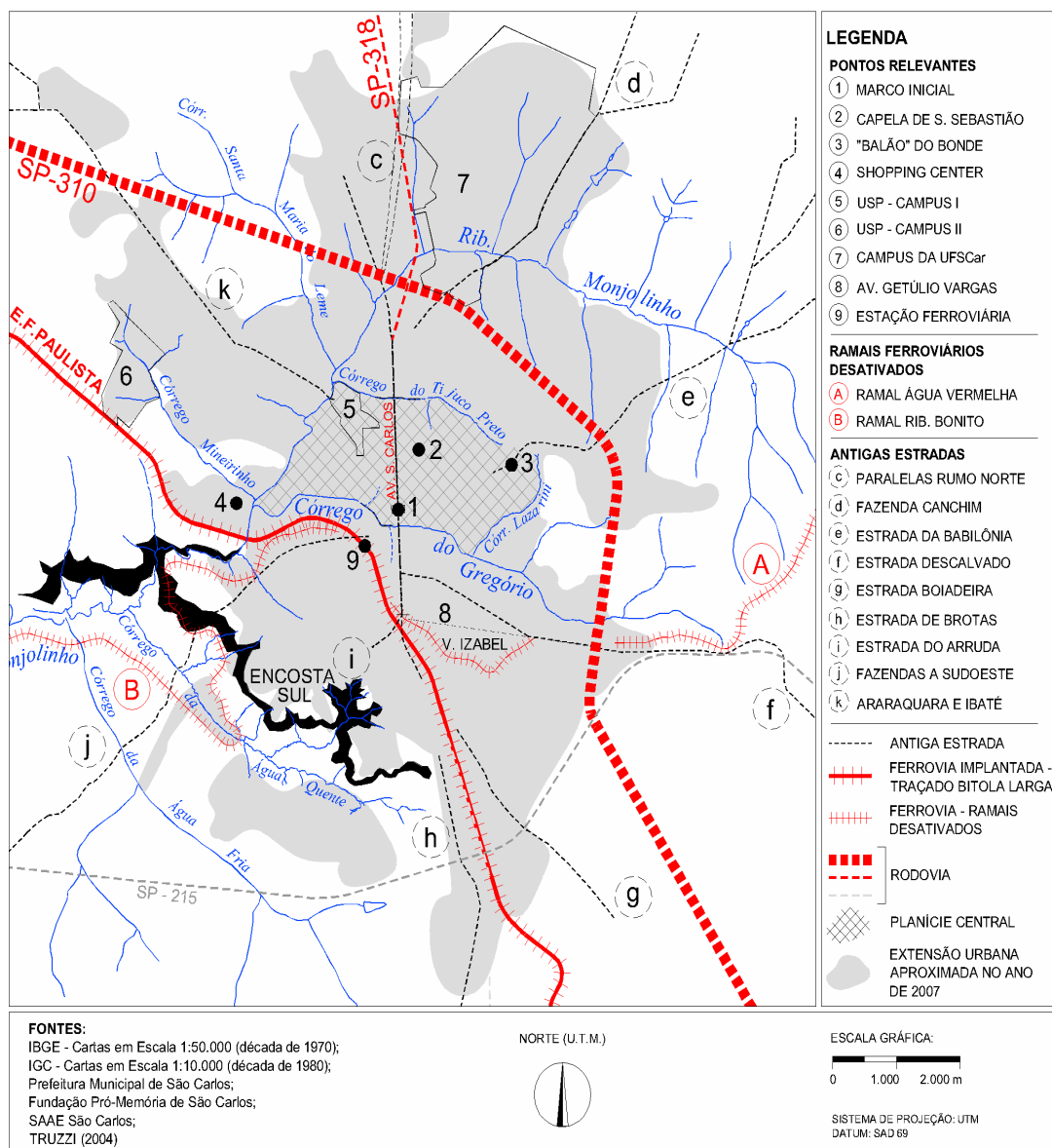


Figura 1. Configuração planimétrica da sede do município de São Carlos, SP, com indicação das maiores barreiras intra-urbanas. Fonte: Ferreira (2007).

No delineamento da presente pesquisa, foi adotada a proposta de Ferreira (2007) para a criação de unidades de localização consideradas básicas para a formulação de um modelo estatístico que exprima de forma potencial o fenômeno da valorização territorial urbana. O sentido da construção destas unidades parte do nível geral em direção ao nível específico. Em uma abordagem geral se lida com características semelhantes compartilhadas por diversas localizações contíguas e semelhantes. Estas variáveis serão utilizadas no modelo a ser proposto. Cada hipótese corresponde a uma variável a ser incluída neste modelo. Estas variáveis serão apresentadas brevemente, a seguir.

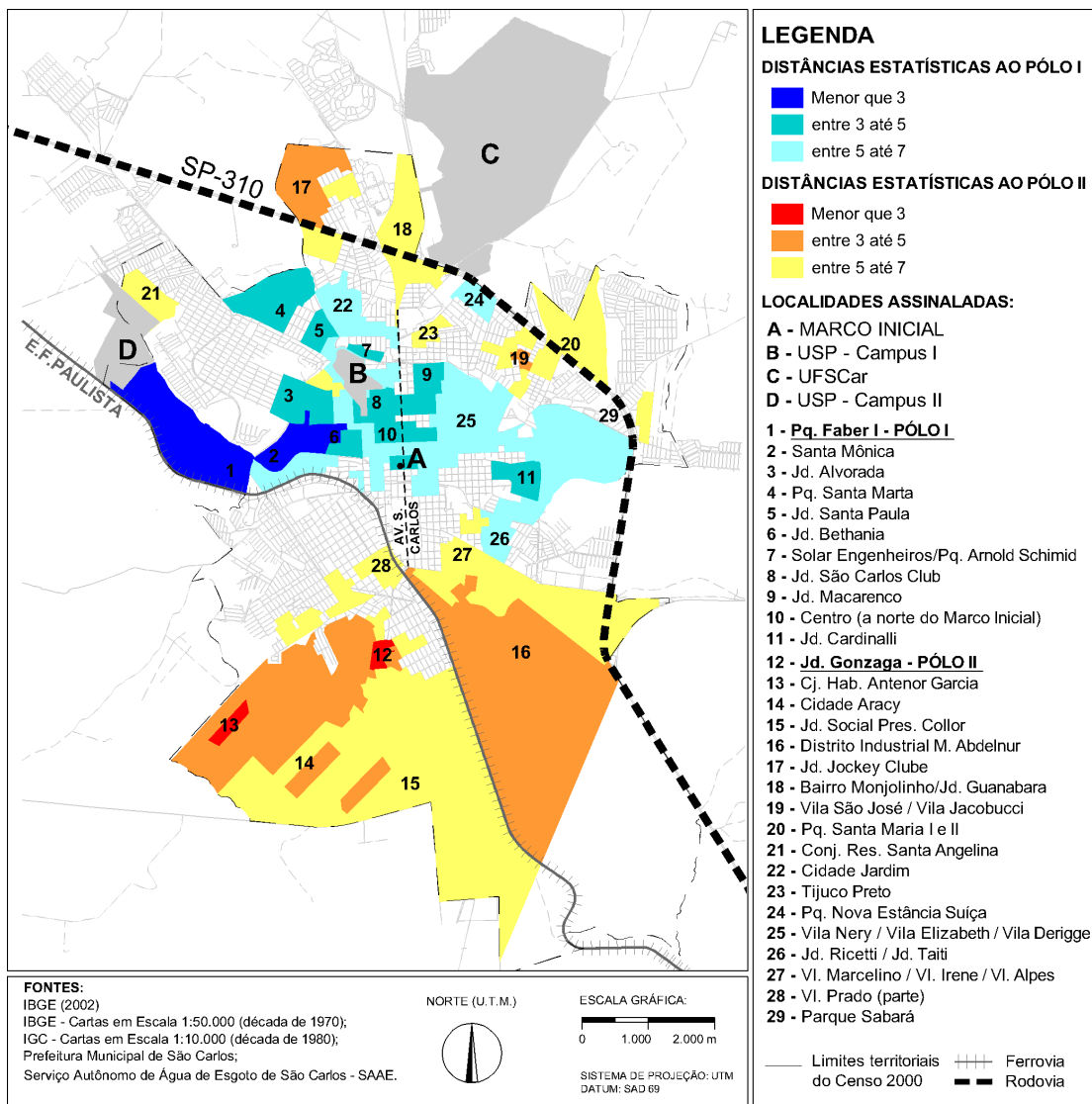


Figura 2. Mapa da classificação dos setores censitários segundo o valor da distância estatística a um dos pólos sócio-econômicos opostos. Fonte: Ferreira (2007).

## 2.4. VARIÁVEIS UTILIZADAS

Conforme Ferreira (2007) foram consideradas oito variáveis binárias, apresentadas a seguir, e uma única variável numérica.

1. **Variável Planície Central** - Esta localização será indicada como uma variável binária, atribuindo-se o valor 1 às áreas que estão contidas no interior desta área e 0 às áreas externas. A hipótese que corresponde a esta variável é que a pertinência a estas áreas é um fator de valorização.

2. **Variável Ferrovia** - A ferrovia representa uma barreira considerável. Atribui-se o valor 1 às áreas que não são afetadas pela barreira ferroviária e 0 às áreas afetadas. A dicotomia desta variável é útil para estimar as médias de desvalorização decorrentes da influência da ferrovia.
3. **Variável Rodovia** - Esta variável refere-se às barreiras de acessibilidade decorrentes da Rodovia SP-310 e segue a mesma especificação que a variável Ferrovia; atribui-se o valor 1 às áreas que não são afetadas pela barreira rodoviária e 0 às áreas afetadas. A exceção ocorre nas localidades cujo acesso é feito sobre o nível do pavimento da SP-310, na continuação da Avenida São Carlos e também na continuação da Avenida Getúlio Vargas. A hipótese que corresponde a esta variável consiste em considerar que as áreas de acessibilidade não afetada pela rodovia SP-310 são mais valorizadas.
4. **Variável Encosta Sul** - A hipótese correspondente a esta variável é que as áreas situadas aquém da barreira da encosta sul são mais valorizadas. Esta variável é do tipo binária, assumindo o valor 1 para as áreas cuja acessibilidade não é afetada pela barreira. Atribui-se o valor 0 para as áreas afetadas.
5. **Variável Fechado** - Um parcelamento fechado é referido popularmente como “condomínio”, pelo seu aspecto murado. A rigor, o condomínio define-se por um conjunto de características além da existência de muro envoltório. De forma geral, o fechamento remete à uma idéia de segurança contra a criminalidade. Uma das hipóteses considera que o fechamento destes parcelamentos seja um fator de valorização.
6. **Variável Condomínio** - Como dito anteriormente, é comum um cidadão confundir um parcelamento simplesmente fechado com um condomínio de fato instituído. A instituição do condomínio define as áreas privativas e comuns, estabelecendo-se frações ideais (porcentagens) da participação. Nestes casos, a manutenção das infra-estruturas de saneamento, pavimentação, drenagem, iluminação, segurança e outros se vincula à taxa de condomínio. Embora a taxa de condomínio represente um custo adicional para a posse de uma parcela de solo contida em seu interior, o que é acessível a uma reduzida parcela de residentes, a hipótese correspondente a esta variável consiste em considerar que a instituição de condomínio é um fator de valorização. Esta localização será indicada como uma variável binária, atribuindo-se o valor 1 às áreas que estão contidas no interior de condomínios fechados registrados e o valor 0 às áreas externas.
7. **Variável Uso Residencial** - Com base na amostragem preliminar e nas análises estatísticas sobre os dados censitários observa-se que existe uma forte correlação entre ocupação residencial de maior renda e os loteamentos com restrições a usos não residenciais. Estas restrições referem-se às cláusulas expressas nos contratos de compra e venda de lotes,

pesquisados na Prefeitura Municipal de São Carlos, SP. Observa-se em termos gerais que existe uma grande quantidade bem como uma variedade expressiva de tipos de restrições. Por exemplo, restrições de usos ou tipologias, a desmembramentos de lotes, recuos frontais e laterais, índices, restrições de altura das edificações e outras. Nota-se que o controle através de restrições urbanísticas tende a aumentar nos parcelamentos fechados, estritamente residenciais ou nos condomínios edifícios. Nestes parcelamentos, seu nível é ainda mais restritivo que o estabelecido pelo conjunto da legislação municipal. 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. No modelo a ser proposto, a variável Uso Residencial assume o tipo binário, atribuindo-se o valor 1 às áreas pertencentes a estes loteamentos e o valor 0 às áreas não pertencentes.

8. **Variável Núcleo Sede** - Esta variável é do tipo binária e assume 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 isolado desta aglomeração. A dicotomia desta variável relaciona-se principalmente às localizações relativamente remotas com relação à sede do município, separados por várias bacias hidrográficas. A dicotomia é útil para comparar localidades fora dos limites administrativos do município de São Carlos, SP, mas estreitamente ligadas ao seu espaço intra-urbano, como por exemplo, a Represa do Broa e o município de Ibaté, onde não é raro seus residentes estabelecerem viagens diárias do tipo casa-trabalho, casa-escola, casa-lazer e outros. É possível comparar contextos incomuns e semelhantes, como a ocupação em torno das represas do 29 e do Broa, que apresentam a mesma distância planimétrica, da ordem dos 20 a 30 quilômetros em relação ao marco inicial de São Carlos, SP. Estes parcelamentos rurais caracterizam-se também pela ausência de infra-estruturas tais como: pavimentação viária, redes gerais de saneamento, iluminação pública.

O produto da sobreposição de todas as oito variáveis binárias e suas abrangências territoriais é demonstrado na Figura 3. Esta sobreposição gera regiões homogêneas quanto a características gerais de acessibilidade e aspectos jurídicos (restrição de uso, regime condominial). Com base nesta regionalização é possível estudar o comportamento de demais variáveis no interior de uma região homogênea.

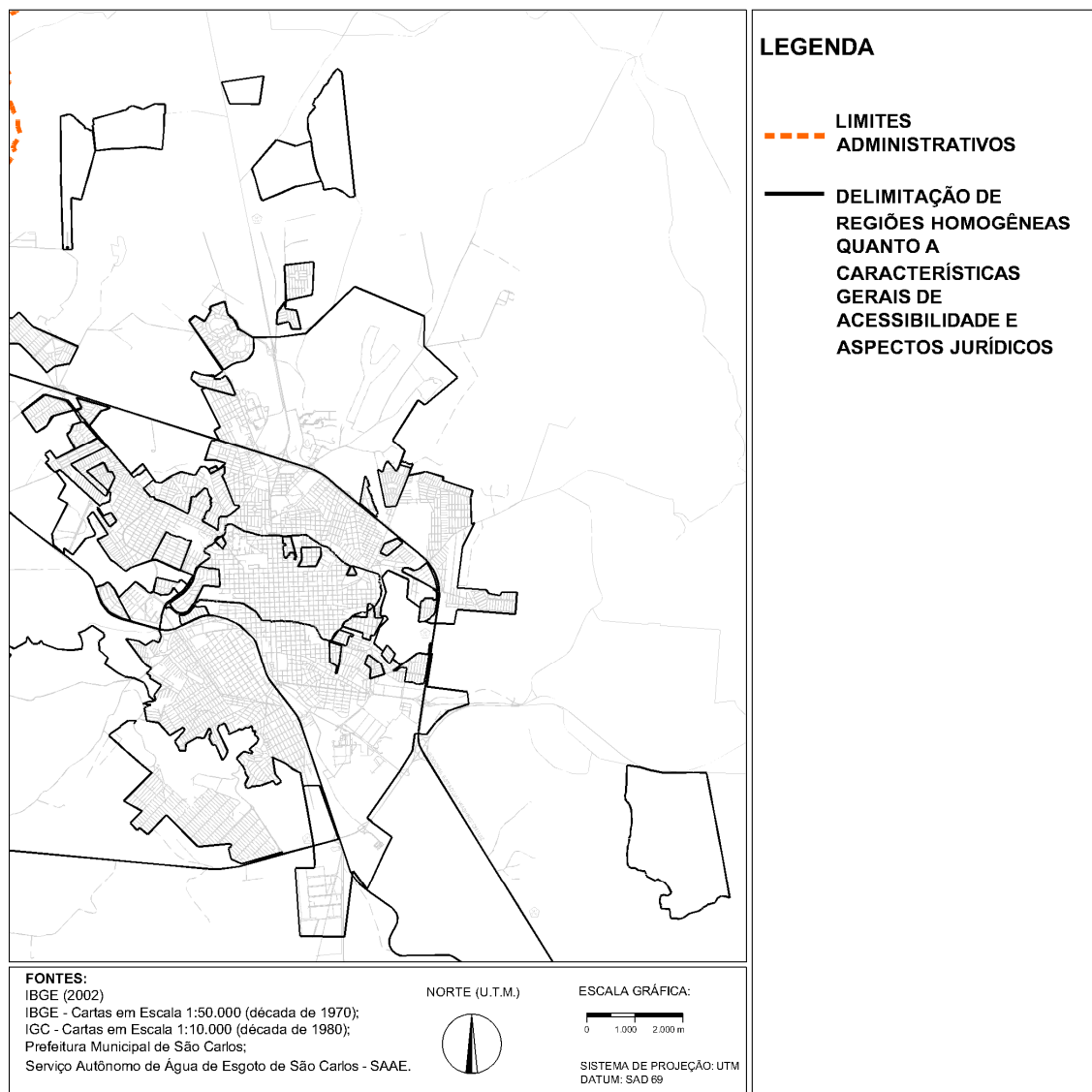


Figura 3. Delimitação de regiões homogêneas com base na sobreposição da abrangência territorial das variáveis selecionadas. Fonte: Ferreira (2007).

Portanto, no modelo a ser proposto, a localização será expressa por oito variáveis dicotômicas, conforme indicado na Tabela 1.

E a nona variável incluída no modelo representa a área territorial do lote amostrado, em metros quadrados.



Tabela 1. Descrição das variáveis dicotômicas indicadoras da localização.

<b>Nome da variável</b>	<b>Descrição</b>
NUC_PRINC	1 = Lote localiza-se contíguo à aglomeração da sede do município; 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.

Outros aspectos importantes que não foram tratados aqui e que podem ser considerados, em pesquisas futuras, na tentativa de caracterizar regiões homogêneas são:

- coleta de lixo;
- saneamento básico;
- iluminação pública;
- pavimentação de ruas;
- distância aos pólos de valorização/desvalorização;

### 3. MODELOS DE REGRESSÃO

A análise de dados através da regressão linear é uma das técnicas mais usadas de estimação, existindo uma ampla literatura sobre o assunto. O modelo clássico de regressão teve origem nos trabalhos de astronomia elaborados por Gauss no período de 1809 a 1821. É a técnica mais adequada quando se deseja estudar o comportamento de uma variável dependente ou variável resposta ( $y$ ) em relação a outras variáveis independentes ou variáveis explicativas ( $X$ ) que podem explicar a variabilidade da variável resposta (Cordeiro e Lima Neto, 2004).

Segundo Demétrio e Zocchi (2007), a utilização de modelos de regressão pode ter por objetivos:

- a) **Predição.** Espera-se que uma parte da variação (o interesse é a maior parte) de  $y$  é explicada pelas variáveis  $X$ , então, pode-se utilizar o modelo para aproximar valores de  $y$  correspondentes a valores de  $X$  que não estavam entre os dados. Esse processo denomina-se **predição** onde os valores de  $X$  para predizer  $y$  devem pertencer ao intervalo original de  $X$ . A utilização de valores fora desse intervalo recebe o nome de extrapolação e, deve ser usada com muito cuidado, pois o modelo adotado pode não ser correto fora do intervalo estudado.
- b) **Seleção de variáveis.** Frequentemente, não se tem idéia de quais são as variáveis que afetam significativamente a variação de  $y$ . Para responder a esse tipo de questão, conduzem-se estudos onde está presente um grande número de variáveis. Posteriormente, geram-se vários modelos onde são escolhidos os melhores por algum critério, adotando-se como variáveis explicativas de  $y$  aquelas incluídas no “melhor” modelo.
- c) **Estimação de parâmetros.** Dado um modelo e um conjunto de dados (amostra) referente às variáveis, resposta e preditoras, estimar parâmetros, ou ainda, ajustar o modelo aos dados, significa obter valores ótimos para os parâmetros, por algum critério adotado.
- d) **Inferência.** O ajuste de um modelo de regressão tem, em geral, por objetivos básicos, além de estimar os parâmetros, realizar inferências sobre eles, tais como testes de hipóteses e intervalos de confiança os quais indicam uma variação confiável desses valores.

#### 4. MODELOS LINEARES USUAIS

O modelo clássico de regressão linear é definido por:

a) repostas  $y_i$  independentes (ou pelo menos não correlacionadas) para  $i = 1, \dots, n$ , tendo, cada  $y_i$ , uma distribuição especificada de média  $\mu_i = E(y_i)$  e variância  $\sigma^2$ ;

b) a média  $\mu_i$  é expressa de forma linear como  $\mu_i = x_i^T \beta$ , onde  $x_i^T$  é um vetor  $1 \times p$  com os valores de  $p$  variáveis explicativas relacionadas a  $i$ -ésima resposta  $y_i$  e  $\beta$  é um vetor  $p \times 1$  de parâmetros a serem estimados.

A estrutura a) e b) pode também ser expressa na forma matricial  $\mu = E(y) = X\beta$ , onde  $y = (y_1, \dots, y_n)^T$  é um vetor  $n \times 1$  cuja  $i$ -ésima componente é  $y_i$  e  $X$  é uma matriz  $n \times p$  formada pelas linhas  $x_1^T, \dots, x_n^T$ . Em geral, os pesquisadores adotam a hipótese de aditividade entre  $y$  e  $\mu$ , isto é,  $y = \mu + \varepsilon$ , onde  $\varepsilon$  é um vetor de erros de média zero e variância  $\sigma^2$ . Os erros são considerados independentes ou pelo menos não-correlacionados. Os efeitos das variáveis explicativas, que formam as colunas da matriz  $X$ , sobre a variável resposta  $y$  são lineares e aditivos. Na formação da matriz modelo, considera geralmente a primeira coluna como um vetor de uns sendo o parâmetro correspondente denominado *intercepto*.

O objetivo inicial é estimar  $\beta$  a partir do vetor  $y$  de dados e da matriz de modelo  $X$  conhecida, suposta de posto completo  $p$ . A estimação pelo *Método de Mínimos Quadrados* não requer qualquer hipótese sobre a distribuição das componentes do vetor  $y$ . Segundo as hipóteses a) e b) o método de mínimos quadrados continua sendo o método preferido entre estes métodos de estimação (Cordeiro e Lima Neto, 2004).

##### 4.1. ESTIMAÇÃO

Adotamos a seguinte notação matricial para representar o modelo clássico de regressão linear

$$y = X\beta + \varepsilon \quad (4.1)$$

em que  $X\beta$  expressa a aditividade entre os efeitos lineares sistemáticos e  $\mathcal{E}$  os efeitos aleatórios, onde  $Cov(\mathcal{E}) = \sigma^2 I$ . A soma de quadrados dos erros  $SQE(\beta) = \sum_i (y_i - \mu_i)^2$  correspondente ao modelo (4.1) é dada em notação matricial por

$$SQE(\beta) = (y - X\beta)^T (y - X\beta) \quad (4.2)$$

Como a matriz modelo  $X$  tem posto completo, a matriz  $X^T X$  é invertível e o *estimador de mínimos quadrados* (EMQ) de  $\beta$  é dado por

$$\hat{\beta} = (X^T X)^{-1} X^T y \quad (4.3)$$

O EMQ  $\hat{\beta}$  em (4.3), segundo o modelo (4.1), tem as seguintes propriedades:

- a)  $\hat{\beta}$  minimiza a soma de quadrados dos erros,  $\sum_i \varepsilon_i^2$ , independentemente da distribuição proposta para os erros. Não é necessário conhecer a distribuição dos erros para estimar  $\beta$ , mas a normalidade é necessária para fazer inferência sobre os parâmetros em  $\beta$ . Esta inferência baseia-se nas distribuições t de Student e F de Snedecor;
- b) as componentes do vetor  $\hat{\beta}$  são funções lineares das observações e são estimadores não-viesados de menor variância dos parâmetros em  $\beta$ .

## 4.2. SOMAS DE QUADRADOS

O valor mínimo da soma de quadrados dos erros é denominado soma de quadrados dos resíduos (SQR), pois mede a discrepância entre o vetor de observações  $y$  e o vetor de valores ajustados (ou médias ajustadas)  $\hat{\mu} = X\hat{\beta}$ . Assim, SQR é expresso por

$$SQR = SQE(\hat{\beta}) = (y - X\hat{\beta})^T (y - X\hat{\beta}) \quad (4.4)$$

Verificamos facilmente que  $\hat{\mu} = X(X^T X)^{-1} X^T y = Hy$ , onde a matriz  $H$  é denominada matriz de projeção. A razão desta terminologia é que o vetor dos valores ajustados é a projeção ortogonal do vetor de dados  $y$  no espaço gerado pelas colunas da matriz  $X$ .

A matriz  $H$  é simétrica ( $H = H^T$ ), idempotente ( $H^2 = H$ ) e tem posto  $p$ .

O vetor de erros não-observados  $\varepsilon = y - X\beta$  é estimado pelo vetor de resíduos  $r$ , dado por

$$r = y - \hat{\mu} = y - X\hat{\beta} \quad (4.5)$$

Tem-se  $r = y - Hy = (I - H)y$ , onde  $I$  representa a matriz identidade de ordem  $n$ .

Cordeiro e Lima Neto (2004) mostram que a soma de quadrados dos dados  $(y^T y)$  é igual a soma de quadrados dos valores ajustados  $(\hat{\mu}^T \hat{\mu})$  mais a soma de quadrados dos resíduos  $(r^T r)$  (ver Eq. (4.6))

$$y^T y = \hat{\mu}^T \hat{\mu} + r^T r \quad (4.6)$$

A equação (4.6) é uma aplicação do teorema de Pitágoras, onde a hipotenusa é o vetor de dados  $y$ , e os catetos são os vetores das médias ajustadas  $\hat{\mu}$  e dos resíduos  $r = y - \hat{\mu}$ . Assim, a soma de quadrados das observações  $y^T y$  pode ser decomposta em duas partes: a soma de quadrados dos valores ajustados  $\hat{\mu}^T \hat{\mu} = \hat{\beta}^T X^T y$  e a soma de quadrados dos resíduos  $SQR = r^T r = (y - \hat{\mu})^T (y - \hat{\mu})$ , que mede a variabilidade dos dados não explicada pela regressão (Cordeiro e Lima Neto, 2004).

### 4.3. PROPRIEDADES DO EMQ E DOS RESÍDUOS

Nesta seção apresentamos algumas propriedades de  $\hat{\beta}$  que são baseadas apenas nas duas hipóteses básicas atribuídas aos dois primeiros momentos dos erros:  $E(\varepsilon) = 0$  e  $Cov(\varepsilon) = \sigma^2 I$ . Aqui seremos breves e apresentaremos os resultados. Um estudo completo é apresentado por Cordeiro e Lima Neto (2004).

a) O EMQ  $\hat{\beta}$  é não-viesado, ou seja,  $E(\hat{\beta}) = \beta$ .

b) A covariância do EMQ  $\hat{\beta}$  é dada por

$$Cov(\hat{\beta}) = \sigma^2 (X^T X)^{-1} \quad (4.7)$$

Assim, a matriz inversa  $(X^T X)^{-1}$  usada para estimar  $\beta$  em (4.3) determina a matriz de covariância de  $\hat{\beta}$ , exceto pelo multiplicador  $\sigma^2$ . Os elementos da diagonal da matriz de

covariância de  $\hat{\beta}$  ( $Cov(\hat{\beta})$ ) são as variâncias dos estimadores de mínimos quadrados dos parâmetros em  $\beta$  e, portanto, representam a precisão destas estimativas.

c) A covariância do vetor  $\hat{\mu}$  é dada por

$$Cov(\hat{\mu}) = XCov(\hat{\beta})X^T = \sigma^2 X(X^T X)^{-1} X^T = \sigma^2 H \quad (4.8)$$

Assim, a matriz de projeção  $H$  representa, exceto pelo escalar  $\sigma^2$ , a matriz de covariância  $\hat{\mu}$ . Logo,  $Cov(\hat{\mu}_i, \hat{\mu}_j) = \sigma^2 h_{ij}$ , onde  $h_{ij}$  é o elemento da matriz  $H$ . As propriedades desta matriz serão detalhadas a seguir.

d) Estimação de  $\sigma^2$ .

Para determinar as covariâncias de  $\hat{\beta}$  e  $\hat{\mu}$  é necessário estimar a variância dos erros. Após alguns cálculos obtemos um estimador não-viesado de  $\sigma^2$ ,

$$\hat{\sigma}^2 = \frac{(y - X\hat{\beta})^T (y - X\hat{\beta})}{n - p} = \frac{SQR}{n - p}. \quad (4.9)$$

Estimando-se  $\sigma^2$  por (4.9), podemos calcular as covariâncias das estimativas dos parâmetros da regressão. A grande maioria dos programas computacionais que realizam a análise de regressão apresenta as estimativas  $\hat{\beta}_1, \dots, \hat{\beta}_p$  e seus erros padrões  $Var(\hat{\beta}_1)^{1/2}, \dots, Var(\hat{\beta}_p)^{1/2}$ , que correspondem às raízes quadradas dos elementos da diagonal da matriz de covariâncias de  $\hat{\beta}$  ( $Cov(\hat{\beta})$ ).

e) Esperança e Covariância do Vetor de Resíduos  $r$ .

A esperança de  $r$  é o vetor nulo, isto é,  $E(r) = 0$ , e a matriz de covariância de  $r$  é dada por  $Cov(r) = \sigma^2(I - H)$ . Logo, a covariância entre os resíduos  $r_i = y_i - \hat{\mu}_i$  e  $r_j = y_j - \hat{\mu}_j$  relativos às observações de ordens  $i$  e  $j$ , é dada por

$$Cov(r_i, r_j) = -\sigma^2 h_{ij} \quad (4.10)$$

Assim, embora os erros aleatórios  $\varepsilon_i$  tenham a mesma variância  $\sigma^2$ , i.e., sejam homocedásticos, o mesmo não ocorre com os resíduos, cujas variâncias dependem dos elementos da diagonal da matriz de projeção  $H$ . Temos,  $Var(r_i) = \sigma^2(1 - h_{ii})$  e, então, os resíduos definidos em (4.5) são heterocedásticos.

f) Covariância entre  $\hat{\beta}$  e  $r$ .

Os vetores  $\hat{\beta}$  e  $r$  são ortogonais, ou seja,  $Cov\left(\hat{\beta}, r\right) = 0$ . Temos,

$$Cov(\hat{\beta}, r) = 0$$

Em outras palavras, os vetores  $\hat{\beta}$  e  $r$  são não-correlacionados.

O vetor de resíduos  $r$  é, também ortogonal ao vetor das médias ajustadas  $\hat{\mu}$ .

$$\hat{\mu}^T r = y^T H^T (I - H)y = y^T (H - H)y = 0$$

#### 4.4. MODELO NORMAL-LINEAR

Vimos anteriormente que não é necessário conhecer a distribuição dos erros para estimar  $\beta$ , mas a normalidade é necessária para fazer inferência sobre os parâmetros em  $\beta$ . Conforme Cordeiro e Lima Neto (2004) para determinarmos a distribuição das estimativas de mínimos quadrados precisamos especificar a distribuição dos erros aleatórios. A suposição de normalidade dos erros é a mais adotada e considera que os erros aleatórios  $\varepsilon_1, \dots, \varepsilon_n$  em (4.1) são independentes e têm distribuição normal  $N(0, \sigma^2)$ .

O modelo (4.1), com esta suposição, é denominado *modelo normal-linear*. Segundo a hipótese de normalidade dos erros, podemos deduzir as seguintes propriedades que são importantes na análise de regressão:

- i) O vetor  $y$  tem distribuição normal  $n$ -variada  $N_n(X\beta, \sigma^2 I)$ .
- ii) O EMQ  $\hat{\beta}$  tem distribuição normal  $p$ -variada  $N_p(\beta, \sigma^2 (X^T X)^{-1})$ .

A média e a estrutura de covariância de  $\hat{\beta}$  foram descritas em 4.3, itens a) e b). A normalidade de  $\hat{\beta}$  decorre do fato de  $\hat{\beta}$  ser uma função linear do vetor  $y$ , cuja distribuição é normal.

iii) O EMQ  $\hat{\beta}$  e a soma de quadrados dos resíduos  $SQR = y^T (I - H)y$  são independentes.

O vetor de resíduos  $r = y - \hat{\mu} = (I - H)y$  tem distribuição normal  $n$ -variada e é ortogonal ao EMQ  $\hat{\beta}$ . Assim, como  $\hat{\beta}$  e  $r$  são ortogonais e tem distribuição normal, estes vetores são independentes. Então, o EMQ  $\hat{\beta}$  e a soma  $SQR$  são independentes.

iv)  $\frac{SQR}{\sigma^2}$  tem distribuição qui-quadrado  $\chi_{n-p}^2$  com  $n-p$  graus de liberdade.

#### 4.5. ANÁLISE DE VARIÂNCIA

A técnica mais usada para verificar a adequação do ajuste do modelo de regressão a um conjunto de dados é a *Análise de Variância* (sigla *ANOVA*) que se baseia na seguinte identidade:

$$\sum_i (y_i - \bar{y})^2 = \sum_i (\hat{\mu}_i - \bar{y})^2 + \sum_i (y_i - \hat{\mu}_i)^2 \quad (4.11)$$

O termo do lado esquerdo de (4.11) é a soma de quadrados das observações em relação ao seu valor médio e representa uma medida da variabilidade total dos dados. Esta soma será denotada por  $SQT = \sum_i (y_i - \bar{y})^2$ . O primeiro termo do lado de direito de (4.11) é a soma de quadrados explicada pelo modelo de regressão sendo denotada por  $SQE = \sum_i (\hat{\mu}_i - \bar{y})^2$ , enquanto o segundo termo é a soma de quadrados residual  $SQR = \sum_i (y_i - \hat{\mu}_i)^2$ , que não é explicada pelo modelo de regressão. O modelo será tanto melhor ajustado quanto maior for a variação explicada  $SQE$  em relação à variação total  $SQT$ . A dedução da equação (4.11) decorre elevando ao quadrado os termos da igualdade  $y_i - \bar{y} = (\hat{\mu}_i - \bar{y}) + (y_i - \hat{\mu}_i)$  e somando sobre as observações. Temos,

$$\sum_i (y_i - \bar{y})^2 = \sum_i (\hat{\mu}_i - \bar{y})^2 + \sum_i (y_i - \hat{\mu}_i)^2 + 2 \sum_i \left( \hat{\mu}_i - \bar{y} \right) \left( y_i - \hat{\mu}_i \right)$$

Cordeiro e Lima Neto (2004) mostram que o último termo é zero,  $\sum_i (\hat{\mu}_i - \bar{y})(y_i - \hat{\mu}_i) = 0$ .

As somas de quadrados explicada,  $SQE = \sum_i (\hat{\mu}_i - \bar{y})^2$ , e não-explicada, pela regressão  $SQR = \sum_i (y_i - \hat{\mu}_i)^2$ , podem ser escritas em notação matricial como:

$SQE = \hat{\beta}^T X^T y - n \bar{y}^2$  e  $SQR = y^T (I - H)y$ . Podemos medir a adequação do ajuste do modelo comparando a soma de quadrados residual  $SQR$  (esperando que seja pequena) com a



soma de quadrados devida à regressão  $SQE$ , ou, alternativamente, comparando  $SQE$  com a soma de quadrados total  $SQT = y^T y - n\bar{y}^2$ . A razão desses dois termos é representada por

$$R^2 = \frac{SQE}{SQT} = \frac{\hat{\beta}^T X^T y - n\bar{y}^2}{y^T y - n\bar{y}^2} \quad (4.12)$$

$R^2$  é denominado *coeficiente de determinação*. Alguns pesquisadores se baseiam erroneamente apenas no valor de  $R^2$  para escolher o melhor modelo. Entretanto, mais importante que termos um  $R^2$  próximo de um, é que a estimativa de  $\sigma^2$  seja também pequena, pois os intervalos de confiança para os parâmetros de interesse são proporcionais a  $\sigma$ .

A equação (4.11) em forma matricial é dada por

$$SQT = SQE + SQR = \hat{\beta}^T X^T y - n\bar{y}^2 + y^T (I - H)y,$$

que é a equação básica de construção da *Tabela de Análise de Variância*. A cada soma de quadrados nesta fórmula está associado um número de graus de liberdade, que é formalmente obtido expressando a soma de quadrados correspondente em forma quadrática, cujo posto é

igual ao número de graus de liberdade. As somas  $\frac{SQE}{\sigma^2} = \frac{\hat{\beta}^T X^T y - n\bar{y}^2}{\sigma^2}$  e  $\frac{SQR}{\sigma^2} = \frac{y^T (I - H)y}{\sigma^2}$  têm distribuições  $\chi_{p-1}^2$  e  $\chi_{n-p}^2$ , respectivamente, que são independentes.

A Tabela 2 apresenta o Quadro de Análise de Variância usada para testar a adequação global do modelo de Regressão  $y = X\beta + \varepsilon$ . Testamos a adequação global do modelo ajustado comparando a estatística  $F = \frac{MQE}{MQR}$  obtida desta tabela com o ponto crítico  $F_{p-1, n-p}(\alpha)$  da distribuição  $F$  de Snedecor com graus de liberdade  $p-1$  e  $n-p$ , respectivamente, supondo um nível significância  $\alpha$ . Se o valor da estatística  $F$  for superior ao ponto crítico, i.e.,  $F > F_{p-1, n-p}(\alpha)$ , o modelo é significativo para explicar a variabilidade da variável resposta. Caso contrário, o modelo não é significativo.

Tabela 2. Tabela de Análise de Variância.

<i>Efeito</i>	<i>Soma de Quadrados</i>	<i>GL</i>	<i>Média de Quadrados</i>	<i>Estatística</i>
Regressão	$SQE = \hat{\beta}^T X^T y - n\bar{y}^2$	$p-1$	$MQE = SQE/p-1$	$F = MQE/MQR$
Residual	$SQR = y^T (I - H)y$	$n-p$	$MQR = SQR/n - p$	
<b>Total</b>	$SQT = y^T y - n\bar{y}^2$	$n-1$		

#### 4.6. SELEÇÃO DE VARIÁVEIS EXPLICATIVAS

Depois do ajustamento preliminar de um modelo de regressão, temos interesse em selecionar as variáveis explicativas em um modelo ótimo parcimonioso para explicar os dados em questão. O teste F da análise de variância permite apenas inferir se algumas das variáveis explicativas são realmente importantes para explicar a variabilidade da variável resposta. Para selecionar as variáveis independentes, que são significativas, precisa determinar a distribuição dos estimadores dos parâmetros  $\beta$  e  $\sigma^2$  do modelo normal-linear (Cordeiro e Lima Neto, 2004).

Neste modelo, o estimador de mínimos quadrados  $\hat{\beta}_r$  tem distribuição normal  $N_p(\beta_r, \sigma^2 v_{rr})$ , onde  $v_{rr}$  é o elemento  $(r, r)$  da diagonal da matriz  $(X^T X)^{-1}$ . Como  $\hat{\beta}$  é independente de  $\hat{\sigma}^2$  e a distribuição de  $\hat{\sigma}^2$  é  $(n-p)^{-1} \sigma^2 \chi_{n-p}^2$ , a variável aleatória

$$T_r = \frac{\hat{\beta}_r - \beta}{\hat{\sigma} \sqrt{v_{rr}}}, \quad (4.13)$$

tem distribuição  $t$  de student, com  $n-p$  graus de liberdade.  $T_r$  é utilizada para testar se a variável explicativa  $x_r$ , correspondente a  $\beta_r$ , deve permanecer no modelo. Na prática, basta dividirmos o valor absoluto de  $\hat{\beta}_r$  pelo seu erro padrão, isto é,  $\hat{\sigma} \sqrt{v_{rr}}$ . Se este quociente for inferior ao valor crítico  $t_{n-p}(\alpha)$  da distribuição  $t$  de Student com  $n-p$  graus de liberdade, a variável independente  $x_r$  não é significativa para explicar a variabilidade da resposta e poderá ser eliminada do modelo; caso contrário,  $x_r$  é estatisticamente significativa para explicar o comportamento da variável resposta e, então, deve ser mantida no modelo.

#### 4.7. INTERVALOS DE CONFIANÇA

Os intervalos de confiança para os componentes de  $\beta$  ou regiões de confiança para subconjuntos e combinações lineares das componentes de  $\beta$  podem ser obtidos, respectivamente, utilizando os elementos da matriz  $(X^T X)^{-1}$ . Aqui trataremos apenas os intervalos de confiança para os componentes de  $\beta$ . Para maiores informações sobre como construir regiões de confiança para subconjuntos e combinações lineares das componentes de  $\beta$  favor consultar Cordeiro e Lima Neto (2004) e Demétrio e Zocchi (2007).

Da estatística pivotal definida em (4.13) podemos construir um intervalo de  $100(1-\alpha)\%$  de confiança para o verdadeiro valor  $\beta_r$ , a partir de

$$\hat{\beta}_r \pm \hat{\sigma} \sqrt{v_{rr}} t_{n-p}(\alpha/2). \quad (4.14)$$

Os sinais menos e mais definem os limites inferior e superior do intervalo, respectivamente. Se o valor de  $\sigma^2$  é conhecido podemos substituir os quantis  $t_{n-p}(\alpha/2)$  da distribuição  $t_{n-p}$  de student, com  $n-p$  graus de liberdade, pelos correspondentes quantis da distribuição normal reduzida ( $N(0,1)$ ).

#### 4.8. SELEÇÃO DE MODELOS

Na literatura existem vários procedimentos para a seleção de modelos de regressão. Os procedimentos mais conhecidos são: maior  $R_p^2$ , menor  $s_p^2$ ,  $C_p$ , *forward*, *backward*, *stepwise* e *AIC* (vide, por exemplo, Neter *et al.*, 1996), além de outros métodos que usam computação intensiva (Paula, 2004). Os conceitos dos métodos *forward*, *backward*, *stepwise* e *AIC* são descritos a seguir, de modo sucinto.

##### MÉTODO STEPWISE

O procedimento constrói iterativamente uma seqüência de modelos de regressão pela adição ou remoção de covariáveis em cada etapa. Após duas variáveis terem sido incluídas no modelo, verificamos se a primeira não sai do modelo. O processo continua até que nenhuma

covariável seja incluída, ou seja, retirada do modelo. Geralmente adota-se  $0,15 \leq f_{entra}$  e  $f_{sai} \leq 0,25$ . Uma sugestão seria usar  $f_{entra} = f_{sai} = 0,20$ .

#### MÉTODO FORWARD

Este método está baseado no princípio de que as covariáveis devem ser adicionadas ao modelo, uma de cada vez, até que não haja mais candidatas a covariável que produzam aumento significativo na soma de quadrados da regressão.

#### MÉTODO BACKWARD

Esse algoritmo começa com todas as K candidatas a covariável no modelo. Então, a covariável com menor estatística parcial F é removida, se essa estatística F for insignificante. Ou seja, se  $f < f_{sai}$ . A seguir, o modelo com K - 1 covariáveis é ajustado e a próxima covariável para potencial eliminação é encontrada. O algoritmo termina quando nenhuma covariável a mais pode ser eliminada.

#### MÉTODO DE AKAIKE

O método proposto por Akaike (1974), basicamente, se diferencia dos procedimentos acima por ser um processo de minimização que não envolve testes estatísticos. A idéia básica é selecionar um modelo que seja parcimonioso, ou em outras palavras, que esteja bem ajustado e tenha um número reduzido de parâmetros. Como o máximo do logaritmo da função de verossimilhança  $L(\beta)$  cresce com o aumento do número de parâmetros do modelo, uma proposta razoável seria encontrar o modelo com menor valor para a função

$$AIC = -2L(\hat{\beta}) + 2p$$

em que  $p$  denota o número de parâmetros. No caso do modelo normal-linear é possível mostrar que  $AIC$  fica expresso, quando  $\sigma^2$  é desconhecido, na forma

$$AIC = n \log \left\{ D(y; \hat{\mu}) / n \right\} + 2p,$$

em que  $D(y; \hat{\mu}) = \sum_i^n (y_i - \hat{\mu}_i)^2 = SQR$ .

Estes métodos descritos anteriormente possuem algumas desvantagens. Tipicamente, eles tendem a identificar um particular conjunto de covariáveis, em vez de possíveis conjuntos igualmente bons para explicar a resposta. Esse fato impossibilita que dois ou mais conjuntos

de covariáveis igualmente bons sejam apresentados para o pesquisador, para a escolha do mais relevante em sua área de aplicação. Isto significa que esses métodos são automáticos e fazem com que o pacote estatístico escolha o melhor modelo. O importante é que o estatístico e o pesquisador tenham uma postura pró-ativa neste processo (Colosimo e Giolo, 2006).

#### 4.9. TÉCNICAS DE DIAGNÓSTICO

A análise de diagnóstico ou técnicas de diagnóstico são usadas para detectar problemas com o ajuste do modelo de regressão. Conforme Paula (2004), tal etapa tem longa data e iniciou-se com a análise de resíduos para detectar a presença de pontos extremos e avaliar a adequação da distribuição proposta para a variável resposta. Uma referência importante nesse tópico é o artigo de Cox e Snell (1968) em que é apresentada uma forma bastante geral de definir resíduos, usada até os dias atuais. Cordeiro e Lima Neto (2004) dizem que esses problemas são de três tipos:

- a) Presença de observações mal ajustadas (pontos aberrantes);
- b) Inadequação das suposições iniciais para os erros aleatórios  $\varepsilon_i$ 's e/ou para a estrutura das médias  $\mu_i$ 's;
- c) Presença de observações influentes.

Demétrio e Zocchi (2007) definem os problemas a) e c) como *falhas isoladas* e o problema b) como *falhas sistemáticas*. Essas falhas podem surgir de várias maneiras. Algumas possibilidades são:

- devido a erros grosseiros na variável resposta ou nas variáveis explanatórias, por medidas erradas ou registro da observação, ou ainda, erros de transcrição;
- observação proveniente de uma condição distinta das demais;
- modelo mal especificado (falta de uma ou mais variáveis explicativas, modelo inadequado etc);
- escala usada errada, talvez os dados sejam mais bem descritos após uma transformação do tipo logarítmica ou raiz quadrada, por exemplo;
- a parte sistemática ( $X\beta$ ) do modelo e a escala estão corretas, mas a distribuição da resposta tem uma cauda mais longa do que a distribuição normal.

As técnicas usadas para a verificação do ajuste de um modelo a um conjunto de dados podem ser formais ou informais. As informais se baseiam em exames visuais de gráficos para

a detecção de padrões, ou então, de pontos discrepantes. As formais envolvem aninhar o modelo sob pesquisa em uma classe maior de modelos pela inclusão de um parâmetro (ou vetor de parâmetros) extra. As mais usadas são baseadas nos testes da razão de verossimilhança e escore. Parâmetros extras podem aparecer devido a:

- Inclusão de uma variável adicional;
- Aninhamento de uma covariável  $X$  em uma família indexada por um parâmetro  $\gamma$ , sendo um exemplo a família de Box-Cox;
- Inclusão de uma variável construída;
- Inclusão de uma variável *dummy* tomando o valor 1 (um) para a unidade discrepante e 0 (zero) para as demais. Isso é equivalente a eliminar essa observação do conjunto de dados, a fazer a análise com a observação discrepante e sem ela e verificar se a mudança no valor da soma de quadrados (*deviance*) é significativa, ou não. Ambos, porém, dependem da localização dos(s) pontos(s) discrepante(s).

Dado um conjunto de dados e ajustado um determinado modelo a ele, para a verificação das pressuposições devem ser considerados como material básico:

- os valores estimados (ou ajustados)  $\hat{\mu}_i = \hat{y}_i$ ;
- os resíduos  $r_i = \hat{y}_i - \hat{\mu}_i$ ;
- a variância residual estimada,  $\hat{\sigma}^2 = s^2 = MQR$ ;
- os elementos da diagonal (*leverage*) da matriz de projeção  $H$ ;

A matriz de projeção  $H$  – definida na Seção 4.2 – é muito usada nas técnicas de diagnóstico em regressão. Uma característica de grande importância da matriz  $H$  é inerente aos elementos  $h_{11}, \dots, h_{mm}$  da sua diagonal. O elemento  $h_{ii}$  mede o quão distante a observação  $y_i$  está das demais  $n-1$  observações no espaço definido pelas variáveis explicativas do modelo. O elemento  $h_{ii}$  só depende dos valores das variáveis explicativas, isto é, da matriz  $X$ , e não envolve as observações em  $y$ . O elemento  $h_{ii}$  representa uma *medida de alavancagem* da  $i$ -ésima observação. Se  $h_{ii}$  é grande os valores das variáveis explicativas associados a  $i$ -ésima observação são atípicos, ou seja, estão distantes do vetor de valores médios das variáveis explicativas. Uma observação com  $h_{ii}$  grande poderá ter influência na determinação dos coeficientes de regressão. Paula (2004) mostra que  $h_{ii}$  representa a variação no valor predito da  $i$ -ésima observação quando o valor observado é acrescido de um infinitésimo. Belsley *et al.*

(1980) sugerem  $h_{ii} \geq 2p/n$  como um indicador de pontos de alta alavancagem que requerem uma investigação adicional. Esta regra funciona bem na prática embora, em geral, irá detectar muitas observações de grande alavancagem. Assim, outras medidas de diagnóstico serão sempre necessárias para confirmar esse primeiro diagnóstico (Cordeiro e Lima Neto, 2004).

Uma idéia importante, também, é a da deleção (*deletion*), isto é, a comparação do ajuste do modelo escolhido, considerando-se todos os pontos, com o ajuste do mesmo modelo sem os pontos atípicos. As estatísticas obtidas pela omissão de certo ponto  $i$  são denotadas com um índice entre parênteses. Assim, por exemplo,  $s_{(i)}^2$  representa a variância residual estimada para o modelo ajustado, excluído o ponto  $i$  (Demétrio e Zocchi, 2007). Paula (2004) cita que um problema que pode ocorrer com a deleção individual de pontos, é o que se denomina *masking effect*, ou seja, deixar de detectar pontos conjuntamente discrepantes.

#### 4.9.1. Resíduos

É importante destacar que os resíduos são valores que possuem papel fundamental na verificação do ajuste de um modelo.

Dos resultados descritos anteriormente temos que  $E(r) = 0$  e  $Cov(r) = \sigma^2(I - H)$ . Isto é,  $r_i$  tem distribuição normal de média zero e variância  $Var(r_i) = \sigma^2(1 - h_{ii})$ . Além disso  $Cov(r_i, r_j) = -\sigma^2 h_{ij}$ . Como os  $r_i$ 's são heterocedásticos (têm variâncias diferentes), é conveniente expressá-los em forma padronizada a fim de permitir uma comparação entre os mesmos. Uma definição natural seria dividir  $r_i$  pelo respectivo desvio padrão, obtendo-se o resíduo estudentizado internamente (resíduo padronizado ou *Studentized residual*)

$$r_i^* = \frac{y_i - \hat{\mu}_i}{\hat{\sigma}(1 - h_{ii})^{1/2}} = \frac{r_i}{\hat{\sigma}(1 - h_{ii})^{1/2}}, i = 1, \dots, n,$$

em que  $\hat{\sigma}^2 = MQR = \sum_i^n r_i^2 / (n - p)$  é a estimativa de  $\sigma^2$  e o denominador

$Var(\hat{r}_i) = \hat{\sigma}^2(1 - h_{ii}) = MQR(1 - h_{ii})$  é um estimador não tendencioso para  $Var(r_i) = \sigma^2(1 - h_{ii})$ . A vantagem dos resíduos padronizados é que se o modelo (4.1) está correto, todos os resíduos têm a mesma variância, mesmo não sendo independentes, e eles são

mais sensíveis do que  $r_i$  por considerarem variâncias distintas. As observações cujos valores absolutos dos resíduos padronizados são maiores do que 2 (dois) podem ser consideradas mal-ajustadas (*pontos aberrantes*). Estes resíduos são também, apropriados para verificar a normalidade dos erros e a homogeneidade das variâncias.

No entanto como  $r_i$  não é independente de  $\hat{\sigma}^2$ ,  $r_i^*$  não segue uma distribuição  $t$  de student como se poderia esperar. Este problema de dependência entre  $r_i^*$  e  $\hat{\sigma}^2$  pode ser contornado substituindo  $\hat{\sigma}^2$  por  $\hat{\sigma}_{(i)}^2$  (média dos quadrados residual livre da influência da observação  $i$ ). Definiremos o resíduo estudentizado externamente (*jackknifed residuals, deletion residuals, externally studentized residual, RStudent*), como:

$$t_i = \frac{y_i - \hat{\mu}_{(i)}}{\hat{\sigma}_{(i)}(1-h_{ii})^{1/2}} = \frac{r_i}{\hat{\sigma}_{(i)}(1-h_{ii})^{1/2}}$$

Paula (2004) mostra que  $t_i = r_i^* \sqrt{\frac{n-p-1}{n-p-(r_i^*)^2}}$ , sendo que  $p$  é o número de parâmetros

independentes. A vantagem de usar  $t_i$  é que, sob normalidade, ele tem distribuição  $t$  de student com  $(n-p-1)$  graus de liberdade. Estes resíduos podem ser usados para testar se há diferenças significativas entre os valores ajustados obtidos *com* e *sem* a  $i$ -ésima observação (Cordeiro e Lima Neto, 2004). Embora não seja recomendada a prática de testes de significância na análise de resíduos, sugere-se que a  $i$ -ésima observação seja merecedora de atenção especial se  $|t_i|$  for maior do que o  $100\left(1 - \frac{\alpha}{2n}\right)$ -ésimo percentil da distribuição  $t$  de student com  $(n-p-1)$  graus de liberdade, sendo que  $\alpha$ , o nível de significância, é dividido por  $n$  por ser este o número de pontos na análise.

#### 4.9.2. Influência

No modelo de regressão é fundamental conhecer o grau de dependência entre o modelo ajustado e o vetor de observações  $y$ . Será preocupante se pequenas perturbações nestas observações produzirem mudanças bruscas nas estimativas dos parâmetros do modelo. Entretanto, se tais observações não alterarem os principais resultados do ajustamento pode-se confiar mais no modelo proposto, mesmo desconhecendo o verdadeiro processo que descreve



o fenômeno em estudo. As técnicas mais conhecidas para detectar esse tipo de influência são baseadas na exclusão de uma única observação e procuram medir o impacto dessa perturbação nas estimativas dos parâmetros (Cordeiro e Lima Neto, 2004). Discrepâncias isoladas (pontos atípicos) podem ser caracterizadas por ter  $h$  e/ou resíduos grandes, ser inconsistente e/ou ser influente (Demétrio e Zocchi, 2007). Em geral, pode-se classificar uma observação como:

- **ponto de alavanca (bom ou ruim)** –  $h$  alto;
- **inconsistente** – o ponto não segue a tendência dos dados;
- **outlier** – resíduo grande e  $h$  pequeno;
- **influyente** – afeta, de forma significativa, o ajuste do modelo.

Assim, Demétrio e Zocchi (2007) dizem que uma observação influente é aquela cuja omissão do conjunto de dados resulta em mudanças substanciais em certos aspectos do modelo. Ela pode ser um *outlier*, ou não. Uma observação pode ser influente de diversas maneiras, isto é,

- no ajuste geral do modelo,
- no conjunto de estimativas dos parâmetros,
- na estimativa de um determinado parâmetro,
- na escolha de uma transformação da variável resposta ou de uma variável explanatória.

As estatísticas mais utilizadas para a verificação de pontos atípicos são:

- a) **Elementos da diagonal da matriz de projeção  $H$  ( $h_{ii}$ , *leverage*)** – Como visto anteriormente, quando uma observação está distante das outras em termos das variáveis explicativas ela pode ser, ou não, influente. A distância de uma observação em relação às demais é medida pelo  $h$  (medida de *leverage*). Segundo Belsley *et al.* (1980), valores de  $h_{ii} \geq 2p/n$  indicam observações que merecem uma análise mais apurada.
- b) **Distância de Cook** – É também uma medida de afastamento do vetor de estimativas provocado pela retirada da observação  $i$ .

$$D_i = (r_i^*)^2 \frac{h_{ii}}{1-h_{ii}} \frac{1}{p}.$$

Paula (2004) fala que a medida  $D_i$  poderá não ser adequada quando o resíduo padronizado  $r_i^*$  for grande e  $h_{ii}$  for próximo de zero. Nesse caso,  $\hat{\sigma}^2$  pode ficar inflacionada, e não ocorrendo nenhuma compensação por parte de  $h_{ii}$ ,  $D_i$  pode ficar

pequeno. As observações serão consideradas influentes quando  $D_i \geq F_{p, n-p}(0.5)$  e recomenda-se examinar as conseqüências da retirada dessas observações no ajustamento do modelo. Como, para a maioria das distribuições  $F$ , o quantil de 50% é próximo de um, sugere-se na prática que se o maior valor de  $D_i$  for muito inferior a 1 (um), então a eliminação de qualquer observação do modelo não irá alterar muito as estimativas dos parâmetros. Entretanto, para investigar mais detalhadamente a influência das observações com maiores valores de  $D_i$ , o analista terá que eliminar estas observações e recalculas as estimativas dos parâmetros (Cordeiro e Lima Neto, 2004). Uma medida, supostamente mais apropriada, foi proposta por Belsley *et al.* (1980), sendo conhecida como *DFFITs*.

c) **DFFITs** – Esta quantidade mede a alteração provocada no valor ajustado pela retirada da observação  $i$ . É definida por

$$DFFITs_i = t_i \left\{ \frac{h_{ii}}{p(1-h_{ii})} \right\}^{1/2}.$$

No caso da estatística  $DFFITs_i$ , os pontos influentes são aqueles em que  $DFFITs_i \geq 2\{p/(n-p)\}^{1/2}$ . Os comentários para a estatística  $D_i$  permanecem válidos para a estatística  $DFFITs_i$ . Geralmente, examinamos as estatísticas  $D_i$  e  $DFFITs_i$  graficamente, dando atenção àquelas observações cujas medidas têm maiores valores.

d) **Distância de Cook modificada** – Atkinson (1981) sugere uma modificação para a distância de Cook

$$C_i = t_i \left( \frac{n-p}{p} \frac{h_{ii}}{1-h_{ii}} \right)^{1/2} |t_i| = \left( \frac{n-p}{p} \right)^{1/2} DFFITs_{(i)}.$$

A vantagem de  $C_i$  é que a mesma pode ser utilizada em gráficos normais de probabilidades (Paula, 2004).

### 4.9.3. Técnicas gráficas

De uma forma geral os problemas de diagnóstico mencionados no início da Seção 4.9, podem ser detectados, respectivamente, através das seguintes técnicas gráficas:

a) um gráfico dos resíduos padronizados  $r_i^*$  versus a ordem das observações para detectar as observações aberrantes;

b) um gráfico dos resíduos padronizados  $r_i^*$  versus os valores ajustados  $\hat{\mu}_i$ . Neste gráfico os pontos devem estar aleatoriamente distribuídos entre as duas retas  $y = -2$  e  $y = 2$ , paralelas ao eixo horizontal, sem exibir uma forma definida. Se, neste gráfico, os pontos exibirem algum padrão, isto pode ser indicativo de heterocedasticidade da variância dos erros ou da não-linearidade dos efeitos das variáveis explicativas nas médias das observações;

c) um gráfico de probabilidades dos resíduos padronizados  $r_i^*$  ordenados versus os quantis da distribuição normal reduzida. Neste gráfico, se os pontos ficarem praticamente dispostos sobre uma reta, as observações podem ser consideradas como tendo, aproximadamente, distribuição normal. Demétrio e Zocchi (2007) citam alguns formatos aproximados comuns que indicam ausência de normalidade:

- **S (Esse)** – indica distribuições com cauda muito curtas, isto é, distribuições cujos valores estão muito próximos da média,
- **S (Esse invertido)** – indica distribuições com caudas muito longas e, portanto, presença de muitos valores extremos,
- **J e J invertido** – indicam distribuições assimétricas, positivas e negativas, respectivamente.

d) gráficos de  $h_{ii}$ ,  $D_i$  e  $DFFITs_i$  versus a ordem das observações para detectar as observações influentes.

#### 4.10. TRANSFORMAÇÃO BOX-COX

O uso do modelo clássico de regressão é justificado admitindo-se:

- i) linearidade da estrutura de  $E(y)$ ;
- ii) variância constante do erro (homocedasticidade),  $Var(y) = \sigma^2$ ;
- iii) normalidade;
- iv) independência das observações.

Se as suposições i) a iii) não são satisfeitas para os dados originais, uma transformação não-linear de  $y$  poderá verificá-las, pelo menos aproximadamente. Em alguns problemas de

regressão deve-se transformar tanto a variável dependente quanto as variáveis explicativas para que as suposições acima sejam satisfeitas (Cordeiro e Lima Neto, 2004).

Cordeiro e Lima Neto (2004) comentam que as dificuldades com o modelo clássico de regressão não só ocorrem devido à violação de uma das hipóteses básicas. Muitas vezes são devidas à problemas fora do contexto da forma dos dados, como por exemplo, a multicolinearidade, quando existem relações aproximadamente lineares entre as variáveis explicativas. Outro tipo de dificuldade ocorre quando se dispõe de um grande número de variáveis explicativas e, portanto, surge um problema de ordem combinatória para selecionar o modelo. Maiores detalhes em Cordeiro e Lima Neto (2004).

A seguir introduzimos a transformação de Box e Cox que tem por objetivo transformar a variável dependente para satisfazer as hipóteses i) a iv) do modelo clássico de regressão. A transformação de Box e Cox (1964) supõe que os dados  $y = (y_1, \dots, y_n)^T$  são independentes e que existe um escalar  $\lambda$  tal que os dados transformados por

$$z = z(\lambda) = \begin{cases} (y^\lambda - 1)/\lambda, & \text{se } \lambda \neq 0 \\ \ln y, & \text{se } \lambda = 0 \end{cases} \quad (4.15)$$

satisfazem  $E(z) = \mu = X\beta$ ,  $Var(z_i) = \sigma^2$  para  $i = 1, \dots, n$  e  $z \sim N(\mu, \sigma^2 I)$ . A transformação (4.15) tem vantagem sobre a transformação potência simples  $y^\lambda$  por ser contínua em  $\lambda = 0$ . Apesar de o modelo admitir a existência de um único  $\lambda$  produzindo linearidade dos efeitos sistemáticos, normalidade e variância constante dos dados transformados, pode ser que diferentes valores de  $\lambda$  sejam necessários para alcançar tudo isso (Cordeiro e Lima Neto, 2004).

## 5. MODELAGEM ATRAVÉS DO MODELO LINEAR USUAL

Foi visto que num modelo de regressão linear utilizam-se variáveis regressoras (covariáveis) que podem influenciar a variabilidade da variável resposta. Uma forma de verificar tal influência é através dos intervalos de confiança assintóticos para os coeficientes estimados do modelo relacionados às essas variáveis. Aqui apresentamos um estudo de simulação para verificar a probabilidade de cobertura de seis parâmetros de um modelo usual de regressão linear considerando 5 tamanhos diferentes de amostras. Para cada amostra, verifica se o intervalo de confiança cobre o verdadeiro valor do parâmetro ou não. Assim calculamos a proporção de vezes em que o intervalo de confiança cobriu o verdadeiro valor dos parâmetros utilizados na geração das amostras.

### 5.1. INTERVALOS DE CONFIANÇA

Intervalos de confiança para os parâmetros podem ser baseados na distribuição normal assintótica. O intervalo assintótico pode ser construído utilizando os estimadores de máxima verossimilhança (EMV) e suas variâncias estimadas.

A utilização desse intervalo é direcionada pelo tamanho da amostra, que deve ser suficientemente grande.

Assim, os intervalos de de 95% confiança são dados por:

$$\hat{\theta}_i \pm 1.96\sqrt{\text{var}(\hat{\theta}_i)}$$

O intervalo de 95% de confiança para a probabilidade de cobertura nominal é dada por

$$0.95 \pm 1.96\sqrt{\frac{(0.95*0.05)}{1000}}$$

Este intervalo fornece os limites inferior e superior da probabilidade de cobertura de cada parâmetro. Se a probabilidade de cobertura de um parâmetro estiver entre esses limites temos um indicativo de uma boa cobertura para tal parâmetro.

## 5.2. ESTUDO DE SIMULAÇÃO PARA O MODELO DE REGRESSÃO LINEAR USUAL

O estudo de simulação realizado baseou-se na geração de 1000 conjuntos de dados trabalhando com amostras de tamanhos: 25, 50, 150, 500 e 1000.

Seja  $V$  a variável aleatória que representa o valor total, em reais (R\$), do imóvel. Esta variável apresenta somente valores positivos e é assimétrica acarretando a não-normalidade. Em geral, preços carregam o problema de heterocedasticidade, imóveis de alto preço têm maior variabilidade em relação aos imóveis de baixo preço. Os profissionais que trabalham na área de avaliação de imóveis estão acostumados a utilizar a transformação Box-Cox para atacar o problema de não-normalidade. Ao realizar qualquer transformação, pertencente a esta família, para os dados originais, conseguimos obter, na maioria dos casos, valores transformados que apresentam normalidade ou no mínimo normalidade aproximada.

Assim, aplicando qualquer transformação pertencente à família Box-Cox, como por exemplo, a transformação raiz quadrada da variável resposta  $V$  (valor total (R\$) do imóvel),  $Y' = \sqrt{V}$ , temos que, aproximadamente,  $Y' \sim N(\mu, \sigma^2)$ .

Aguirre (1997) mostra que após aplicar a transformação raiz quadrada da variável resposta, é necessário retransformar os valores calculados com as regressões para voltar às unidades originais do problema. Ao realizar esses cálculos obtemos exatamente o mesmo valor para a variável resposta. Assim, geramos a variável resposta da simulação,  $Y'$ , a partir de uma distribuição normal com média igual ao preditor linear da equação de regressão  $(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5)$  e fixamos a variância de três maneiras diferentes:  $10^2$  (baixa),  $100^2$  (variância moderada) e  $1000^2$  (alta), onde quatro das covariáveis ( $x_1, x_2, x_3$  e  $x_4$ ) foram geradas a partir de uma distribuição bernoulli, devido a dicotomia das covariáveis de localização presentes no nosso estudo, e uma única variável ( $x_5$ ), representando a área, foi gerada a partir de uma distribuição Weibull. Devido à presença de assimetria na distribuição dos valores de área dos lotes observados em São Carlos, SP, optamos pela distribuição Weibull para simular a área do lote sendo que esta distribuição pode apresentar assimetria. Veja a seguir:  $x_1 \sim \text{bernoulli}(0.4)$ ;  $x_2 \sim \text{bernoulli}(0.7)$ ;  $x_3 \sim \text{bernoulli}(0.5)$ ;  $x_4 \sim \text{bernoulli}(0.3)$ ;  $x_5 \sim \text{Weibull}(3, 800)$ , onde o valor 3 é o valor do parâmetro de forma e 800 o valor do parâmetro de escala da distribuição Weibull.

Para cada tamanho de amostra foi verificado se o intervalo de confiança cobria o verdadeiro valor do parâmetro ou não, ou seja, foi calculada a proporção de vezes em que o intervalo de confiança cobriu os 1000 conjuntos de dados. A discrepância entre um valor estimado e o valor real foi medida por  $|\hat{\beta} - \beta|$ , onde  $\hat{\beta}$  é o valor estimado e  $\beta$  é o valor real.

Considerando todas as situações consideradas acima, tem-se a tabela e os gráficos a seguir:

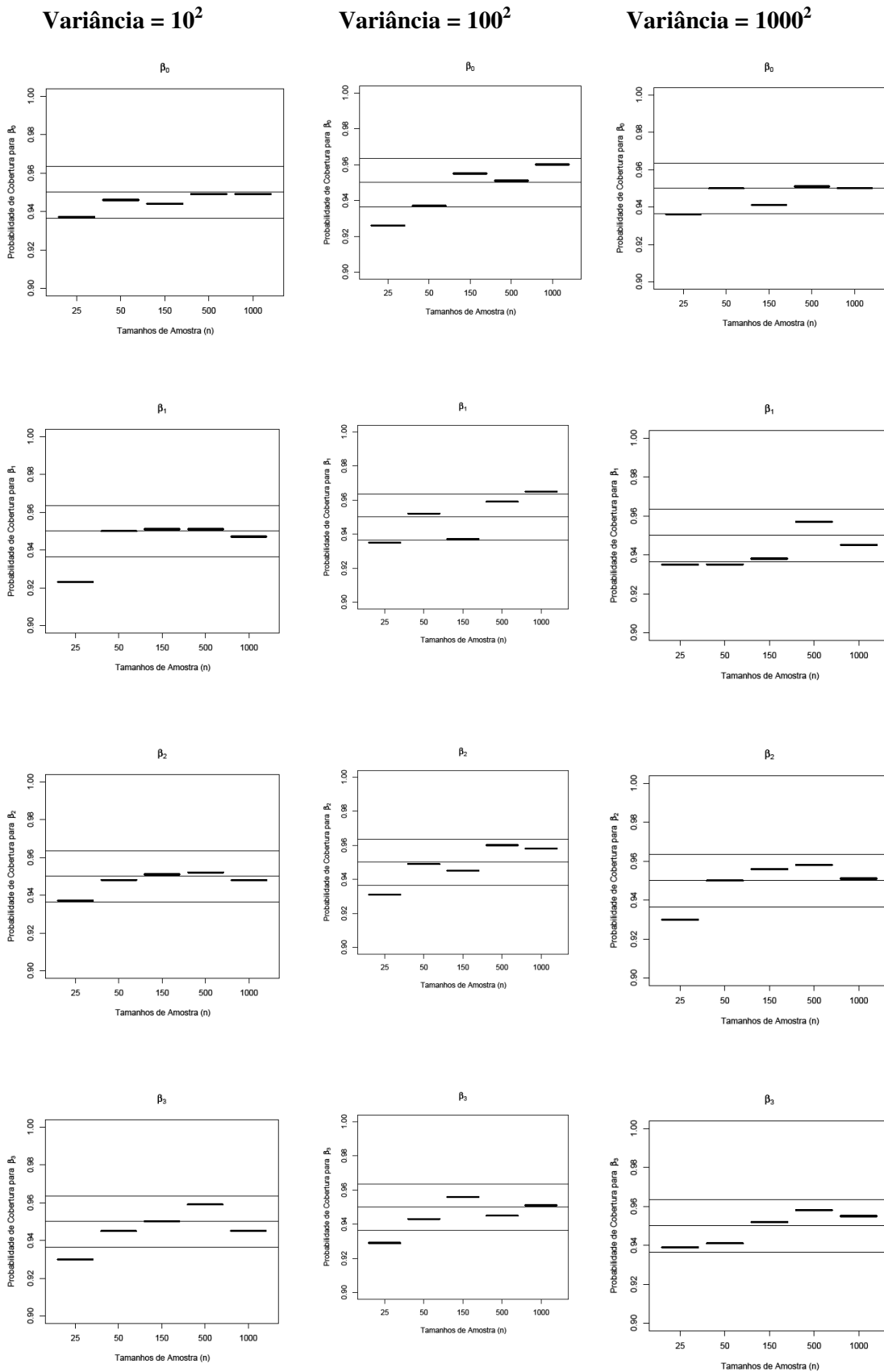
Tabela 3. Probabilidades de cobertura nominais das estimativas dos parâmetros para os 5 tamanhos diferentes de amostra.

Variância	Parâmetro	n=25	n=50	n=150	n=500	n=1000
$10^2$	$\beta_0$	0.937	0.946	0.944	0.949	0.949
$100^2$	$\beta_0$	0.926	0.937	0.955	0.951	0.96
$1000^2$	$\beta_0$	0.936	0.95	0.941	0.951	0.95
$10^2$	$\beta_1$	0.923	0.95	0.951	0.951	0.947
$100^2$	$\beta_1$	0.935	0.952	0.937	0.959	0.965
$1000^2$	$\beta_1$	0.935	0.935	0.938	0.957	0.945
$10^2$	$\beta_2$	0.937	0.948	0.951	0.952	0.948
$100^2$	$\beta_2$	0.931	0.949	0.945	0.96	0.958
$1000^2$	$\beta_2$	0.93	0.95	0.956	0.958	0.951
$10^2$	$\beta_3$	0.93	0.945	0.95	0.959	0.945
$100^2$	$\beta_3$	0.929	0.943	0.956	0.945	0.951
$1000^2$	$\beta_3$	0.939	0.941	0.952	0.958	0.955
$10^2$	$\beta_4$	0.941	0.945	0.948	0.947	0.954
$100^2$	$\beta_4$	0.93	0.943	0.939	0.954	0.952
$1000^2$	$\beta_4$	0.929	0.953	0.942	0.953	0.952
$10^2$	$\beta_5$	0.942	0.945	0.948	0.95	0.944
$100^2$	$\beta_5$	0.929	0.938	0.958	0.945	0.951
$1000^2$	$\beta_5$	0.934	0.941	0.949	0.967	0.943

Tabela 4. Discrepâncias das estimativas dos parâmetros para os 5 tamanhos diferentes de amostra.

Variância	Parâmetro	n=25	n=50	n=150	n=500	n=1000
$10^2$	$\beta_0$	0.642	0.122	0.018	0.052	0.042
$100^2$	$\beta_0$	6.42	1.217	0.18	0.522	0.417
$1000^2$	$\beta_0$	64.230	12.173	1.804	5.216	4.175
$10^2$	$\beta_1$	0.065	0.174	0.024	0.003	0.007
$100^2$	$\beta_1$	0.648	1.743	0.238	0.029	0.072
$1000^2$	$\beta_1$	6.483	17.425	2.382	0.287	0.724
$10^2$	$\beta_2$	0.203	0.011	0.119	0.067	0.053
$100^2$	$\beta_2$	2.03	0.11	1.185	0.67	0.531
$1000^2$	$\beta_2$	20.298	1.097	11.855	6.698	5.309
$10^2$	$\beta_3$	0.094	0.085	0.035	0.036	0.003
$100^2$	$\beta_3$	0.943	0.846	0.353	0.365	0.034
$1000^2$	$\beta_3$	9.434	8.464	3.534	3.649	0.344
$10^2$	$\beta_4$	0.172	0.088	0.019	0.005	0.037
$100^2$	$\beta_4$	1.721	0.881	0.191	0.051	0.37
$1000^2$	$\beta_4$	17.208	8.809	1.907	0.510	3.705
$10^2$	$\beta_5$	0.001	0.000	0.000	0.000	0.000
$100^2$	$\beta_5$	0.009	0.002	0.001	0.001	0.000
$1000^2$	$\beta_5$	0.089	0.017	0.007	0.006	0.002





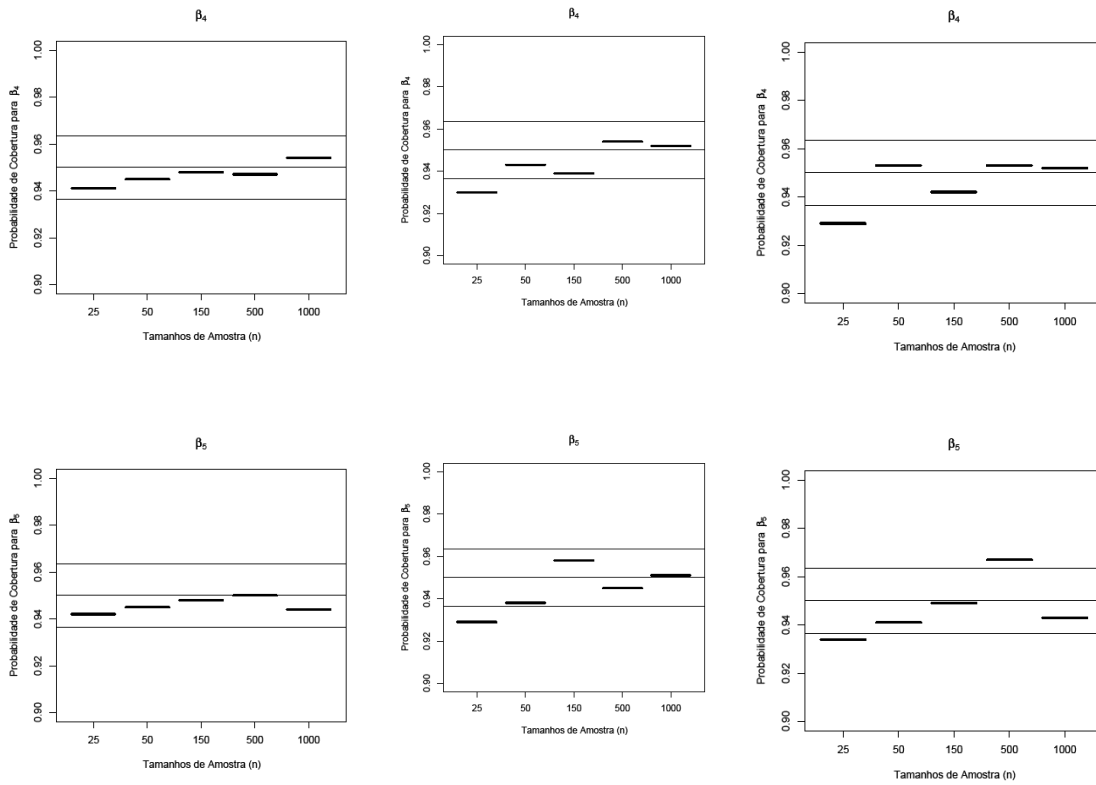
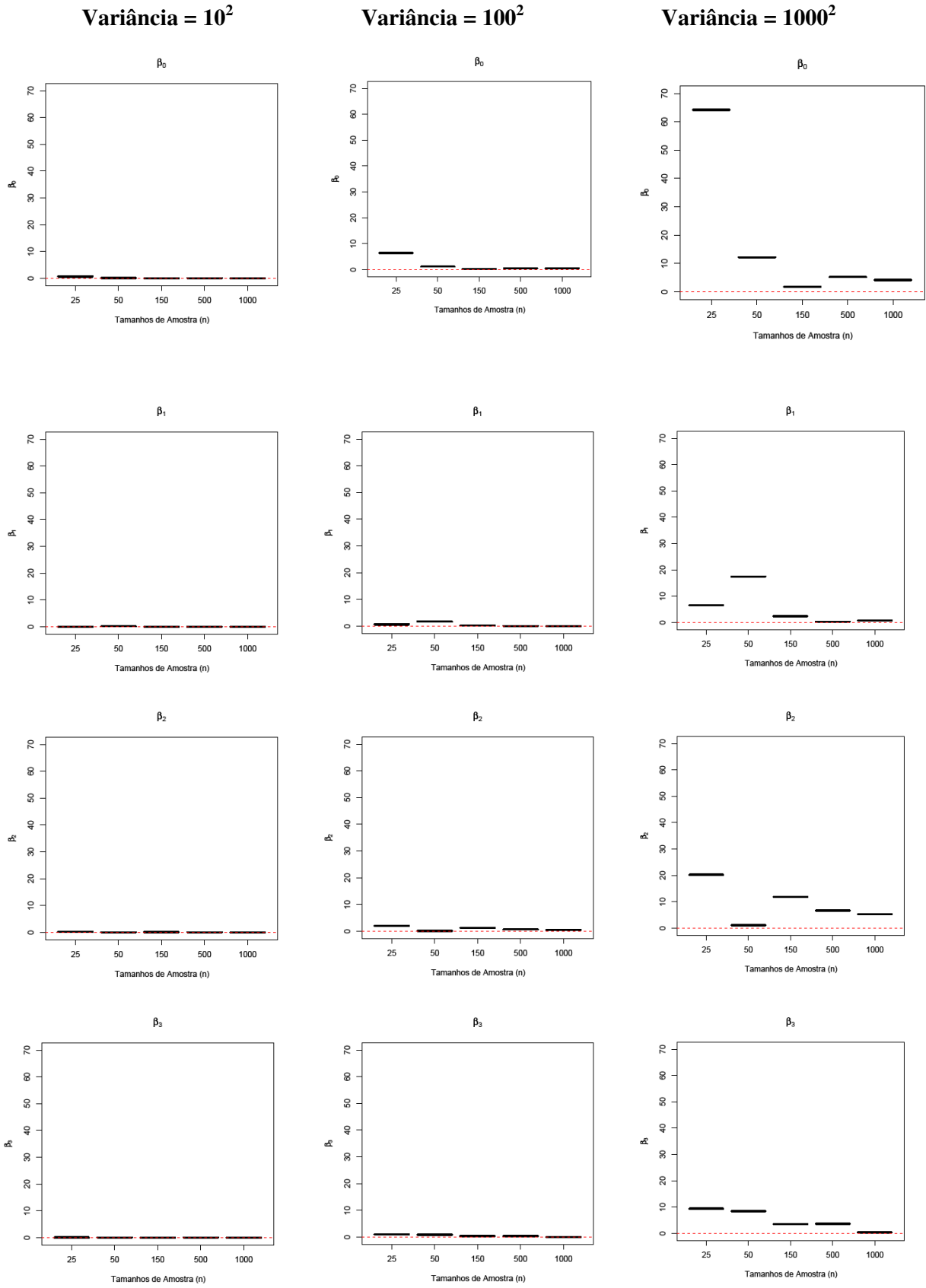


Figura 4. Limite inferior nominal, limite superior nominal e valor central nominal (95%) dos intervalos assintóticos normais das probabilidades de cobertura nominais das estimativas dos parâmetros para os 5 tamanhos diferentes de amostra e 3 diferentes variâncias.



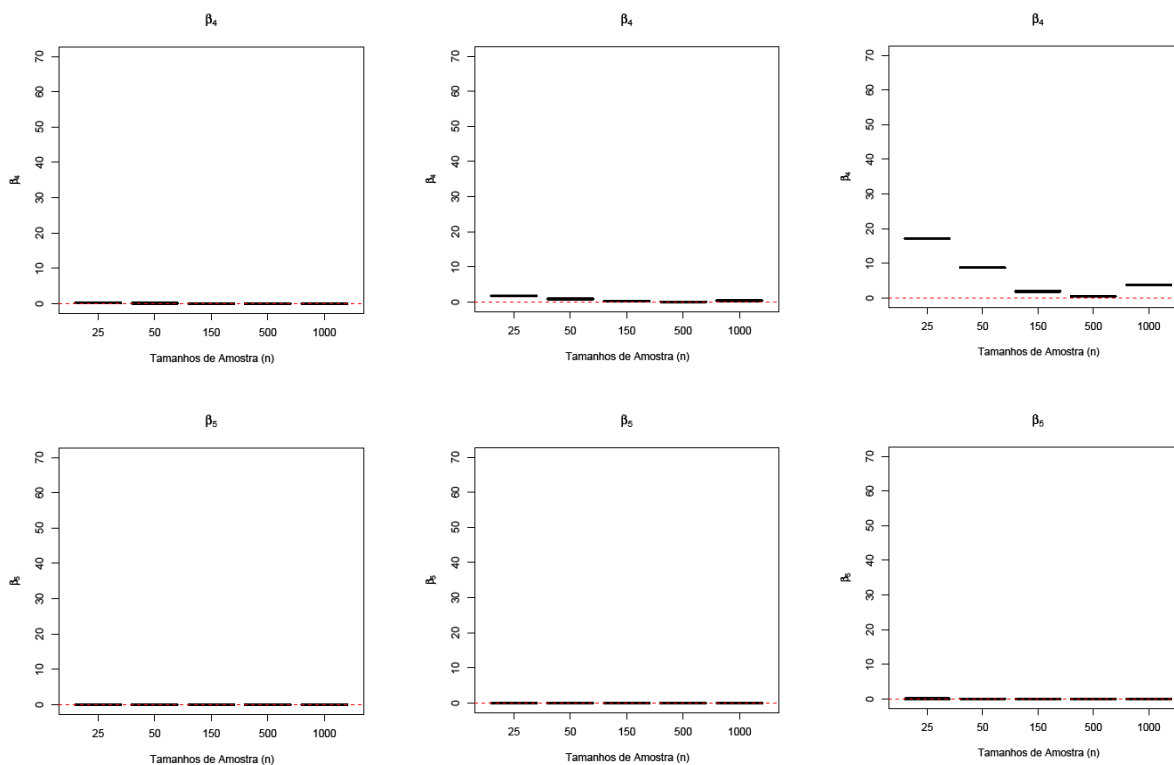


Figura 5. Discrepâncias das estimativas dos parâmetros para os 5 tamanhos diferentes de amostra e 3 diferentes variâncias.

Através da verificação da Tabela 3 e da Figura 4 pode ser visualizado que, quando se considera uma variância baixa,  $\beta_1$  e  $\beta_3$  apresentam cobertura dentro dos limites nominais a partir das amostras de tamanho 50 e os outros parâmetros a partir das amostras de tamanho 25. Ao considerar uma variância moderada, todos os parâmetros apresentam cobertura dentro dos limites nominais a partir das amostras de tamanho 50. Quando a variância é alta  $\beta_2$ ,  $\beta_4$  e  $\beta_5$  apresentam cobertura dentro dos limites nominais a partir das amostras de tamanho 50 e  $\beta_1$  apresenta cobertura dentro dos limites nominais a partir das amostras de tamanho 150 e  $\beta_0$  e  $\beta_3$  apresentam cobertura dentro dos limites nominais a partir das amostras de tamanho 25. E, através da Tabela 4 e da Figura 5, verificou-se que a qualidade das estimativas dos parâmetros depende da variância amostral. Um aumento na variância amostral levou a uma maior discrepância das estimativas dos parâmetros, exceto para  $\beta_5$ . O  $\beta_5$ , parâmetro responsável pelo efeito da única variável numérica, apresentou a menor discrepância das estimativas, independente da variância e do tamanho amostral. Para os demais parâmetros, a discrepância diminuiu à medida que o tamanho da amostra aumentou.

Como esperado, pequenas amostras (tamanho 25) apresentaram baixa performance na simulação de probabilidade de cobertura dos parâmetros. Assim sendo, o pesquisador deverá sempre ficar atento ao tamanho de amostra a ser utilizado na sua pesquisa. O planejamento amostral feito com qualidade ajuda a contornar esse problema.

Pelo fato de se trabalhar com simulação, o valor do parâmetro que está sendo testado para verificação são valores amostrais, assim, a simulação foi realizada com a escolha de um valor coerente para verificar se este estava ou não contido nos intervalos.

### 5.3. APLICAÇÃO DO MODELO LINEAR USUAL

Para a construção do modelo, a variável resposta será o valor total do lote (R\$) e, como variáveis explicativas, as variáveis de localização e a área do lote (m<sup>2</sup>). Considerou-se apenas os lotes com área igual ou inferior a 800m<sup>2</sup> e comercializados efetivamente em 2005, ou seja, apenas os lotes cuja venda foi efetivamente realizada resultando numa amostra contendo 284 lotes. Para tratarmos a variável localização, o espaço urbano foi subdividido conforme descrito na Tabela 1. Quando o imóvel em avaliação pertenceu a uma determinada região foi atribuído o valor 1 e, caso contrário, o valor 0 (não pertenceu). Entretanto, esta abstração implica no problema do critério para a delimitação destas regiões. Esta maneira demonstra uma possibilidade de abstração da modelagem da realidade espacial, no âmbito do método científico para obtenção de uma avaliação, mais especificamente a embasada na análise de regressão múltipla não-espacial. Desta maneira teremos o seguinte modelo inicial composto pelas variáveis e pelas interações entre cada uma delas com a variável área do lote (m<sup>2</sup>):

$$V_i = \beta_0 + \beta_1 \text{AREA}_i + \beta_2 \text{NUC\_PRINC}_i + \beta_3 \text{PLN\_CENTRAL}_i + \beta_4 \text{FERROVIA}_i + \beta_5 \text{RODOVIA\_WL}_i + \beta_6 \text{ENCOSTA}_i + \beta_7 \text{CONDO}_i + \beta_8 \text{FECHADO}_i + \beta_9 \text{ESTRIT\_RESID}_i + \beta_{10} (\text{NUC\_PRINC}_i * \text{AREA}_i) + \beta_{11} (\text{PLN\_CENTRAL}_i * \text{AREA}_i) + \beta_{12} (\text{FERROVIA}_i * \text{AREA}_i) + \beta_{13} (\text{RODOVIA\_WL}_i * \text{AREA}_i) + \beta_{14} (\text{ENCOSTA}_i * \text{AREA}_i) + \beta_{15} (\text{CONDO}_i * \text{AREA}_i) + \beta_{16} (\text{FECHADO}_i * \text{AREA}_i) + \beta_{17} (\text{ESTRIT\_RESID}_i * \text{AREA}_i) + \varepsilon_i$$

Tratamos as variáveis de localização (*Núcleo Principal, Planície Central, Ferrovia, Rodovia Washington Luís, Condomínio, Condomínio Fechado e Estritamente Residencial*) como uma variável *dummy* sendo 1 para “pertence” e 0 para “não pertence”. As unidades da

variável resposta, Valor Total do Imóvel ( $V_i$ ), e da covariável, área do terreno, estão em reais (R\$) e  $m^2$ , respectivamente.

Inicialmente separamos o banco de dados em duas partes. Utilizamos 70% dos dados para o ajuste do modelo e 30 % para a validação.

O processo de modelagem seguiu as seguintes etapas:

i) Em geral, preços carregam o problema de heterocedasticidade, preços de imóveis de valores altos têm maior variabilidade em relação aos preços de imóveis de valores baixos. Desta maneira, foi necessária a aplicação da transformação de Box e Cox nos 70% dos dados separados para a construção do modelo;

ii) Utilizamos a metodologia de STEPWISE para selecionarmos o modelo que iniciaremos o nosso processo de modelagem. O modelo escolhido neste passo foi:

$$(V_i)^{1/2} = \beta_0 + \beta_1 \text{AREA}_i + \beta_2 \text{NUC\_PRINC}_i + \beta_3 \text{PLN\_CENTRAL}_i + \beta_4 \text{FERROVIA}_i + \beta_5 \text{RODOVIA\_WL}_i + \beta_6 \text{CONDO}_i + \beta_7 \text{FECHADO}_i + \beta_8 \text{ESTRIT\_RESID}_i + \beta_9 (\text{NUC\_PRINC}_i * \text{AREA}_i) + \beta_{10} (\text{FECHADO}_i * \text{AREA}_i) + \beta_{11} (\text{ESTRIT\_RESID}_i * \text{AREA}_i) + \beta_{12} (\text{PLN\_CENTRAL}_i * \text{AREA}_i) ;$$

iii) no modelo acima verificamos a significância estatística das variáveis. Nessa etapa, para obter o modelo final, devemos combinar a significância estatística com o interesse prático. O princípio da parcimônia deve guiar a nossa busca pelo modelo final;

iv) Após algum tempo, obtemos o seguinte modelo final:

$$(V_i)^{1/2} = \beta_0 + \beta_1 \text{NUC\_PRINC}_i + \beta_2 \text{PLN\_CENTRAL}_i + \beta_3 \text{FERROVIA}_i + \beta_4 \text{RODOVIA\_WL}_i + \beta_5 \text{CONDO}_i + \beta_6 \text{FECHADO}_i + \beta_7 \text{ESTRIT\_RESID}_i + \beta_8 (\text{NUC\_PRINC}_i * \text{AREA}_i) + \beta_9 (\text{FECHADO}_i * \text{AREA}_i) + \beta_{10} (\text{ESTRIT\_RESID}_i * \text{AREA}_i) + \beta_{11} (\text{PLN\_CENTRAL}_i * \text{AREA}_i),$$

e percebe-se que esse modelo é o mesmo modelo sugerido pelo método STEPWISE exceto pela variável ÁREA.

v) o próximo passo é o processamento das análises de diagnóstico para verificação das suposições do modelo proposto. Alguns pontos, considerados influentes, foram retirados das análises. Após a retirada desses pontos vê-se, na Figura 6, que as medidas de diagnósticos apresentaram os seguintes valores:  $h_{ii} < 0,5$ , a distância de Cook  $< 0,2$  e DFFITS  $< 1$ , indicando que não existe observação influente. O teste de Shapiro-Wilk, para a verificação da normalidade dos resíduos, não rejeita a normalidade ao nível de 10%.

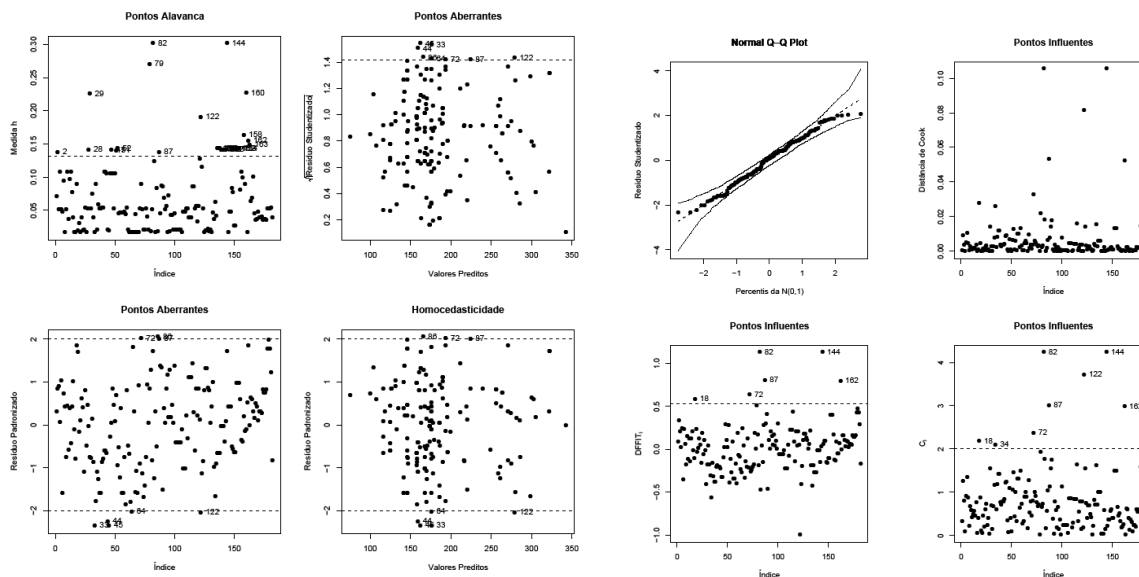


Figura 6. Gráficos de diagnóstico para o modelo normal com transformação raiz quadrada para a variável resposta.

Aqui terminamos a parte de ajuste do modelo. A próxima etapa é a validação do modelo proposto.

Assim, aplicando o modelo encontrado aos dados separados para validação apresentamos as seguintes estimativas dos parâmetros com os respectivos intervalos de confiança (95%) (Tabela 5). Desta maneira, o preditor linear final (“*pred*”) se apresenta como:

$$\begin{aligned} \text{pred} = & 96,71 - 81,05 \cdot \text{NUC\_PRINC} + 27,63 \cdot \text{CONDO} - 4,87 \cdot \text{PLN\_CENTRAL} + 17,98 \cdot \text{FERROVIA} + 42,94 \cdot \text{RODOVIA\_WL} - \\ & 91,28 \cdot \text{FECHADO} + 39,70 \cdot \text{ESTRIT\_RESID} + 0,33 \cdot (\text{NUC\_PRINC} \cdot \text{AREA}) + 0,08 \cdot (\text{PLN\_CENTRAL} \cdot \text{AREA}) - \\ & 0,13 \cdot (\text{ESTRIT\_RESID} \cdot \text{AREA}) + 0,28 \cdot (\text{FECHADO} \cdot \text{AREA}) \end{aligned}$$

Como foi aplicada a transformação raiz quadrada na variável resposta é necessário aplicar a transformação inversa para obtermos os valores preditos na escala da variável original. Para isso basta elevarmos o preditor linear ao quadrado, isto é,

$$\text{valor total do imóvel}(V_i) = (\text{pred})^2$$

A diferença relativa média para o modelo é de 25%, ou seja, a taxa de aceitação total é de 75%.

Tabela 5. Estimativa dos parâmetros, respectivos limites: inferior e superior, do intervalo de confiança de 95% e amplitude do intervalo, para cada covariável.

Variável	Estimativa dos parâmetros	LI	LS	Amplitude
Intercepto	96.71	66.47	127	60.48
NUC_PRINC	-81.05	-111.1	-51.04	60.03

<b>CONDO</b>	27.63	5.33	49.92	44.59
<b>PLN_CENTRAL</b>	-4.87	-45.94	36.2	82.14
<b>FERROVIA</b>	17.98	5.12	30.85	25.73
<b>RODOVIA_WL</b>	42.94	25.59	60.28	34.69
<b>FECHADO</b>	-91.28	-138.9	-43.68	95.21
<b>ESTRIT_RESID</b>	39.7	0.81	78.59	77.78
<b>NUC_PRINC*AREA</b>	0.33	0.26	0.41	0.15
<b>PLN_CENTRAL*AREA</b>	0.08	-0.03	0.18	0.21
<b>ESTRIT_RESID*AREA</b>	-0.13	-0.23	-0.02	0.21
<b>FECHADO*AREA</b>	0.28	0.14	0.42	0.28

A Figura 7 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, ou seja, estatisticamente não há distúrbios para não utilizar o modelo final para atender o objetivo geral. Portanto, a equação de regressão representativa da formação do valor de mercado do solo urbano do município de São Carlos, SP, através dos valores de seus lotes no ano de 2005, está representada pelo nosso modelo linear usual final dado por

$$V_i = [96.71 - 81.05*NUC\_PRINC_i - 4.87*PLN\_CENTRAL_i + 17.98*FERROVIA_i + 42.94*RODOVIA\_WL_i + 27.63*CONDO_i - 91.28*FECHADO_i + 39.7*ESTRIT\_RESID_i + 0.33*(NUC\_PRINC_i*AREA_i) + 0.28*(FECHADO_i*AREA_i) - 0.13*(ESTRIT\_RESID_i*AREA_i) + 0.08*(PLN\_CENTRAL_i*AREA_i)]^2,$$

Este modelo poderá ser utilizado como ferramenta na elaboração da PVG do município de São Carlos, SP.

O modelo mostrou que os lotes pertencentes a parcelamentos estritamente residenciais apresentam valores 4 vezes superiores aos valores dos lotes que pertencem a parcelamentos que não são estritamente residenciais.

Para as barreiras, que têm muita influência na desvalorização de áreas territoriais, os lotes que possuem a sua acessibilidade ao centro não prejudicada pela ferrovia apresentam uma valorização de 58% em relação aos lotes que possuem a sua acessibilidade ao centro prejudicada pela ferrovia. As áreas mais valorizadas têm em comum a característica de otimizarem suas localizações em função do conjunto de deslocamentos diários casa-escola, trabalho, lazer, consumo, minimizando ao máximo os efeitos das barreiras em termos de tempo de deslocamento e segurança.



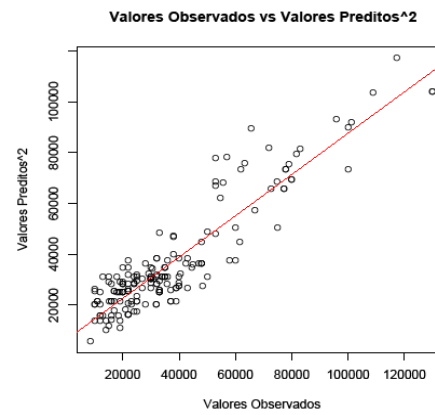


Figura 7. Valores observados versus os valores preditos elevados ao quadrado para o modelo normal usual.

## 6. ANÁLISE DE SOBREVIVÊNCIA

Louzada-Neto *et al.* (2001) diz que “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. Geralmente, a análise de sobrevivência diz respeito a dados biomédicos, enquanto que a análise de confiabilidade refere-se à pesquisa industrial”.

Em estudos médicos ou industriais, em geral, o evento de interesse é o tempo decorrido até o momento da falha. No primeiro, a falha pode ser o óbito do paciente e no segundo, a falha de um determinado produto manufaturado, geralmente um sistema ou parte do mesmo. Entretanto, o termo tem sido estendido além deste limite com o intuito de aplicá-lo aos mais variados tipos de eventos, incluindo os não fatais (Louzada-Neto *et al.*, 2001). Em nosso estudo é tratado como falha a venda em reais, do lote. Em outras palavras, com esta metodologia, por exemplo, podemos estar interessados em saber qual o preço ( $V$ ) mais provável que o lote poderá ser vendido sendo que ele foi ofertado por um valor ( $Z$ ), sendo que  $V \leq Z$ .

A principal característica contida nos bancos de dados para a análise de sobrevivência é a presença de censuras, que é a observação parcial da resposta. Isto se refere às situações em que, por alguma razão, o acompanhamento da unidade em estudo (o lote urbano) foi interrompido. Sem a presença de censura, as técnicas estatísticas clássicas, como análise de regressão e planejamento de experimentos, poderiam ser utilizadas na análise deste tipo de dados, provavelmente usando uma transformação para a resposta. Portanto, o uso dos métodos de análise de sobrevivência possibilita incorporar na análise estatística a informação contida nos dados censurados (Colosimo e Giolo, 2006).

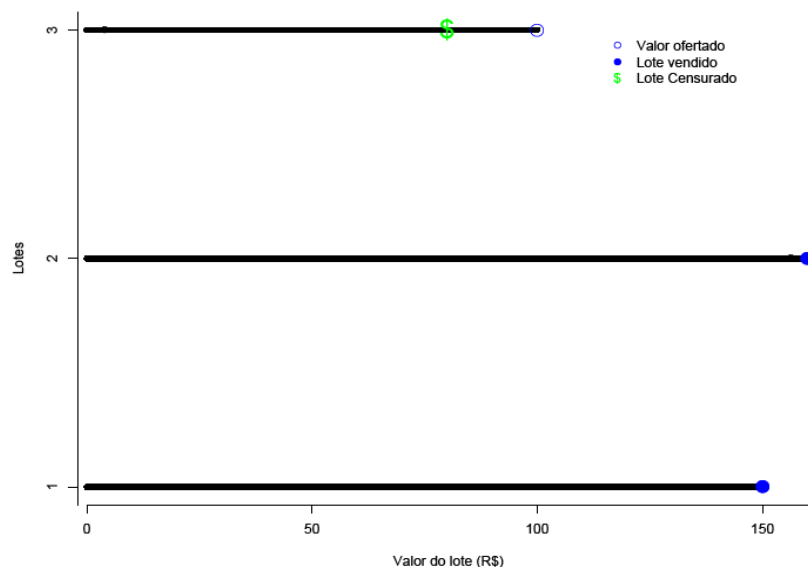


Figura 8. Gráfico mostrando o mecanismo da censura à esquerda no valor do lote urbano.

## 6.1. PARTICULARIDADES DA ANÁLISE DE SOBREVIVÊNCIA

A análise de sobrevivência e confiabilidade se faz peculiar devido às características especiais, inerentes aos tipos de dados que são normalmente disponíveis para análise. Aqui, discutimos apenas duas características principais dos dados de sobrevivência e confiabilidade: a presença de censuras e a presença de covariáveis. As outras duas características, a quantidade de causas de falha e o número de eventos recorrentes, podem ser vistas em Louzada-Neto *et al.* (2001).

### 6.1.1. Presença de Censuras

Vimos que, na análise de sobrevivência, existe a necessidade da introdução de uma variável extra na análise, que indica se a resposta de interesse foi ou não observada. No nosso caso, se a venda de um determinado lote foi, ou não, realizada. Essa variável é conhecida na literatura de análise de sobrevivência e confiabilidade como variável indicadora de censura, ou simplesmente censura (Louzada-Neto *et al.*, 2001).

Colosimo e Giolo (2006) ressaltam o fato que, mesmo censurados, todos os resultados provenientes de um estudo de sobrevivência devem ser usados na análise estatística. Duas razões justificam tal procedimento: (i) mesmo sendo incompletas, as observações censuradas fornecem informações sobre a variável resposta, por exemplo, o valor, em reais, do lote; (ii) a omissão das censuras no cálculo das estatísticas de interesse pode acarretar conclusões viciadas.

As censuras podem ocorrer de várias formas, de acordo com diferentes mecanismos, dentre os quais podemos citar: censuras a direita, censuras a esquerda, censuras de tipo I, censuras de tipo II e censuras aleatórias (Lawless, 1982).

Em análise de sobrevivência é utilizado frequentemente o mecanismo de censura à direita, pois 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. Os mecanismos relativos às censuras de tipo I (quando o estudo termina em um tempo pré-estabelecido e alguns dos tempos de sobrevivência não são observados) e aleatórias (quando um paciente deixa o estudo sem ter experimentado o evento de interesse) são observadas com mais frequência em estudos biomédicos. Entretanto, em experimentos industriais, as censuras de tipo II (quando o estudo termina após a ocorrência de uma determinada quantidade de falhas pré-estabelecida, dentre os itens em estudo) são predominantes (Louzada-Neto *et al.*, 2001).

Neste estudo adotamos o conceito de censura a esquerda. Este tipo de censura 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).

No nosso estudo consideramos a censura à esquerda. Por exemplo, um determinado lote é ofertado por R\$50.000,00. Ele será vendido (falha) por no máximo R\$50.000,00, considerando que será difícil alguém pagar qualquer valor acima do valor ofertado.

### **6.1.2. Presença de Covariáveis**

Em análise de sobrevivência, como nos modelos de regressão lineares usuais, vistos anteriormente, além do tempo de sobrevivência (representado pelo Valor Total, em reais, do Imóvel) e da variável indicadora de censura, também podemos observar nos dados, variáveis que representam à heterogeneidade existente na população. Aqui temos as variáveis de localização (*Núcleo Principal, Planície Central, Ferrovia, Rodovia Washington Luís,*

*Condomínio, Condomínio Fechado e Estritamente Residencial*) como uma variável *dummy* sendo 1 para “pertence” e 0 para “não pertence”. A variável resposta, Valor Total do Imóvel ( $V_i$ ) está em reais (R\$) e a área do terreno em  $m^2$ .

Desta maneira, do ponto de vista estatístico, temos a variável tempo de sobrevivência que nesse estudo será representada pelo Valor Total, em reais, do Imóvel,  $V_i$ , a variável indicadora de censura e um vetor de variáveis explicativas disponíveis para a análise.

## 6.2. DESCRIÇÃO DO COMPORTAMENTO DO VALOR DO IMÓVEL

O comportamento da variável aleatória contínua valor do imóvel,  $V > 0$  pode ser expresso através de várias funções matematicamente equivalentes, tais que, se uma delas é especificada, as outras podem ser derivadas. Entre elas temos a função densidade de probabilidade,  $f(v)$ , a função de sobrevivência,  $S(v)$ , e a função de risco,  $h(v)$ , que serão descritas em detalhes (Louzada-Neto *et al.*, 2001).

### 6.2.1. A função densidade de probabilidade

A função densidade de probabilidade é definida como o limite da probabilidade de um lote ser vendido no intervalo de valor  $[v, v + \Delta v)$  por unidade de valor (R\$, por exemplo), e é expressa por (Lee, 2000 p. 11)

$$f(v) = \lim_{\Delta v \rightarrow 0} \frac{P(v \leq V < v + \Delta v)}{\Delta v}, \quad (6.1)$$

onde  $f(v) \geq 0$  para todo  $v$ , e tem a área abaixo da curva igual a 1.

### 6.2.2. A função de sobrevivência ou função de permanência à venda

Em termos de lote, o termo sobreviver significa permanecer à venda. A função de sobrevivência é definida como sendo a probabilidade de um lote não ser vendido por menos que determinado valor,  $v$ , dada por (Lawless, 1982 p. 8)

$$S(v) = P(V > v) = 1 - F(v), \quad (6.2)$$

tal que  $S(v) = 1$  quando  $v = 0$  e  $S(v) = 0$  quando  $v = \infty$ , e  $F(v) = \int_0^v f(u) du$  representa a função de densidade de probabilidade acumulada ou função de distribuição.  $S(v)$  também é conhecida como a taxa de sobrevivência acumulada (taxa de valor de permanência à venda acumulada). Esta função é geralmente utilizada para determinarmos o  $p$ -ésimo percentil do valor de permanência à venda, por exemplo, o 50-ésimo percentil, que corresponde à mediana do valor de permanência à venda.

### 6.2.3. A função de risco

A função de risco é definida como o limite da probabilidade de um lote ser vendido no intervalo de valor  $[v, v + \Delta v)$ , dado que o mesmo não foi vendido até o valor  $v$ , e é expressa por (Cox e Oakes, 1984 p. 14)

$$h(v) = \lim_{\Delta v \rightarrow 0} \frac{P(v \leq V < v + \Delta v | V \geq v)}{\Delta v}. \quad (6.3)$$

Esta função também pode ser definida em termos de (6.1) e (6.2) por meio da expressão

$$h(v) = \frac{f(v)}{S(v)}, \quad (6.4)$$

descrevendo assim, o relacionamento entre as três funções que geralmente são utilizadas para representar o comportamento do tempo de sobrevivência ou valor de permanência à venda.

## 6.3. A IMPORTÂNCIA DA FUNÇÃO DE RISCO

Devido a sua interpretação, a função de risco (6.4) tem sido preferida por muitos autores para descrever o comportamento do tempo de sobrevivência, no caso, o valor de permanência à venda. A função de risco descreve como a probabilidade instantânea de falha (taxa de falha) se modifica com o passar do tempo (Louzada-Neto *et al.*, 2001). Ela é também

conhecida como taxa de falha instantânea, força de mortalidade e taxa de mortalidade condicional (Cox e Oakes, 1984).

Além disso, através da função de risco podemos caracterizar classes especiais de distribuições de tempo de sobrevivência (valor de permanência à venda), de acordo com o seu comportamento como função do valor de venda. A função de risco pode ser constante, crescente, decrescente ou mesmo não monótona. Algumas distribuições usuais de tempo de sobrevivência são, por exemplo, a exponencial, a Weibull, a log-normal, a log-logística (Louzada-Neto *et al.*, 2001).

A distribuição exponencial acomoda funções de risco constantes, enquanto que se a função de risco for monotonicamente crescente ou decrescente em  $v$ , temos uma distribuição Weibull (Lawless, 1982). As distribuições log-logística e log-normal acomodam funções de risco unimodais (Kalbfleisch e Prentice, 1980). Também, apesar de não muito comuns, funções de riscos multimodais e em forma de “U” também podem ser observadas. Louzada-Neto (1999) propôs uma função de múltiplos riscos que acomoda estas formas de funções de risco (Louzada-Neto *et al.*, 2001).

#### 6.4. PROCEDIMENTOS NÃO-PARAMÉTRICOS

Por mais complexo que seja um estudo, as respostas às perguntas de interesse são dadas a partir de um conjunto de dados de sobrevivência (permanência à venda), e o passo inicial de qualquer análise estatística consiste em uma descrição dos dados. A presença de observações censuradas é, contudo, um problema para as técnicas convencionais de análise descritiva, envolvendo média, desvio-padrão e técnicas gráficas, como histograma, box-plot, entre outros (Colosimo e Giolo, 2006). Louzada-Neto *et al.* (2001) descrevem de maneira simples a estimação da função densidade de probabilidade,  $f(v)$ , da função de sobrevivência (permanência à venda),  $S(v)$ , e a função de risco,  $h(v)$ , definidas em (6.1), (6.2) e (6.3), respectivamente. Estas funções podem ser estimadas diretamente a partir dos dados amostrais.

Os estimadores não-paramétricos mais utilizados na literatura são: o estimador de Kaplan-Meier (o mais utilizado), o estimador de Nelson-Aalen e o estimador da tabela de vida ou atuarial. Neste trabalho será descrito apenas o estimador de Kaplan-Meier. O leitor

interessado nos outros dois estimadores não-paramétricos pode consultar Louzada-Neto *et al.* (2001) e Colosimo e Giolo (2006).

#### 6.4.1. O estimador de Kaplan-Meier

A seguir é descrito como as funções de sobrevivência e de risco podem ser estimadas por meio do estimador de Kaplan-Meier (Kaplan e Meier, 1958), o qual permite a presença de observações censuradas. Este estimador é também conhecido na literatura como estimador produto-limite e foi proposto inicialmente por Böhmer (1912) (Louzada-Neto *et al.* 2001).

Considere um estudo envolvendo  $n$  lotes, e que os valores de permanência à venda, incluindo as censuras, são ordenados, isto é,  $v_1 \leq v_2 \dots \leq v_n$ . A função empírica de permanência à venda é estimada por

$$\hat{S}_{KM}(v) = \frac{n_1 - d_1}{n_1} \frac{n_2 - d_2}{n_2} \dots \frac{n_r - d_r}{n_r} = \prod_{r: v_r \leq v} \frac{n_i - d_i}{n_i}, \quad (6.5)$$

onde  $v_r$  é o maior valor de permanência à venda menor ou igual a  $v$ ,  $n_i$  é o número de lotes não vendidos até o valor  $v_i$  (que representa o valor do lote à venda ordenado  $i$ ) e  $d_i$  representa o número de lotes vendidos no valor  $v_i$  ( $d_i = 0$  para valores de permanência à venda censurados), onde  $i$  pode ser qualquer valor inteiro entre 1 e  $r$ . Na ausência de censuras, o estimador de Kaplan-Meier da função de permanência à venda se reduz a,

$$\hat{S}(v) = \frac{\text{Número de lotes com valores de permanência à venda} > v}{\text{Número total de lotes}}. \quad (6.6)$$

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

$$\hat{H}_{KM}(v) = -\ln \{ \hat{S}_{KM}(v) \}. \quad (6.7)$$

sendo que  $\hat{S}_{KM}(v)$  não pode ser igual à zero.



## 6.5. COMPARAÇÃO DE DUAS FUNÇÕES DE SOBREVIVÊNCIA

Embora possamos comparar grupos de lotes visualmente por meio dos gráficos das funções de permanência à venda estimadas é desejável que tenhamos um teste estatístico para apoiar decisões sobre a igualdade ou não destas curvas.

Um teste muito utilizado na comparação de curvas de sobrevivência de dois grupos de indivíduos é conhecido como teste de *logrank*, nome este, relacionado à utilização dos logaritmos dos postos dos dados em seu cálculo. Este teste foi inicialmente proposto por Mantel e Haenszel (1959) para a comparação de dois conjuntos de proporções, e foi estendido para o teste de tempos de sobrevivência por Mantel (1966) (Louzada-Neto *et al.*, 2001).

Considere dois grupos de lotes, *A* e *B*, a serem comparados. Em uma primeira etapa devemos combinar, ordenar os dados dos dois grupos e calcular os números esperados de vendas no grupo *A*, até um determinado valor  $v_i$ , isto é,

$$E_{A_i} = \frac{r_{A_i}}{r_i} d_i, \quad (6.8)$$

onde  $r_i$  representa o número total de lotes à venda até o valor  $v_i$ ,  $r_{A_i}$  representa o número total de lotes à venda no grupo *A* até o valor  $v_i$  e  $d_i$  representa o número total de vendas até o valor  $v_i$ .

A partir de (6.8) obtemos os números totais esperados de vendas nos dois grupos (*A* e *B*), dados por

$$E_A = \sum_{i=1}^n E_{A_i} \text{ e } E_B = 1 - E_A, \quad (6.9)$$

onde  $n$  é o número total de vendas observado considerando os dois grupos.

A estatística de *logrank* para testar a hipótese de igualdade entre as duas funções de permanência à venda é dada por

$$U^2 = \frac{(O_A - E_A)^2}{E_A} + \frac{(O_B - E_B)^2}{E_B}, \quad (6.10)$$

onde  $O_A$  e  $O_B$  representam os números totais observados de vendas em cada grupo.

O valor obtido em (6.10) deve, então, ser comparado com o quantil da distribuição qui-quadrado com um grau de liberdade.

Uma vez observada a existência da desigualdade entre as curvas de sobrevivência, uma outra quantidade de interesse é dada pelo risco relativo, estimado por

$$RR = \frac{O_A/E_A}{O_B/E_B}. \quad (6.11)$$

A quantidade (6.11) reflete o desempenho de um grupo em relação ao outro.

O teste de *logrank* é muito utilizado em análise de sobrevivência e é particularmente apropriado quando a razão das funções de risco dos grupos a serem comparados é aproximadamente constante. Isto é, as populações têm a propriedade de riscos proporcionais (Colosimo e Giolo, 2006).

## 6.6. PROCEDIMENTOS PARAMÉTRICOS

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. Estas distribuições são descritas em detalhes em Kalbfleisch e Prentice (1980, p. 25-28), entretanto, neste estudo, apenas a distribuição Weibull será descrita (Louzada-Neto *et al.*, 2001).

Além disso, em um contexto prático, a modelagem de dados de sobrevivência está vinculada à forma da função de risco (ver Seção 6.3).

### 6.6.1. Determinação empírica da forma da função de risco

Louzada-Neto *et al.* (2001) dizem que na análise de sobrevivência os modelos citados acima são concorrentes entre si. Para ajustar um determinado conjunto de valores de venda de imóveis (os mesmos podem apresentar diferentes formas de funções de risco) torna-se necessário a utilização de alguma metodologia para selecionar o modelo mais apropriado, mesmo antes de qualquer ajuste.

Em muitas aplicações existe informação qualitativa e, muitas vezes, estrutural a respeito do fenômeno em questão, que pode ser utilizada na determinação empírica da forma da função de risco. Informações estruturais estão diretamente vinculadas ao conhecimento do pesquisador sobre o fenômeno, enquanto que informações qualitativas podem ser extraídas por meio de uma análise gráfica. Neste contexto, um gráfico conhecido como gráfico do

tempo total em teste (curva TTT) é de grande utilidade. Este gráfico foi inicialmente proposto por Aarset (1987) e é construído a partir das quantidades

$$G(r/n) = \left[ \left( \sum_{i=1}^r T_{i:n} \right) + (n-r)T_{r:n} \right] / \left( \sum_{i=1}^r T_{i:n} \right) \text{ versus } A = r/n, \quad (6.12)$$

onde  $r = 1, \dots, n$  e  $T_{i:n}, i = 1, \dots, n$  são as estatísticas de ordem da amostra (Mudholkar *et al.*, 1996).

Caso tenhamos informações sobre covariáveis para cada indivíduo e uma quantidade significativa de indivíduos em cada nível ou combinação destas covariáveis, a curva TTT (6.12) pode ser construída considerando cada nível de covariável ou combinação das mesmas, separadamente (Louzada-Neto *et al.*, 2001).

## 6.7. DISTRIBUIÇÃO WEIBULL

A distribuição Weibull foi proposta originalmente por Wallodi Weibull em (1951) e desde então, devido em grande parte à sua simplicidade, tem sido uma das distribuições de probabilidade mais utilizadas na modelagem de dados biomédicos bem como industriais e aqui é avaliada a sua utilização para modelar o valor total, em reais, do lote (Louzada-Neto *et al.*, 2001). Sua densidade pode ser escrita na forma

$$f(v) = \frac{\gamma}{\alpha} \left( \frac{v}{\alpha} \right)^{\gamma-1} \exp \left\{ - \left( \frac{v}{\alpha} \right)^{\gamma} \right\}, \quad (6.13)$$

onde  $\gamma > 0$  e  $\alpha > 0$  são os parâmetros de forma e escala, respectivamente.

Quando  $\gamma = 1$  em (6.13), obtemos a distribuição exponencial como caso particular. É comum encontrarmos na literatura a distribuição Weibull escrita sob diferentes parametrizações (Louzada-Neto *et al.*, 2001).

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

$$h(v) = \frac{\gamma}{\alpha} \left( \frac{v}{\alpha} \right)^{\gamma-1}, \quad (6.14)$$

$$S(v) = \exp \left\{ - \left( \frac{v}{\alpha} \right)^{\gamma} \right\}, \quad (6.15)$$

$$v_p = \alpha \left[ -\log(1-p) \right]^{\frac{1}{\gamma}}, \quad (6.16)$$

Como visto anteriormente, 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  $\gamma < 1$  temos funções de risco monótonas decrescentes, para  $\gamma > 1$  as funções de risco são monótonas crescentes e para  $\gamma = 1$  temos a distribuição exponencial com função de risco constante (Louzada-Neto *et al.*, 2001).

### 6.7.1. Regressão Weibull

Consideramos a situação em que temos disponível uma amostra aleatória  $v_1, \dots, v_n$  de tempos de sobrevivência e os valores da variável indicadora  $\delta_i$ ,  $\delta_i = 1$  se  $v_i$  é exatamente observado ou  $\delta_i = 0$  se  $v_i$  é censurado à esquerda. Baseado nas informações  $(v_1, \delta_1), \dots, (v_n, \delta_n)$ , que o esquema de censura é não informativo e que os  $v_i$  são provenientes da mesma distribuição de probabilidade indexada pelo parâmetro  $\theta$ , a função de verossimilhança é genericamente escrita na forma

$$L(\theta) = \prod_{\delta_i=1} f(v_i|\theta) \prod_{\delta_i=0} F(v_i|\theta) \quad (6.17)$$

Para a distribuição Weibull (6.13) com parâmetros  $\alpha$  e  $\gamma$ , considerando uma amostra aleatória  $v_1, \dots, v_n$  e a variável indicadora de censura  $\delta_i$ ,  $\delta_i = 1$  se  $v_i$  é exatamente observado ou  $\delta_i = 0$  se  $v_i$  é censurado à esquerda, a função de verossimilhança, na presença de covariáveis, é escrita na forma

$$L(\alpha; \gamma | v_i; \delta_i; \mathbf{x}_i) = \prod_{i=1}^n [f(v_i | \mathbf{x}_i)]^{\delta_i} [1 - S(v_i | \mathbf{x}_i)]^{1-\delta_i} = \prod_{i=1}^n [f(v_i | \mathbf{x}_i)]^{\delta_i} [F(v_i | \mathbf{x}_i)]^{1-\delta_i} \quad (6.18)$$

onde  $\mathbf{x}_i$  é o vetor de covariáveis referente ao  $i$ -ésimo lote e a variável indicadora de censura é definida como

$$\delta_i = \begin{cases} 1, & \text{se o lote foi vendido e} \\ 0, & \text{caso contrário.} \end{cases} \quad (6.19)$$

Aplicando o logaritmo natural na função de verossimilhança e após alguns passos obtém-se a seguinte função,

$$l(\alpha; \gamma|v_i; \delta_i; \mathbf{x}_i) = \sum_{i=1}^n \left\{ \delta_i \left[ (\gamma-1)\ln(v_i) - \gamma \ln(\alpha) + \ln(\gamma) - \left(\frac{v_i}{\alpha}\right)^\gamma \right] + (1-\delta_i) \ln \left( 1 - \exp \left\{ - \left(\frac{v_i}{\alpha}\right)^\gamma \right\} \right) \right\}, \quad (6.20)$$

sendo que,  $l(\alpha; \gamma|v_i; \delta_i; \mathbf{x}_i) = \log[L(\alpha; \gamma|v_i; \delta_i; \mathbf{x}_i)]$

### 6.7.2. Estratégia para a seleção de covariáveis

Uma das etapas de um processo de modelagem de dados consiste em identificar um particular conjunto de covariáveis que consiga reter variabilidade relevante da variável resposta. Conforme aumenta o número de covariáveis potencialmente importantes para descrever o comportamento da resposta aumenta o número de possíveis modelos formados pela combinação de todas estas covariáveis. No nosso caso temos nove covariáveis, portanto, existem  $2^9 = 512$  possíveis modelos formados pela combinação de todas estas covariáveis. É certamente impraticável ajustar todos estes possíveis modelos a fim de ser selecionado o que melhor explique a resposta. Nessas situações, as rotinas automáticas para seleção de covariáveis descritas em (4.8) podem ser utilizadas, entretanto, possuem limitações às quais já descritas anteriormente em (4.8).

Frente a estas limitações das rotinas automáticas, Colosimo e Giolo (2006) sugerem utilizar métodos que envolvem a interferência do analista. A filosofia do método é essencialmente a mesma para qualquer classe de modelos. Aqui se optou por utilizar uma estratégia de seleção de modelos derivada da proposta de Collet (1994). Os passos utilizados no processo de seleção são, segundo Colosimo e Giolo (2006), descritos a seguir:

1. Ajustar todos os modelos contendo uma única covariável. Incluir todas as covariáveis que forem significativas ao nível de 0,10. É aconselhável utilizar o teste da razão de verossimilhança neste passo.
2. As covariáveis significativas no passo 1 são, então, ajustadas conjuntamente. Na presença de certas covariáveis, outras podem deixar de ser significativas. Conseqüentemente, ajustam-se os modelos reduzidos, excluindo uma única covariável de cada vez. Verificam-se quais covariáveis provocam um aumento estatisticamente significativo na estatística da razão de verossimilhanças. Somente aquelas que atingirem a significância permanecem no modelo.

3. Ajusta-se um novo modelo com as covariáveis retidas no passo 2. Neste passo, as covariáveis excluídas no passo 2 retornam ao modelo para confirmar que elas não são estatisticamente significativas.
4. As eventuais covariáveis significativas no passo 3 são incluídas ao modelo juntamente com aquelas do passo 2. Neste passo, retorna-se com as covariáveis excluídas no passo 1 para confirmar que elas não são estatisticamente significativas.
5. Ajusta-se um modelo incluindo-se as covariáveis significativas no passo 4. Neste passo é testado se alguma delas pode ser retirada do modelo.
6. Utilizando as covariáveis que sobreviveram ao passo 5, ajusta-se o modelo final para os efeitos principais. Para completar a modelagem, deve-se verificar a possibilidade de inclusão de termos de interação dupla entre as covariáveis incluídas no modelo. O modelo final fica determinado pelos efeitos principais identificados no passo 5 e os termos de interação significativos identificados neste passo.

Ao ser utilizado este procedimento de seleção, deve-se incluir as informações inerentes à área de estudo, tais como, avaliação de imóveis, no processo de decisão e evitar muito rigor ao testar cada nível individual de significância. Na decisão da inclusão de um termo no modelo o nível de significância não deve ser muito baixo, sendo recomendado um valor próximo de 0,10. Variações deste método de seleção de covariáveis podem ser encontradas na literatura. Hosmer e Lemeshow (1999) discutem claramente estes métodos com bastante elegância (Colosimo e Giolo, 2006).

## 6.8. TESTES DE HIPÓTESES

Para um modelo com um vetor  $\theta = (\theta_1, \dots, \theta_p)$  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. A seguir é apresentada uma breve descrição do Teste da Razão de Verossimilhanças (Colosimo e Giolo, 2006).

### 6.8.1. Teste da razão de verossimilhanças

Este teste é baseado na função de verossimilhança e envolve a comparação dos valores do logaritmo da função de verossimilhança maximizada sem restrição e sob  $H_0$ , ou seja, a comparação de  $\log L(\hat{\theta})$  e  $\log L(\hat{\theta}_0)$ . A estatística para este teste é dada por

$$TRV = -2 \log \left[ \frac{L(\hat{\theta}_0)}{L(\hat{\theta})} \right] = 2 \left[ \log L(\hat{\theta}) - \log L(\hat{\theta}_0) \right], \quad (6.21)$$

que sob  $H_0 : \theta = \theta_0$ , segue aproximadamente uma distribuição qui-quadrado com  $p$  graus de liberdade, onde  $p$  é o número de parâmetros. Para amostras grandes,  $H_0$  é rejeitada, a um nível  $100\alpha\%$  de significância, se  $TRV > \chi_{p,1-\alpha}^2$  (Colosimo e Giolo, 2006).

## 7. RESULTADOS DA ANÁLISE DE SOBREVIVÊNCIA

### 7.1. ANÁLISE DESCRITIVA E EXPLORATÓRIA

A primeira etapa de qualquer análise estatística de dados consiste de análises descritivas das variáveis em estudo. Em análise de sobrevivência, esta etapa consiste em utilizar os métodos não-paramétricos apresentados em (6.4). Na Figura 9(a) apresenta-se a função empírica de permanência à venda, estimada através do método de Kaplan-Meier e na Figura 9(b) apresenta-se a função de risco acumulado empírico. Verifica-se que à medida que o valor do lote aumenta o risco de vendê-lo também aumenta. Em outras palavras, conforme aumenta o valor do lote diminui a chance de não vendê-lo.

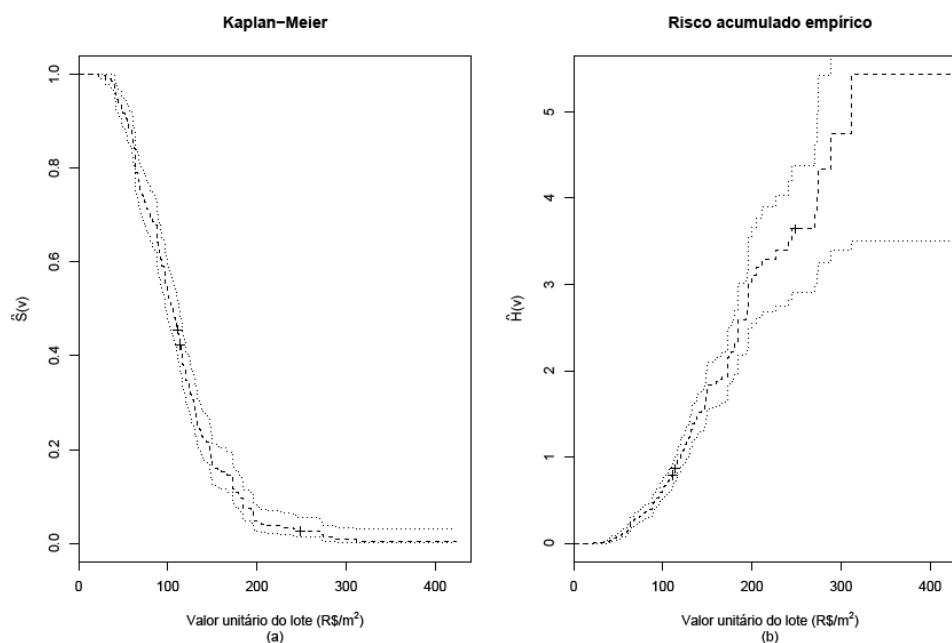


Figura 9. (a) Curva de permanência à venda estimada pelo método de Kaplan-Meier; (b) Risco acumulado de venda empírico e os respectivos intervalos 95% de confiança.

### 7.2. ESTUDO DE SIMULAÇÃO PARA O MODELO DE REGRESSÃO WEIBULL

Na simulação utilizando o modelo de regressão Weibull repetiu-se o mesmo procedimento da simulação do modelo de regressão linear usual (ver Seção 5.2), alternando apenas o modelo. Na primeira simulação utiliza-se o modelo de regressão linear usual e na



segunda o modelo de regressão Weibull. Para o modelo de regressão Weibull serão consideradas 6 porcentagens de censuras diferentes: 0%, 1%, 5%, 15%, 30% e 60% para os tamanhos de amostra 25, 50, 150, 500 e 1000.

Tabela 6. Probabilidades de cobertura nominais das estimativas dos parâmetros para os 5 tamanhos diferentes de amostra.

<b>Censuras</b>	<b>Parâmetro</b>	<b>n=25</b>	<b>n=50</b>	<b>n=150</b>	<b>n=500</b>	<b>n=1000</b>
<b>0%</b>	$\beta_0$	0.94	0.94	0.95	0.95	0.95
<b>1%</b>	$\beta_0$	0.94	0.95	0.95	0.95	0.95
<b>5%</b>	$\beta_0$	0.95	0.95	0.95	0.95	0.94
<b>15%</b>	$\beta_0$	0.93	0.93	0.93	0.92	0.90
<b>30%</b>	$\beta_0$	0.89	0.91	0.90	0.86	0.81
<b>60%</b>	$\beta_0$	0.82	0.80	0.76	0.61	0.41
<b>0%</b>	$\beta_1$	0.94	0.96	0.95	0.95	0.95
<b>1%</b>	$\beta_1$	0.94	0.95	0.95	0.95	0.95
<b>5%</b>	$\beta_1$	0.93	0.95	0.94	0.95	0.95
<b>15%</b>	$\beta_1$	0.92	0.93	0.93	0.94	0.92
<b>30%</b>	$\beta_1$	0.88	0.90	0.91	0.92	0.91
<b>60%</b>	$\beta_1$	0.85	0.84	0.85	0.84	0.83
<b>0%</b>	$\beta_2$	0.95	0.95	0.95	0.94	0.94
<b>1%</b>	$\beta_2$	0.93	0.95	0.94	0.96	0.94
<b>5%</b>	$\beta_2$	0.93	0.94	0.94	0.95	0.93
<b>15%</b>	$\beta_2$	0.91	0.93	0.92	0.94	0.92
<b>30%</b>	$\beta_2$	0.88	0.90	0.90	0.92	0.90
<b>60%</b>	$\beta_2$	0.80	0.85	0.84	0.83	0.83
<b>0%</b>	$\beta_3$	0.94	0.94	0.95	0.95	0.95
<b>1%</b>	$\beta_3$	0.94	0.95	0.94	0.96	0.94
<b>5%</b>	$\beta_3$	0.94	0.95	0.94	0.96	0.93
<b>15%</b>	$\beta_3$	0.92	0.94	0.92	0.95	0.93
<b>30%</b>	$\beta_3$	0.90	0.92	0.91	0.92	0.92
<b>60%</b>	$\beta_3$	0.83	0.83	0.82	0.83	0.84

<b>0%</b>	$\beta_4$	0.94	0.95	0.96	0.95	0.94
<b>1%</b>	$\beta_4$	0.95	0.94	0.95	0.95	0.94
<b>5%</b>	$\beta_4$	0.95	0.94	0.95	0.95	0.94
<b>15%</b>	$\beta_4$	0.94	0.92	0.93	0.94	0.93
<b>30%</b>	$\beta_4$	0.92	0.90	0.92	0.91	0.90
<b>60%</b>	$\beta_4$	0.82	0.84	0.83	0.85	0.87
<b>0%</b>	$\beta_5$	0.93	0.95	0.95	0.95	0.95
<b>1%</b>	$\beta_5$	0.94	0.96	0.96	0.96	0.95
<b>5%</b>	$\beta_5$	0.94	0.96	0.95	0.96	0.94
<b>15%</b>	$\beta_5$	0.92	0.94	0.94	0.94	0.94
<b>30%</b>	$\beta_5$	0.89	0.92	0.91	0.93	0.91
<b>60%</b>	$\beta_5$	0.87	0.84	0.85	0.84	0.84

Tabela 7. Discrepâncias das estimativas dos parâmetros para os 5 tamanhos diferentes de amostra.

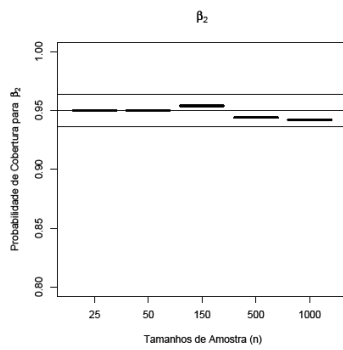
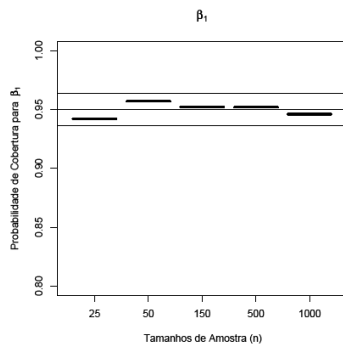
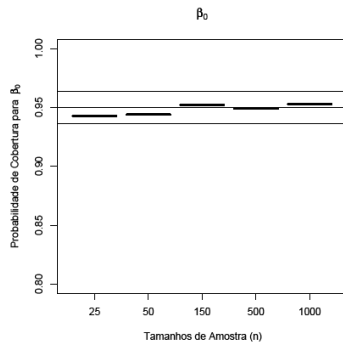
<b>Censuras</b>	<b>Parâmetro</b>	<b>n=25</b>	<b>n=50</b>	<b>n=150</b>	<b>n=500</b>	<b>n=1000</b>
<b>0%</b>	$\beta_0$	0.365	0.126	0.058	0.006	0.010
<b>1%</b>	$\beta_0$	0.536	0.299	0.087	0.033	0.023
<b>5%</b>	$\beta_0$	0.668	0.235	0.187	0.074	0.087
<b>15%</b>	$\beta_0$	0.778	0.326	0.258	0.235	0.259
<b>30%</b>	$\beta_0$	1.025	0.66	0.528	0.455	0.468
<b>60%</b>	$\beta_0$	2.688	1.547	1.358	1.144	1.139
<b>0%</b>	$\beta_1$	0.063	0.033	0.015	0.002	0.005
<b>1%</b>	$\beta_1$	0.028	0.003	0.006	0.012	0.002
<b>5%</b>	$\beta_1$	0.024	0.006	0.007	0.000	0.006
<b>15%</b>	$\beta_1$	0.046	0.011	0.016	0.007	0.006
<b>30%</b>	$\beta_1$	0.051	0.073	0.016	0.000	0.001
<b>60%</b>	$\beta_1$	0.072	0.065	0.017	0.003	0.006
<b>0%</b>	$\beta_2$	0.079	0.028	0.017	0.005	0.008
<b>1%</b>	$\beta_2$	0.108	0.03	0.019	0.001	0.001
<b>5%</b>	$\beta_2$	0.116	0.046	0.017	0.004	0.007

---

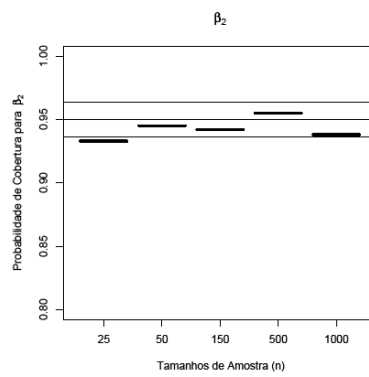
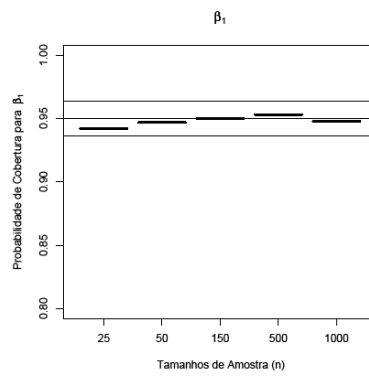
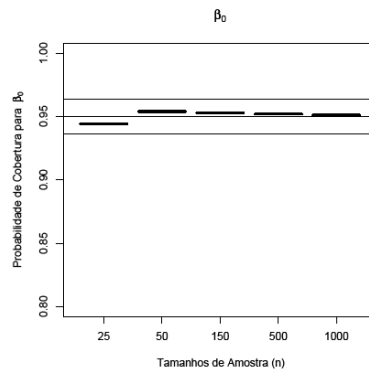
<b>15%</b>	$\beta_2$	0.095	0.017	0.018	0.003	0.002
<b>30%</b>	$\beta_2$	0.148	0.055	0.021	0.015	0.001
<b>60%</b>	$\beta_2$	0.342	0.117	0.016	0.008	0.006
<b>0%</b>	$\beta_3$	0.024	0.002	0.012	0.005	0.001
<b>1%</b>	$\beta_3$	0.017	0.024	0.002	0.003	0.003
<b>5%</b>	$\beta_3$	0.036	0.015	0.004	0.008	0.001
<b>15%</b>	$\beta_3$	0.016	0.029	0.019	0.006	0.001
<b>30%</b>	$\beta_3$	0.068	0.019	0.003	0.014	0.000
<b>60%</b>	$\beta_3$	0.020	0.008	0.009	0.014	0.005
<b>0%</b>	$\beta_4$	0.159	0.056	0.003	0.019	0.006
<b>1%</b>	$\beta_4$	0.011	0.024	0.012	0.002	0.001
<b>5%</b>	$\beta_4$	0.010	0.016	0.012	0.001	0.004
<b>15%</b>	$\beta_4$	0.142	0.020	0.006	0.006	0.006
<b>30%</b>	$\beta_4$	0.141	0.007	0.000	0.000	0.009
<b>60%</b>	$\beta_4$	0.507	0.112	0.028	0.012	0.006
<b>0%</b>	$\beta_5$	0.000	0.000	0.000	0.000	0.000
<b>1%</b>	$\beta_5$	0.000	0.000	0.000	0.000	0.000
<b>5%</b>	$\beta_5$	0.000	0.000	0.000	0.000	0.000
<b>15%</b>	$\beta_5$	0.000	0.000	0.000	0.000	0.000
<b>30%</b>	$\beta_5$	0.000	0.000	0.000	0.000	0.000
<b>60%</b>	$\beta_5$	0.001	0.000	0.000	0.000	0.000

---

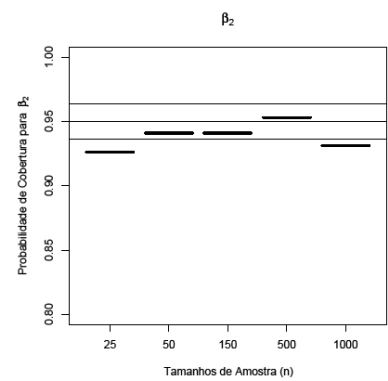
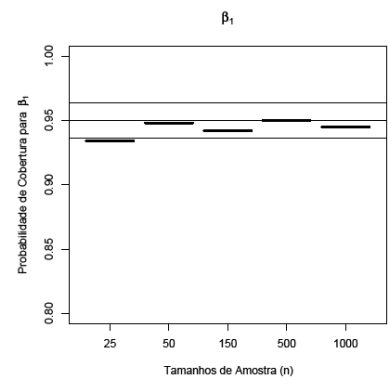
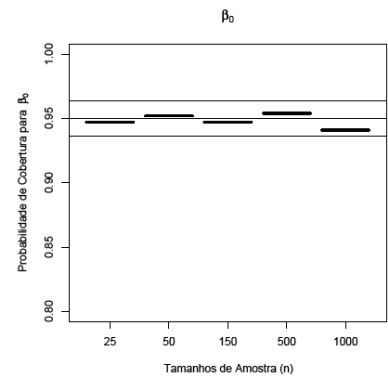
## 0% censuras



## 1% censuras



## 5% censuras



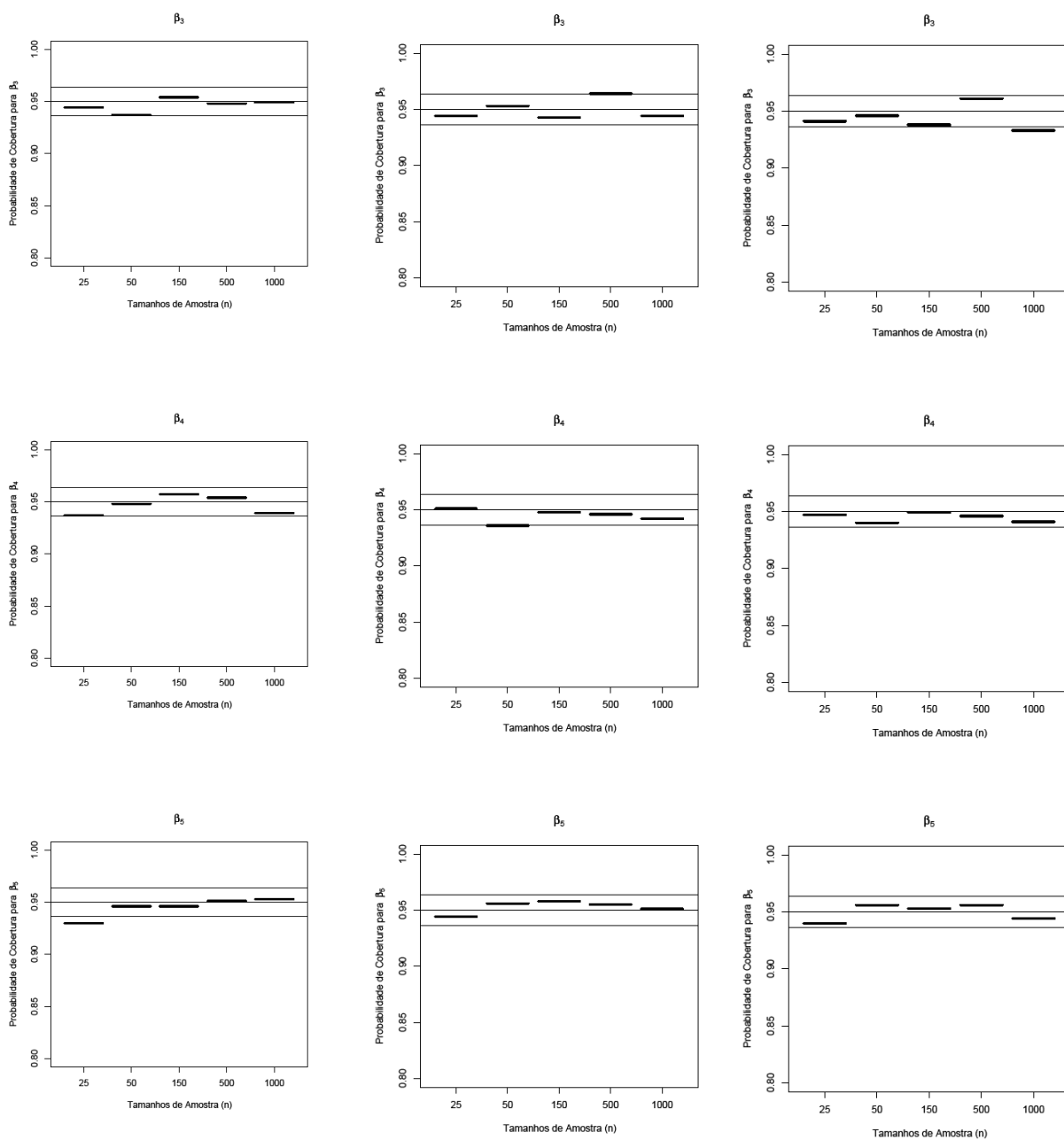
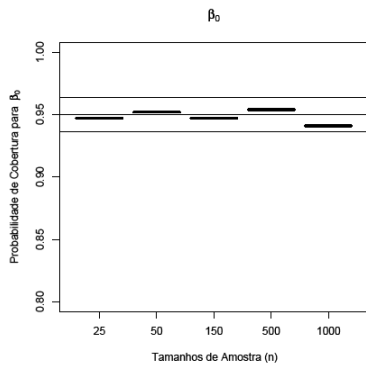
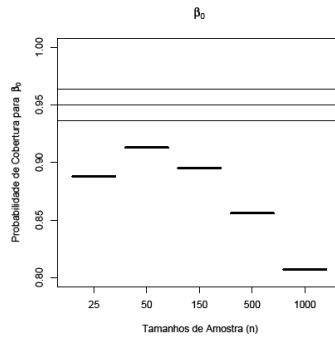


Figura 10. Limite inferior nominal, limite superior nominal e valor central nominal (95%) dos intervalos assintóticos normais das probabilidades de cobertura nominais das estimativas dos parâmetros para os 5 tamanhos diferentes de amostra e 3 diferentes porcentagens de censuras (0%, 1% e 5%).

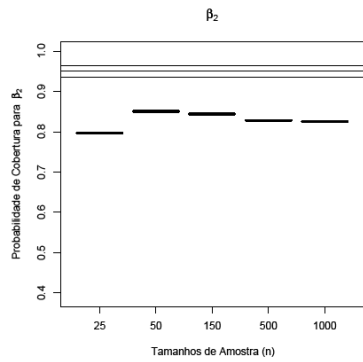
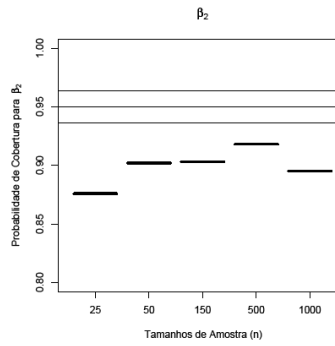
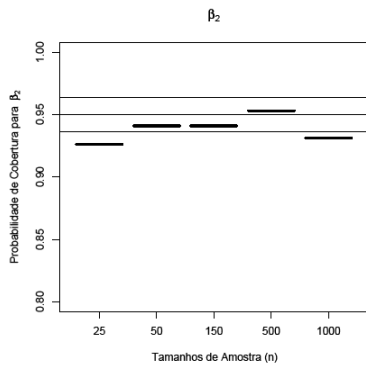
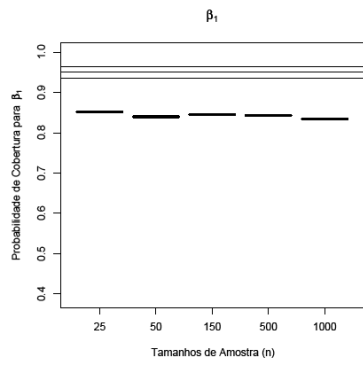
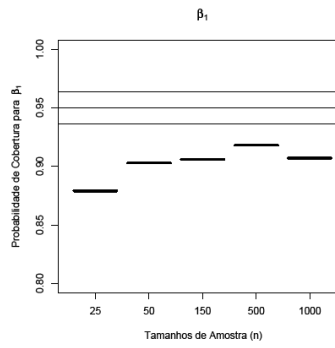
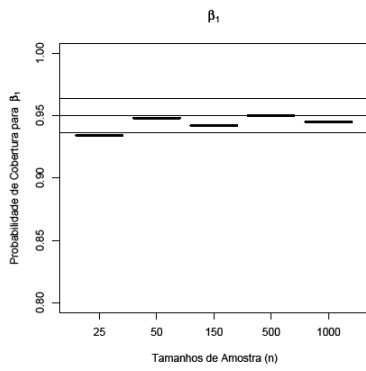
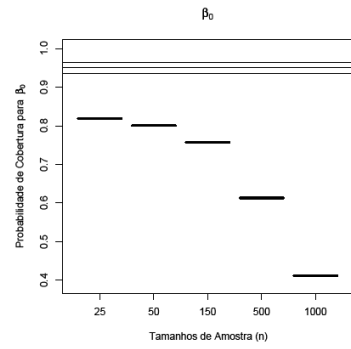
**15% censuras**



**30% censuras**



**60% censuras**



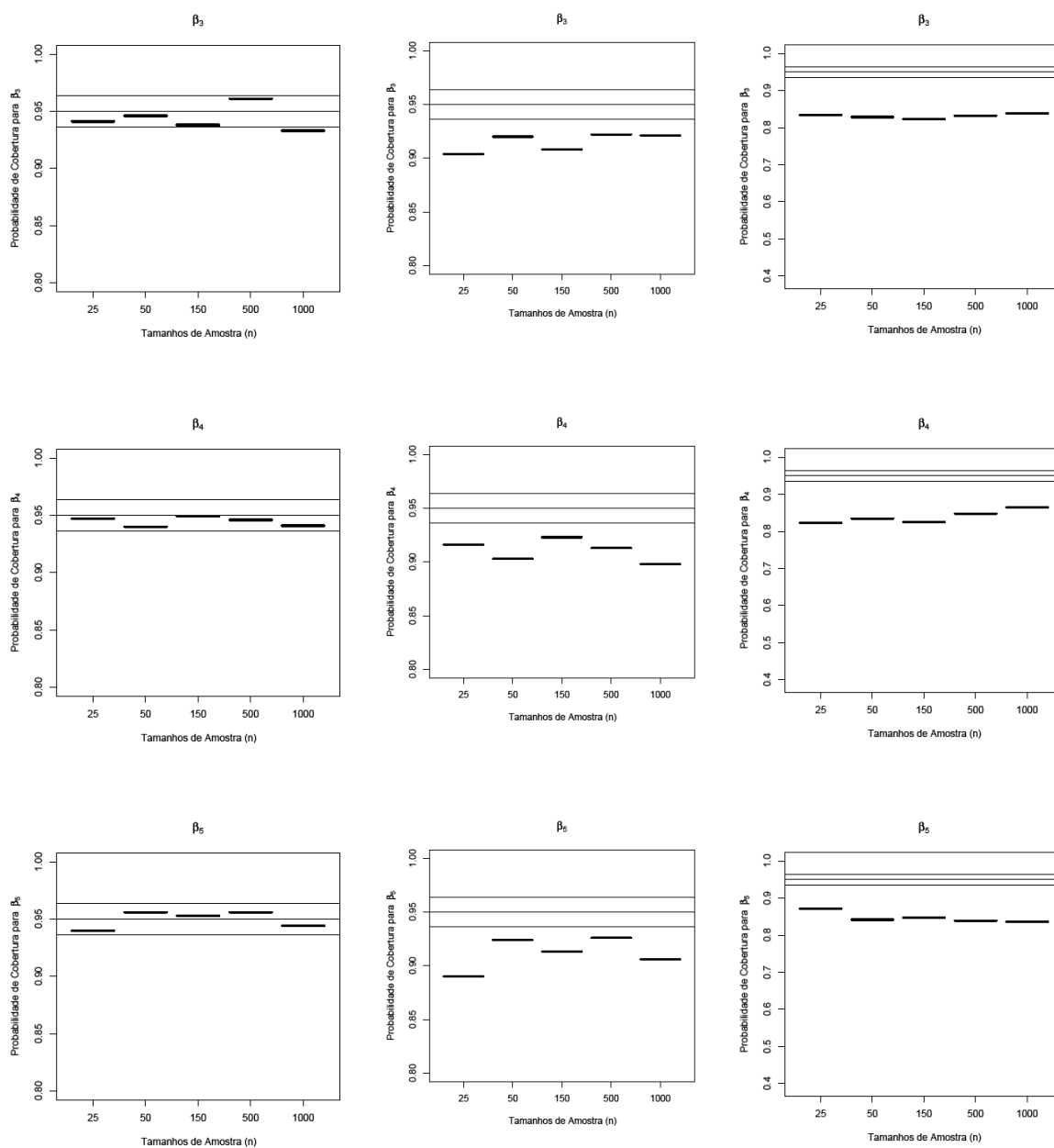
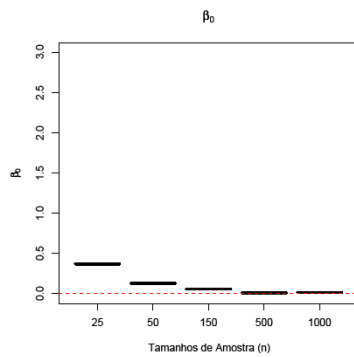
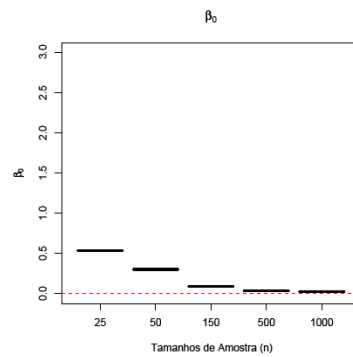


Figura 11. Limite inferior nominal, limite superior nominal e valor central nominal (95%) dos intervalos assintóticos normais das probabilidades de cobertura nominais das estimativas dos parâmetros para os 5 tamanhos diferentes de amostra e 3 diferentes percentagens de censuras (15%, 30% e 60%).

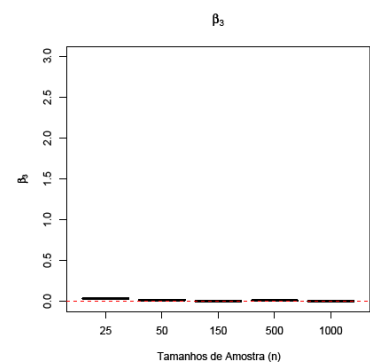
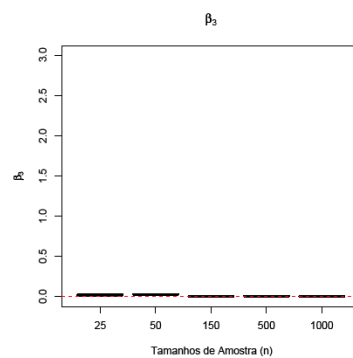
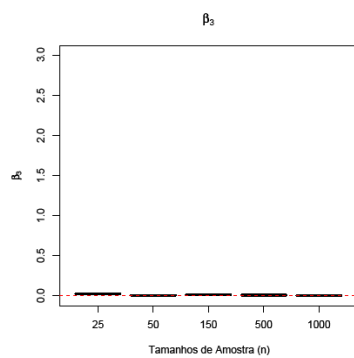
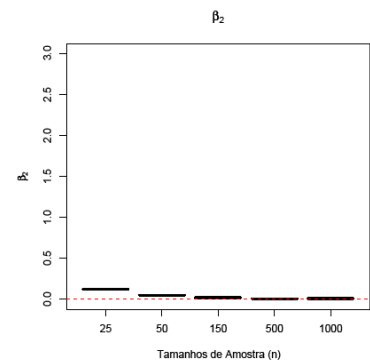
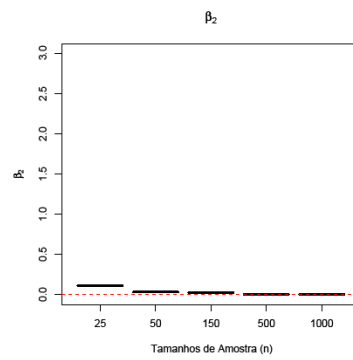
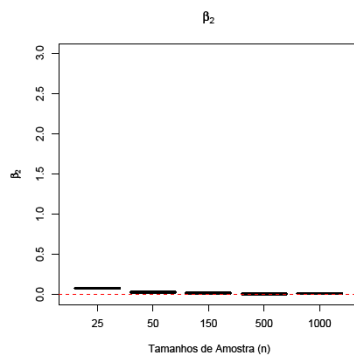
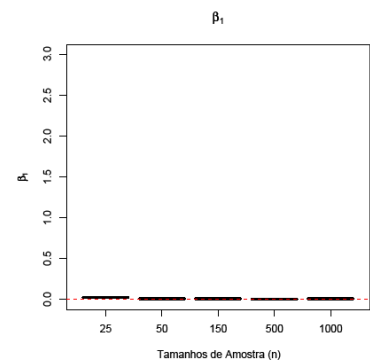
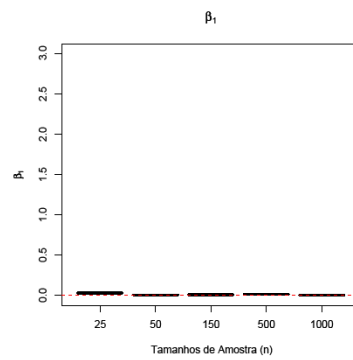
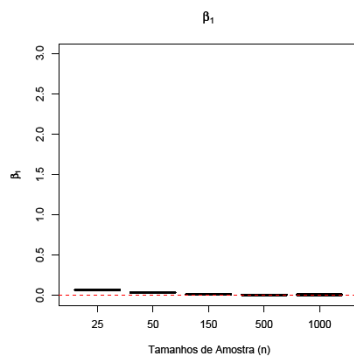
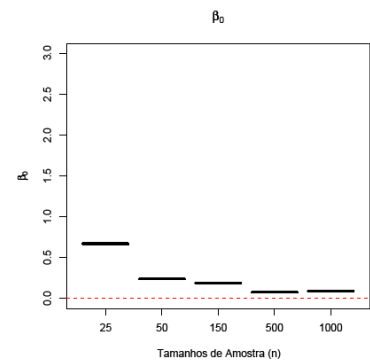
## 0% censuras



## 1% censuras



## 5% censuras





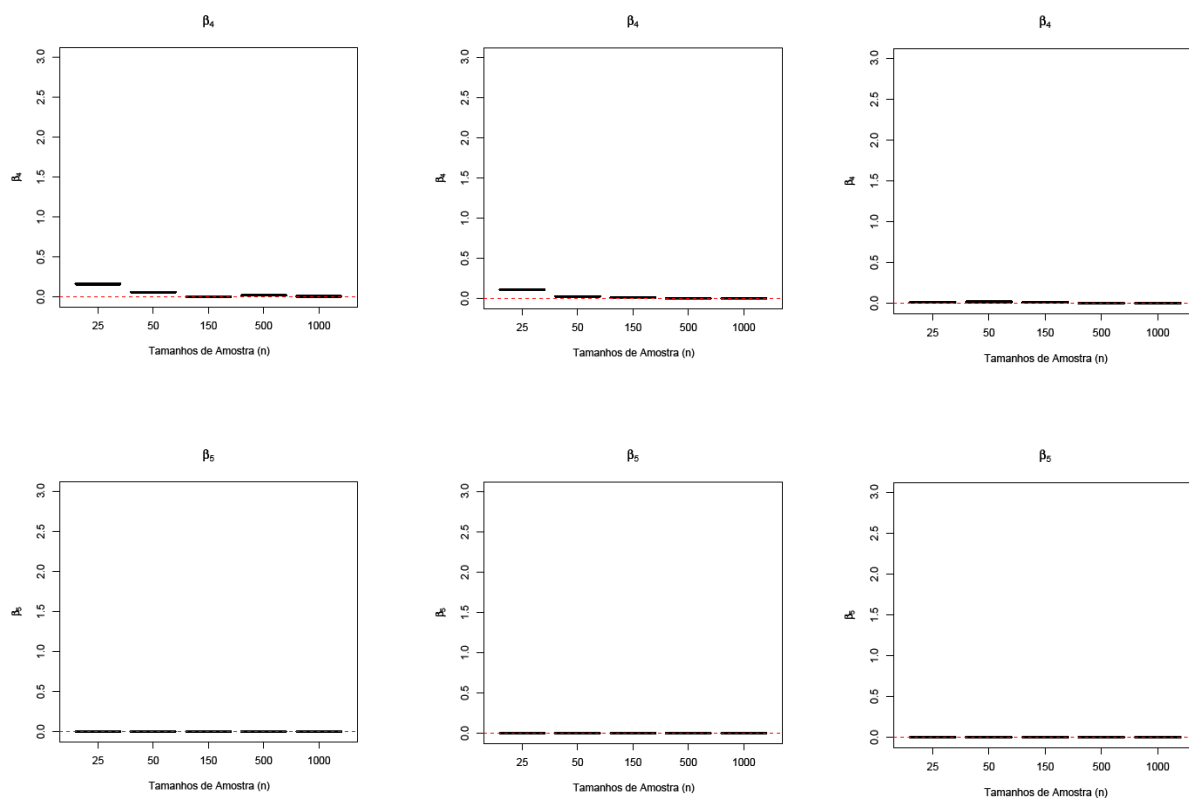
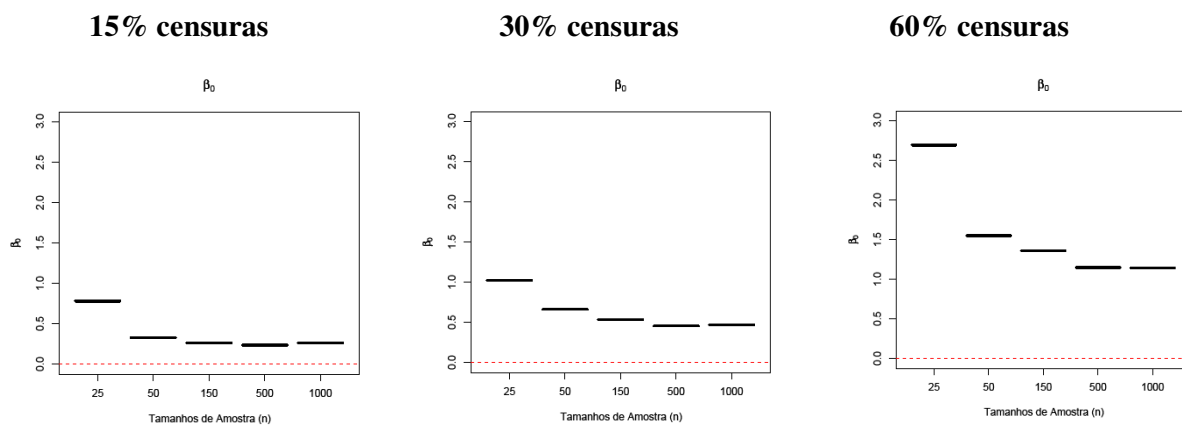
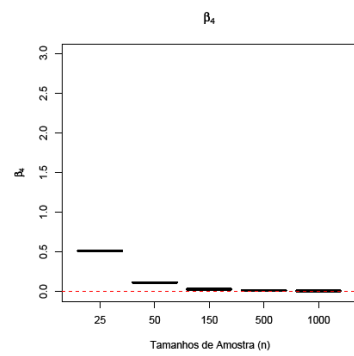
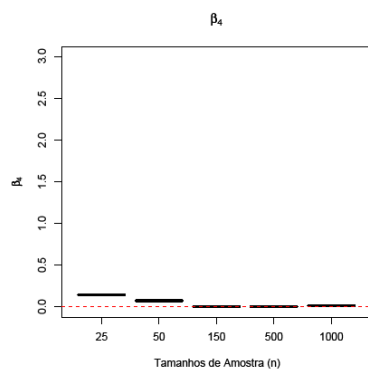
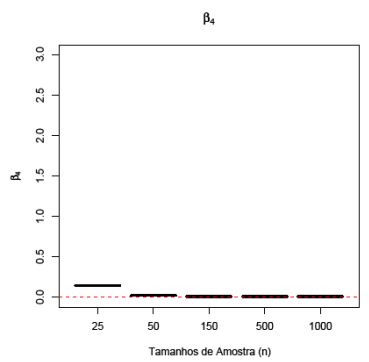
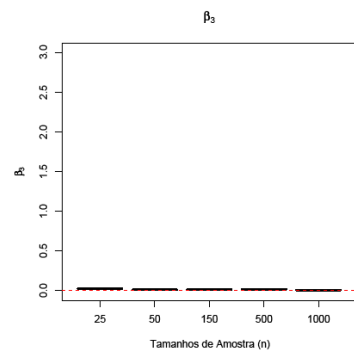
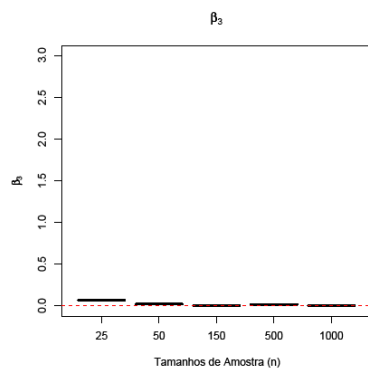
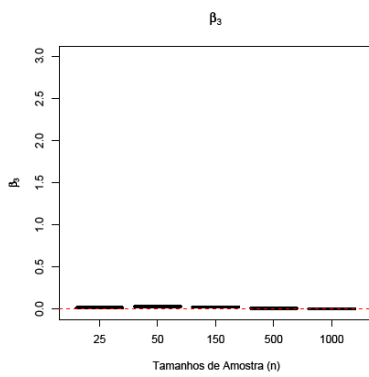
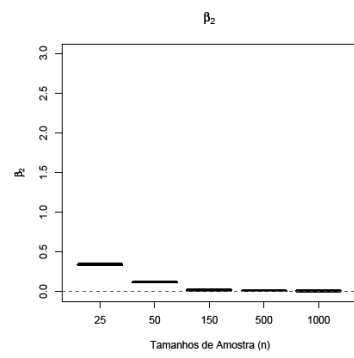
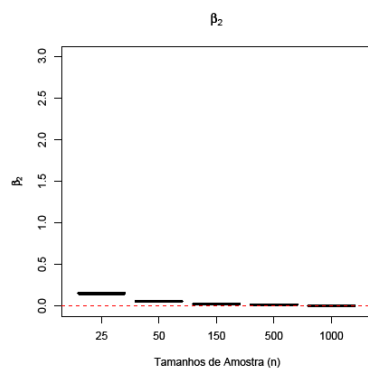
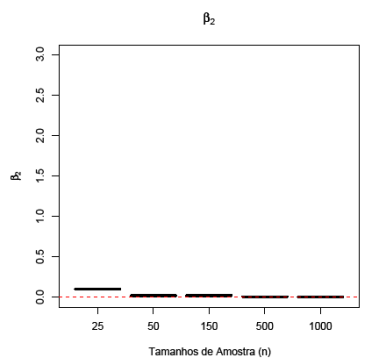
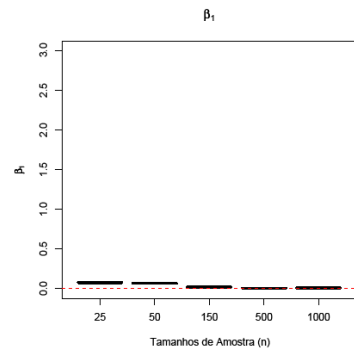
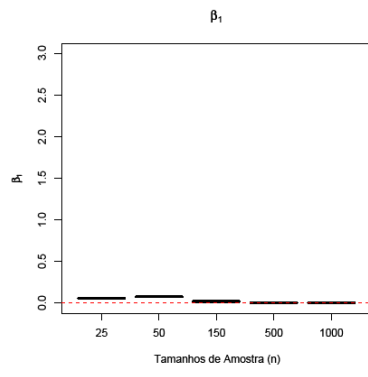
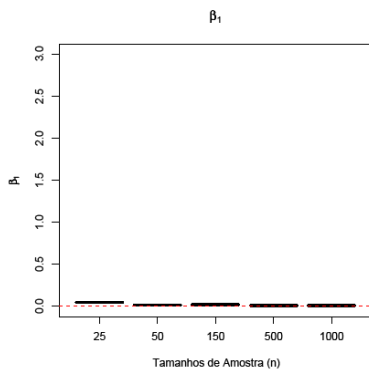


Figura 12. Discrepâncias das estimativas dos parâmetros para os 5 tamanhos diferentes de amostra e 3 diferentes porcentagens de censuras (0%, 1% e 5%).





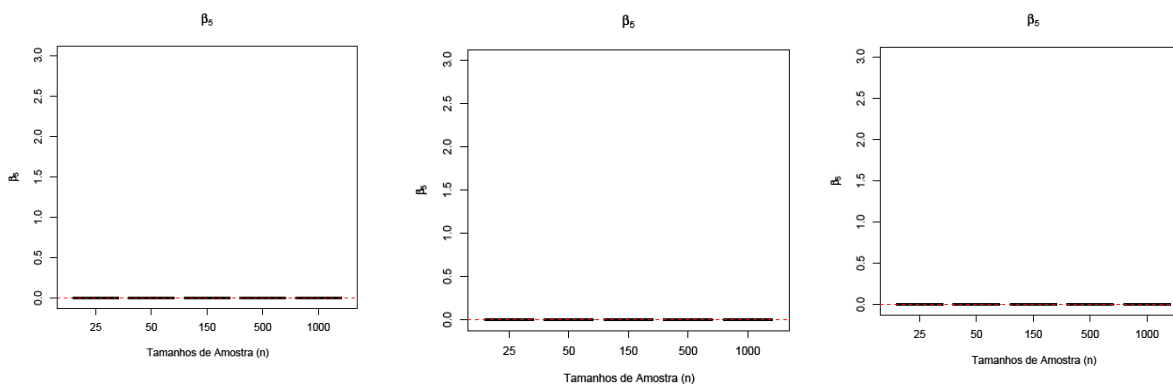


Figura 13. Discrepâncias das estimativas dos parâmetros para os 5 tamanhos diferentes de amostra e 3 diferentes porcentagens de censuras (15%, 30% e 60%).

Através da Tabela 6 e das Figuras 10 e 11 onde se considera a porcentagem de censuras entre 0% e 15% e com amostras de tamanho superior a 50 unidades de observação, os parâmetros apresentaram probabilidade de cobertura dentro dos limites nominais. O aumento de censura, a partir de 30%, direciona a uma diminuição da probabilidade de cobertura dos parâmetros, para qualquer tamanho de amostra, mas esta diminuição não é maior do que 10% abaixo do limite nominal, exceto para o parâmetro  $\beta_0$  (intercepto).

A probabilidade de cobertura do intercepto ( $\beta_0$ ) mostrou que, a partir de 30% de censuras nas amostras, a sua probabilidade de cobertura diminuiu à medida que o tamanho da amostra aumentou.

A partir da Tabela 7 e das Figuras 12 e 13 verifica-se que os estimadores, exceto  $\beta_5$ , apresentaram discrepâncias em dois casos, quando o tamanho de amostra é pequeno e quando aumenta o número de censuras na amostra. O intercepto ( $\beta_0$ ) foi o parâmetro que apresentou a maior discrepância das estimativas e o  $\beta_5$ , parâmetro responsável pelo efeito da única variável numérica, apresentou a menor discrepância, independente do número de censuras para ambos os parâmetros  $\beta_0$  e  $\beta_5$ .

### 7.3. AJUSTE DE UM MODELO DE REGRESSÃO PARAMÉTRICO

Nesta seção serão utilizados métodos paramétricos para modelar o valor de venda de lotes em função das covariáveis consideradas. A utilização desses métodos requer a

especificação de uma distribuição de probabilidade para a variável resposta. Nessa situação, o passo mais importante da modelagem é encontrar uma distribuição de probabilidade adequada para os dados em estudo. Somente após encontrar esta distribuição será possível estimar e testar as quantidades de interesse (Colosimo e Giolo, 2006).

Para determinar qual distribuição de probabilidade melhor se ajusta aos dados utilizou-se o TTT-Plot para os valores unitários dos lotes de São Carlos, SP, no ano de 2005.

Ainda na exploração dos dados, como as variáveis de localização são todas dicotômicas é possível construir as estimativas de Kaplan-Meier para comparar as duas categorias. Isto foi feito para as oito covaráveis de localização, incluindo um teste de hipótese para avaliar a igualdade ou não das duas curvas utilizando-se o teste de *logrank*. Não apresentaremos os oito gráficos, mas, a título de ilustração, a Figura 14(a) apresenta a curva de Kaplan-Meier para a covariável Ferrovia e o p-valor do teste *logrank* e o risco relativo para tal covariável. A Figura 14(b) apresenta a curva do risco acumulado empírico para a covariável Ferrovia. Novamente, segmentando pela covariável Ferrovia, verifica-se que a medida que o valor do lote aumenta o risco de vendê-lo também aumenta. Em outras palavras, conforme aumenta o valor do lote diminui a chance de não vendê-lo. Os lotes que não têm a sua acessibilidade ao centro prejudicado pela ferrovia possuem um risco de 73% maior de serem vendidos em relação aos lotes que possuem a acessibilidade ao centro prejudicada pela ferrovia.

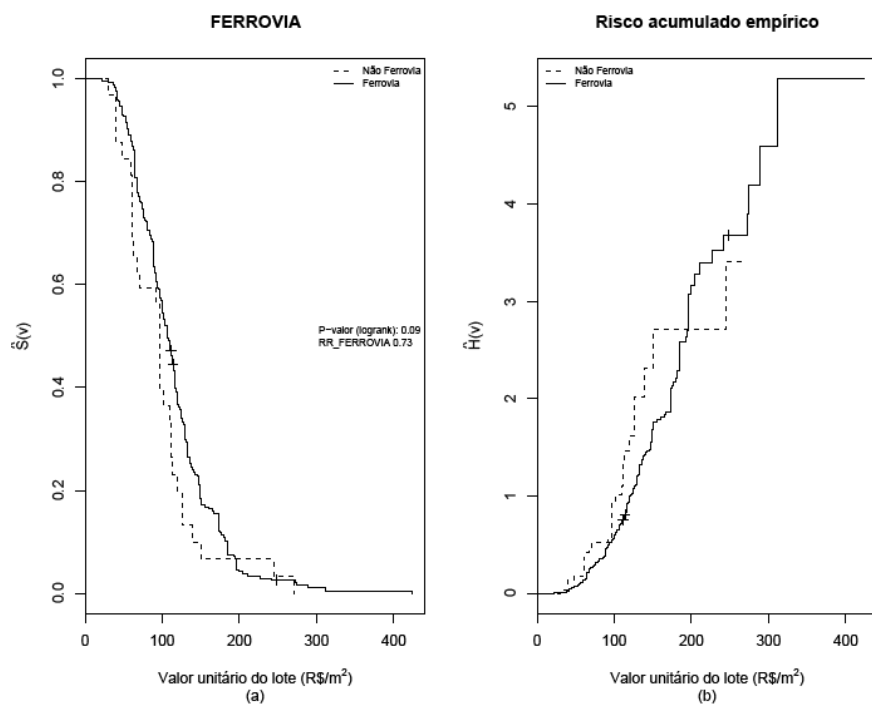


Figura 14. (a) Curva de permanência à venda estimada pelo método de Kaplan-Meier, o p-valor da estatística *logrank* e o risco relativo para a covariável Ferrovias, (b) Risco acumulado de venda empírico considerando a covariável Ferrovias.

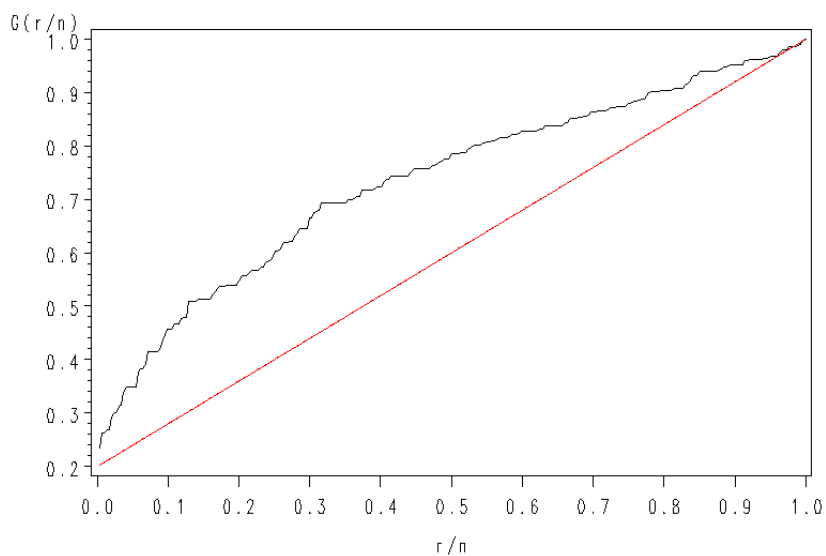


Figura 15. TTT-Plot para os valores unitários dos lotes de São Carlos, SP, no ano de 2005.

A Figura 15 apresenta uma concavidade na curva, indicando que a forma da função de risco é monótona crescente. Desta forma, uma distribuição candidata ao ajuste é a Weibull com parâmetro de forma maior que um. O TTT-Plot apresentado na Figura 15 foi gerado no

software SAS 9.0. Os comandos para gerar o TTT-Plot foram fornecidos pelo meu amigo André Yoshizumi.

A partir do TTT-Plot (Figura 15) adotamos o modelo de regressão Weibull no ajuste dos valores unitários dos lotes urbanos (R\$/m<sup>2</sup>) da cidade de São Carlos, SP, ano de 2005.

Para a construção do modelo de regressão Weibull, a variável resposta será o valor unitário do lote (R\$/m<sup>2</sup>) e como variáveis explicativas: as variáveis de localização e a área do lote (m<sup>2</sup>). Foram considerados apenas lotes com área igual ou inferior a 800m<sup>2</sup> enquanto que os lotes não vendidos em 2005 foram considerados censurados já que não experimentaram o evento de interesse (a venda), isto é, permaneceram à venda. Ressalta-se que estamos considerando censura à esquerda. Utilizamos as funções *survreg* e *Surv* do pacote *survival*, intrínseca ao software R. A função *Surv* permite especificarmos o tipo de censura, no nosso caso, censura à esquerda e a função *survreg* permite especificarmos a distribuição de interesse, no nosso caso, Weibull.

O próximo passo é a seleção das covariáveis que melhor explicam o valor unitário do lote (R\$/m<sup>2</sup>). Neste trabalho optou-se por utilizar uma estratégia de seleção de modelos derivada da proposta de Collet (1994). Ela é baseada no teste da razão de verossimilhanças (ver Seção 6.8.1). A Tabela 8 apresenta os passos dessa estratégia de seleção.

Tabela 8. Seleção de covariáveis usando o modelo Weibull considerando censura à esquerda.

Passos	Modelo	Log(Verossimilhança)	Estatística de teste (TRV)	Valor p
	NULO	-1444.9	-	-
	NUC PRINC	-1432.7	24.4	0.00000
	PLN CENTRAL	-1426.8	36.2	0.00000
	FERROVIA	-1443.9	2	0.17000
Passo 1	RODOVIA_WL	-1432	25.8	0.00000
	CONDO	-1436.3	17.2	0.00003
	FECHADO	-1440.6	8.6	0.00330
	ESTRIT_RESID	-1444.2	1.4	0.24000
	AREA	-1437.8	14.2	0.00016
	NUC_PRINC + PLN_CENTRAL + RODOVIA_WL + CONDO + FECHADO + AREA	-1382.6	-	-
	PLN_CENTRAL + RODOVIA_WL + CONDO + FECHADO + AREA	-1395.2	25.2	0.0000
	NUC_PRINC + RODOVIA_WL + CONDO + FECHADO + AREA	-1408	50.8	0.0000
Passo 2	NUC_PRINC + PLN_CENTRAL + CONDO + FECHADO + AREA	-1394.6	24	0.0000
	NUC_PRINC + PLN_CENTRAL + RODOVIA_WL + FECHADO + AREA	-1386.7	8.2	0.0042
	NUC_PRINC + PLN_CENTRAL + RODOVIA_WL + CONDO + AREA	-1383	0.8	0.3711
	NUC_PRINC + PLN_CENTRAL + RODOVIA_WL + CONDO + FECHADO	-1382.9	0.6	0.4386
Passo 3	NUC_PRINC + PLN_CENTRAL + RODOVIA_WL + CONDO	-1383.7	-	-

	NUC_PRINC + PLN_CENTRAL + RODOVIA_WL + CONDO + FECHADO	-1382.9	1.6	0.2059
	NUC_PRINC + PLN_CENTRAL + RODOVIA_WL + CONDO + AREA	-1383	1.4	0.2367
	NUC_PRINC + PLN_CENTRAL + RODOVIA_WL + CONDO	-1383.7	-	-
Passo 4	NUC_PRINC + PLN_CENTRAL + RODOVIA_WL + CONDO + FERROVIA	-1383.4	0.3	0.5839
	NUC_PRINC + PLN_CENTRAL + RODOVIA_WL + CONDO + ESTRIT_RESID	-1383.7	0	1.0000
	NUC_PRINC + PLN_CENTRAL + RODOVIA_WL + CONDO	-1383.7	-	-
	PLN_CENTRAL + RODOVIA_WL + CONDO	-1396.3	12.6	0.0004
Passo 5	NUC_PRINC + RODOVIA_WL + CONDO	-1412.1	28.4	0.0000
	NUC_PRINC + PLN_CENTRAL + CONDO	-1396.2	12.5	0.0004
	NUC_PRINC + PLN_CENTRAL + RODOVIA_WL	-1399.8	16.1	0.0000
Modelo Final	NUC_PRINC + PLN_CENTRAL + RODOVIA_WL + CONDO	-1383.7		

Ao final da estratégia de seleção de covariáveis (Tabela 8) obtém-se o seguinte modelo final

$$V_i = \text{Intercepto} + \text{NUC\_PRINC} + \text{PLN\_CENTRAL} + \text{RODOVIA\_WL} + \text{CONDO}$$

As estimativas dos parâmetros e os respectivos intervalos de confiança do modelo Weibull considerando censura à esquerda encontram-se na Tabela 9.

Tabela 9. Estimativa dos parâmetros, respectivos limites inferior e superior, do intervalo de confiança de 95% e amplitude do intervalo, para cada covariável.

Variável	Estimativa dos parâmetros	LI	LS	Amplitude
<b>Intercepto</b>	30.26	22.88	40.02	17.14
<b>NUC_PRINC</b>	2.06	1.66	2.56	0.9
<b>PLN_CENTRAL</b>	1.66	1.44	1.9	0.46
<b>RODOVIA_WL</b>	1.84	1.52	2.21	0.69
<b>CONDO</b>	1.4	1.26	1.59	0.33

A estimativa do parâmetro de forma do nosso modelo final é 2,73, maior que um, concordando com a informação obtida através do TTT-Plot (Figura 15) que nos mostrava a distribuição Weibull, com parâmetro de forma maior que um, uma distribuição plausível para o ajuste.

Podemos verificar que, para a regressão Weibull, a amplitude dos intervalos de confiança das estimativas dos parâmetros das covariáveis estatisticamente significativas é menor, comparada à amplitude dos intervalos de confiança das estimativas dos parâmetros das covariáveis estatisticamente significativas pelo modelo linear usual.

O modelo final através da regressão Weibull apresentou um número menor de covariáveis (mais parcimonioso) estatisticamente significativas quando comparado com o modelo final da regressão linear usual.



## 8. CONCLUSÃO

Na aplicação do ferramental de modelos lineares, o modelo linear usual com transformação raiz quadrada na variável resposta (valor total, em reais, do imóvel) mostrou ser adequado quanto à proposição de uma equação de regressão representativa da formação do valor de mercado dos lotes urbanos do município de São Carlos, SP, ano de 2005, cuja capacidade preditiva foi de aproximadamente 75%.

No estudo de simulação para o modelo linear usual, conclui-se que ao considerar uma variância baixa ou moderada, para as amostras com tamanhos iguais ou superiores a 50 todos os parâmetros apresentaram uma cobertura dentro dos limites nominais. Para uma variância alta, todos os parâmetros apresentaram uma cobertura dentro dos limites nominais a partir das amostras com tamanhos iguais ou superiores a 150.

Para a simulação do modelo Weibull, considerando a porcentagem de censuras entre 0% e 15% e amostras com tamanhos iguais ou superiores a 50 unidades de observação os parâmetros apresentaram probabilidade de cobertura dentro dos limites nominais. O aumento de censura, a partir de 30%, direciona a uma diminuição da probabilidade de cobertura dos parâmetros, para qualquer tamanho de amostra, mas esta diminuição não é maior do que 10% abaixo do limite nominal, exceto para o parâmetro  $\beta_0$  (intercepto).

A probabilidade de cobertura do intercepto ( $\beta_0$ ), a partir de 30% de censuras nas amostras, diminuiu à medida que aumentamos o tamanho da amostra.

Como esperado, pequenas amostras (tamanho 25) apresentaram baixo desempenho na simulação de probabilidade de cobertura dos parâmetros de ambos os modelos (linear usual e Weibull). O pesquisador que utilizar o modelo Weibull deve ficar atento, também, para o número de censuras em sua amostra. Assim sendo, o pesquisador deverá sempre ficar atento ao tamanho de amostra (ambos os modelos) e o número de censuras (modelo Weibull) a serem utilizados em sua pesquisa. O planejamento amostral feito com qualidade ajuda a contornar o problema.

A seguir apresentamos, para título de ilustração, as estimativas referentes aos valores de dois lotes utilizando ambas as metodologias.

Tabela 10. Valor estimado, considerando os dois modelos construídos anteriormente, para dois lotes urbanos em condições distintas de negociação.

<b>Censuras</b>	<b>val_unit</b>	<b>Estim_sobrev</b>	<b>Estim_mod_lin_usual</b>	<b>Incremento_sobrev</b>	<b>Incremento_lin_usual</b>
<b>1</b>	100.00	114.58	84.10	15%	-16%
<b>0</b>	34.18	62.43	118.11	83%	245%

Através da Tabela 10 verificamos que a metodologia de análise sobrevivência apresentou um preço mais real para lotes em negociação (censurado). Sendo assim, a modelagem linear usual superestimou o preço para lotes em negociação (censurado). Para os lotes efetivamente negociados (não censurado) os dois modelos apresentaram estimativas similares.

A metodologia de análise de sobrevivência é mais flexível, pois permite incluir no processo de modelagem os lotes efetivamente negociados (não censurado) e os lotes em negociação (censurado).

Ainda é cedo para afirmar sobre a efetiva melhora dos modelos baseados em análise de sobrevivência comparados aos modelos lineares usuais. Há muito trabalho a ser feito, porém já visualizamos boas perspectivas nesse caminho que estamos seguindo. As metodologias se complementam cabendo ao pesquisador saber qual irá utilizar.

Cumpramos observar que, apesar do tamanho da amostra ser moderado (284 lotes), o poder de predição do modelo é considerável. Esta afirmação é especialmente válida quando consideramos que existem variáveis “micro” de importância expressiva que não foram consideradas, tais como: formato do lote, posição no interior da quadra, proximidade a sub-centros comerciais e serviços, dentre outras características que se referem à individualidade do lote urbano.

Assim, o desenvolvimento de futuras pesquisas aponta para o estudo da influência destas variáveis menores, no interior das regiões homogêneas propostas e ilustradas na Figura 3. Outras propostas de estudos futuros pode ser a utilização de outras técnicas estatísticas, matemáticas e computacionais, tais como, a inferência Bayesiana, os Modelos Lineares Generalizados, Algoritmos Genéticos e modelos heurísticos, tais como, as Redes Neurais Artificiais.

## REFERÊNCIAS BIBLIOGRÁFICAS

AARSET, M. V. **How to identify a bathtub hazard rate.** 1987. IEEE Transactions on Reliability, v. 36, p106–108.

ASSOCIAÇÃO BRASILEIRA DE NORMAS TÉCNICAS. **NBR 5676: avaliação de imóveis urbanos.** Rio de Janeiro, 1989.

ASSOCIAÇÃO BRASILEIRA DE NORMAS TÉCNICAS. **NBR 14653-2: avaliação de bens – Parte 2 – Imóveis Urbanos.** Rio de Janeiro, 2004.

AGUIRRE, A. **Uma nota sobre a transformação Box-Cox.** 1997. 21p. Belo Horizonte: Cedeplar, Universidade Federal de Minas Gerais.

AKAIKE, H. **A new look at statistical model identification.** 1974. IEEE Transactions on Automatic Control, AU-19, p 716-722.

ATKINSON, A.C. **Two graphical display for outlying and influential observations in regression.** 1981. Biometrika, v. 68, p 13-20.

BELSLEY, D. A.; KUH, E.; WELSCH, R. E. **Regression Diagnostics.** 1980. New York: John Wiley.

BOX, G.E.P.; COX, D.R. **An analysis of transformation.** 1964. J.R Statist. Soc. B, v. 26, p 211-252.

BÖHMER, P. E. **Theorie der unabhängigen.** 1912. Wahrscheinlichkeiten Rapports Memories et Proces-verbaux de Septieme Congres International d'Actuaires, v. 2, p 327–343.

BRONDINO, N. C. M. **Estudo da influência da Acessibilidade no Valor de Lotes Urbanos Através do Uso de Redes Neurais**. 1999. 146p. Tese (Doutorado em Engenharia Civil) – Departamento de Transportes, Escola de Engenharia de São Carlos.

CHARNES, A.; COOPER, W. W.; LEWIN, A. Y.; SEIFORD, L.M. **Data Envelopment Analysis: Theory, Methodology and Applications**. USA: Kluwer Academic Publishers, 1995.

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

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

CORDEIRO, G.M.; LIMA NETO, E. A. **Modelos Paramétricos**. Caxambu, MG: XVI SINAPE, 2004, 246 p.

COX, D. R.; OAKES, D. **Analysis of Survival Data**. London: Chapman and Hall, 1984.

COX, D. R.; SNELL, E. J. **A general definition of residuals (with discussion)**. Journal of the Royal Statistical Society B, 1968, v. 30, p 248-275.

DANTAS, R. A. **Engenharia de Avaliações, Uma Introdução à Metodologia Científica**. São Paulo, SP: PINI, 1998, v. 1 , 242 p.

DEMÉTRIO, C.G.B. e ZOCCHI, S.S. Modelos de regressão na experimentação agrônômica. Apostila. Departamento de Ciências Exatas, ESALQ/USP. 2007.

EVERITT, B. S.; HOTHORN, T. **A Handbook of Statistical Analyses Using R**. London: Chapman and Hall, 2006.

FERREIRA, J. F. **Proposta de tratamento da variável localização em modelos inferenciais de avaliação imobiliária para municípios médios.** 2007. Dissertação (Mestrado em Engenharia Urbana) – Departamento de Engenharia Civil, Universidade Federal de São Carlos. Exame de defesa realizado em 18/12/2007.

INSTITUTO BRASILEIRO DE AVALIAÇÕES E PERÍCIAS DE ENGENHARIA NO ESTADO DE SÃO PAULO - IBAPE/SP. **Normas de perícia.** São Paulo, 1984.

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA – IBGE. **Base de Informações Municipais** [CD-ROM]. 3. ed. Rio de Janeiro: IBGE, 2002.

JOHNSON, R. A.; WICHERN, D. W. **Applied Multivariate Statistical Analysis.** 4 ed. Upper Saddle River, NJ; Prentice-Hall, Inc, 1998.

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

KAPLAN, E. L. e MEIER, P. (1958). **Nonparametric estimation from incomplete observations.** Journal of the American Statistical Association, 1958, v. 53, p 457-481.

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

LEE, E. T. **Statistical Methods for Survival Data Analysis.** New York: John Wiley and Sons, 2000.

LOUZADA-NETO, F. **Polyhazard regression models for lifetime data.** Biometrics, 1999, v. 55, p 1281–1285.

LOUZADA-NETO, F.; MAZUCHELI, J.; ACHCAR, J. A. **Introdução à Análise de Sobrevivência e Confiabilidade**. Minicurso: XXVIII Jornadas Nacionales de Estadística, Antofagasta, Chile, 2001.

MANTEL, N. **Evaluation of survival data and two new rank order statistics arising in its consideration**. Cancer Chemotherapy Reports, 1966, v. 50, p 163–170.

MANTEL, N. E HAENSZEL, W. **Statistical aspects of the analysis of data from retrospective studies of disease**. Journal of the National Cancer Institute, 1959, v. 22, p 719–748.

MYERS, R. H.; MONTGOMERY, D. C.; VINING, G. G. **Generalized Linear Models: with applications in engineering and the sciences**. New York: John Wiley and Sons. 2002. 342 p.

MUDHOLKAR, G. S., SRIVASTAVA, D. K., E KOLLIA, G. D. **A generalization of the Weibull distribution with application to the analysis of survival data**. Journal of the American Statistical Association, 1996, 91(436):1575-1583.

NETER, J.; KUTNER, M. H.; NACHTSEIM, C. J.; WASSERMAN, W. Applied Linear Regression Models, 3rd Edition. Irwin, Illinois, 1996.

PAULA, G. A. Modelos de Regressão com apoio computacional (versão preliminar). São Paulo, SP: Instituto de Matemática e Estatística, 253 p., 2004.

R Development Core Team (2008). 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>.

SÃO CARLOS. Prefeitura Municipal – Divisão de Cadastro Imobiliário. Base de dados cadastrais, 2006.

SÃO CARLOS. Site da Prefeitura Municipal. Disponível em <http://www.saocarlos.sp.gov.br/>. Acesso em 14 de outubro de 2008.

SAS INSTITUTE (2005). Cary, NC, USA.

WEIBULL, W. A. A statistical distribution of wide applicability. *Journal of Applied Mechanics*, 1951, v. 18, p 293-297.

## APÊNDICE A – Código para a modelagem dos dados

```

# entrada dos dados

dados = read.table('E:\\Mestrado 2007\\MLG_ajuste\\amostra_treino.txt',
head=T)
attach(dados)

require(MASS)

#ESTOU CODIFICANDO AS VARIÁVEIS NOVAMENTE POIS TIREI OS PONTOS INFLUENTES
VAL_TOT = subset(dados[c(-7,-9,-14,-41,-86,-158,-152,-40,-93,-137,-91,-68,-
34,-47,-48)],,select=VAL_TOT)
AREA = subset(dados[c(-7,-9,-14,-41,-86,-158,-152,-40,-93,-137,-91,-68,-
34,-47,-48)],,select=AREA)
NUC_PRINC = subset(dados[c(-7,-9,-14,-41,-86,-158,-152,-40,-93,-137,-91,-
68,-34,-47,-48)],,select=NUC_PRINC)
PLN_CENTRAL = subset(dados[c(-7,-9,-14,-41,-86,-158,-152,-40,-93,-137,-91,-
68,-34,-47,-48)],,select=PLN_CENTRAL)
FERROVIA = subset(dados[c(-7,-9,-14,-41,-86,-158,-152,-40,-93,-137,-91,-
68,-34,-47,-48)],,select=FERROVIA)
RODOVIA_WL = subset(dados[c(-7,-9,-14,-41,-86,-158,-152,-40,-93,-137,-91,-
68,-34,-47,-48)],,select=RODOVIA_WL)
ENCOSTA = subset(dados[c(-7,-9,-14,-41,-86,-158,-152,-40,-93,-137,-91,-68,-
34,-47,-48)],,select=ENCOSTA)
CONDO = subset(dados[c(-7,-9,-14,-41,-86,-158,-152,-40,-93,-137,-91,-68,-
34,-47,-48)],,select=CONDO)
FECHADO = subset(dados[c(-7,-9,-14,-41,-86,-158,-152,-40,-93,-137,-91,-68,-
34,-47,-48)],,select=FECHADO)
ESTRIT_RESID = subset(dados[c(-7,-9,-14,-41,-86,-158,-152,-40,-93,-137,-
91,-68,-34,-47,-48)],,select=ESTRIT_RESID)

##PARTE 1. Ajuste do modelo

VAL_TOT = VAL_TOT[,1]
AREA = AREA[,1]
NUC_PRINC = NUC_PRINC[,1]
PLN_CENTRAL = PLN_CENTRAL[,1]
FERROVIA = FERROVIA[,1]
RODOVIA_WL = RODOVIA_WL[,1]
ENCOSTA = ENCOSTA[,1]
CONDO = CONDO[,1]
FECHADO = FECHADO[,1]
ESTRIT_RESID =ESTRIT_RESID[,1]

dados.ajuste = data.frame(VAL_TOT, AREA, NUC_PRINC, PLN_CENTRAL, FERROVIA,
RODOVIA_WL, ENCOSTA,
CONDO, FECHADO, ESTRIT_RESID)

##### MODELAGEM VIA MODELOS LINEARES USUAIS
fit.model.inicial = glm(VAL_TOT ~ AREA + NUC_PRINC + PLN_CENTRAL + FERROVIA
+ RODOVIA_WL + ENCOSTA + CONDO + FECHADO + ESTRIT_RESID + NUC_PRINC*AREA +
PLN_CENTRAL*AREA + FERROVIA*AREA + RODOVIA_WL*AREA + ENCOSTA*AREA +
CONDO*AREA + FECHADO*AREA + ESTRIT_RESID*AREA )

## TRANSFORMAÇÃO BOX-COX

boxcox(fit.model.inicial, lambda = seq(-0.5, 1.5, 1/10) )

## MODELAGEM VIA MODELOS LINEARES USUAIS COM TRANSFORMAÇÃO ##RAÍZ QUADRADA
NA VARIÁVEL RESPOSTA (SUGERIDA POR BOX COX)

```



```

fit.model.step = glm((VAL_TOT)^(0.5) ~ AREA + NUC_PRINC + PLN_CENTRAL +
FERROVIA + RODOVIA_WL + ENCOSTA + CONDO + FECHADO + ESTRIT_RESID +
NUC_PRINC*AREA + PLN_CENTRAL*AREA + FERROVIA*AREA + RODOVIA_WL*AREA +
ENCOSTA*AREA + CONDO*AREA + FECHADO*AREA + ESTRIT_RESID*AREA)

# SELEÇÃO DE VARIÁVEIS - STEPWISE
model.step = stepAIC(fit.model.step)
model.step$anova

## Modelo sugerido para iniciarmos o processo de modelagem via modelo
NORMAL
fit.model.s = glm(VAL_TOT^(0.5) ~ AREA + NUC_PRINC + PLN_CENTRAL + FERROVIA
+ RODOVIA_WL + CONDO + FECHADO + ESTRIT_RESID + AREA:NUC_PRINC +
AREA:ESTRIT_RESID + AREA:PLN_CENTRAL + AREA:FECHADO)
summary(fit.model.s)

## Modelo FINAL via modelo NORMAL
fit.model = glm(VAL_TOT^(0.5) ~ NUC_PRINC + PLN_CENTRAL + FERROVIA +
RODOVIA_WL + CONDO + FECHADO + ESTRIT_RESID + AREA:NUC_PRINC +
AREA:ESTRIT_RESID + AREA:PLN_CENTRAL + AREA:FECHADO)
summary(fit.model)

#Construção dos Intervalos de Confiança via matriz de informação de Fisher
observada#
p = ncol(model.matrix(fit.model))
alfa = 0.05

IC_inf = matrix(NA, 1, p)
IC_sup = matrix(NA, 1, p)

for (j in 1:length(diag(vcov(fit.model)))){
  IC_inf[1,j]<-(fit.model$coeff)[j] + qnorm(alfa/2) *
sqrt(diag(vcov(fit.model))[j])
  IC_sup[1,j]<-(fit.model$coeff)[j] + qnorm(1-(alfa/2)) *
sqrt(diag(vcov(fit.model))[j])
}

IC = data.frame(Inf=t(round(IC_inf,2)), sup=t(round(IC_sup,2)), row.names =
names(fit.model$coef))

# Graficos - Diagnóstico
#-----
# Para rodar a função a seguir, especifique no objeto "fit.model" a
# saída do ajuste do modelo de regressão linear simples.
#
# A saída terá oito gráficos: de pontos alavanca, influentes e
# de resíduos.
#
# Desenvolvida por: Fernando Lucambio e Fabiola Araújo Costa
#
# Para salvar em pdf e importar no Microsoft Word
# eu dividi os oitos gráficos em dois arquivos em pdf com
# 4 gráficos cada arquivo.
#-----
#

graficos.diagnostico1=function(fit.model)
{
X = model.matrix(fit.model)
n = nrow(X)

```

```

p = ncol(X)
H=X%*(chol2inv(t(X)%*%X))%*%t(X)
h = hat(X)
tst = rstudent(fit.model)
tsi = rstandard(fit.model)
#
# Primeiro grupo de gráficos
#
pdf('C:\\Documents and Settings\\User\\Meus documentos\\Gui
Ferraudo\\Estatística\\Mestrado
2007\\MLG_ajuste\\gráfico_diagnostics_1.pdf', width=10, height=10)
par(mfrow=c(2,2))
plot(h, xlab="Índice", ylab="Medida h", main="Pontos Alavanca",
ylim=range(h,na.rm=T,finite=T), pch=16)
abline(2*p/n,0, lty=2)
pontos = ifelse(h>2*p/n,labels(h)," ")
text(h,pontos,pos=4,xpd=TRUE)
#
plot(fit.model$fitted.values, sqrt(abs(tst)), xlab="Valores Preditos",
ylab=expression(sqrt(abs("Residuo Studentizado"))), main="Pontos
Aberrantes",
ylim=range(sqrt(abs(tst)),na.rm=T,finite=T), pch=16)
abline(sqrt(2), 0, lty=2)
pontos = ifelse(sqrt(abs(tsi))>sqrt(2),labels(fit.model$fitted.values)," ")
text(fit.model$fitted.values, sqrt(abs(tsi)),pontos,pos=4,xpd=TRUE)
#
plot(tsi, xlab="Índice", ylab="Resíduo Padronizado", main="Pontos
Aberrantes",
ylim=range(tsi,na.rm=T,finite=T), pch=16)
abline(2,0,lty=2)
abline(-2,0,lty=2)
pontos = ifelse((tsi> -2) & (tsi<2)," ",labels(tsi))
text(tsi,pontos,pos=4,xpd=TRUE)
#
plot(fitted(fit.model), tsi, xlab="Valores Preditos", ylab="Resíduo
Padronizado",
main="Homocedasticidade", ylim=range(tsi,na.rm=T,finite=T), pch=16)
abline(2,0,lty=2)
abline(-2,0,lty=2)
pontos = ifelse((tsi> -2) & (tsi<2)," ",labels(fitted(fit.model)))
text(fitted(fit.model),tsi,pontos,pos=4,xpd=TRUE)
dev.off()
}

graficos.diagnostics1(fit.model)

graficos.diagnostics2=function(fit.model)
{#
# Segundo grupo de gráficos
#
X = model.matrix(fit.model)
n = nrow(X)
p = ncol(X)
H=X%*(chol2inv(t(X)%*%X))%*%t(X)
h = hat(X)
tst = rstudent(fit.model)
tsi = rstandard(fit.model)

pdf('C:\\Documents and Settings\\User\\Meus documentos\\Gui
Ferraudo\\Estatística\\Mestrado
2007\\MLG_ajuste\\gráfico_diagnostics_2.pdf', width=10, height=10)

```

```

readline("Presione ENTER para continuar .....")
ident = diag(n)
epsilon = matrix(0,n,1000)
e = matrix(0,n,1000)
e1 = numeric(n)
e2 = numeric(n)
#
for(i in 1:1000){
  epsilon[,i] = rnorm(n,0,1)
  e[,i] = (ident - H)%*%epsilon[,i]
  u = diag(ident - H)
  e[,i] = e[,i]/sqrt(u)
  e[,i] = sort(e[,i]) }
#
for(i in 1:n){
  eo = sort(e[i,])
  e1[i] = eo[5]
  e2[i] = eo[995] }
#
med = apply(e,1,mean)
faixa = range(tsi,e1,e2)
#

par(pty="s")
par(mfrow=c(2,2))
qqnorm(tsi, xlab="Percentis da N(0,1)", ylab="Residuo Studentizado",
ylim=faixa, pch=16)
par(new=T)
qqnorm(e1, axes=F, xlab="", ylab="", type="l", ylim=faixa, lty=1)
par(new=T)
qqnorm(e2, axes=F, xlab="", ylab="", type="l", ylim=faixa, lty=1)
par(new=T)
qqnorm(med, axes=F, xlab="", ylab="", type="l", ylim=faixa, lty=2)
#
distancia.cook = cooks.distance(fit.model)
plot(distancia.cook, xlab="Índice", ylab="Distância de Cook", main="Pontos
Influentes",
ylim=range(distancia.cook,na.rm=T,finite=T), pch=16)
abline(1,0,lty=2)
pontos = ifelse(distancia.cook>1,labels(distancia.cook)," ")
text(distancia.cook,pontos,pos=4,xpd=TRUE)
#
dfits = dffits(fit.model)
plot(dfits, xlab="Índice", ylab=expression(plain(DFFIT)[i]), main="Pontos
Influentes",
ylim=range(dfits,na.rm=T,finite=T), pch=16)
abline(2*(p/(n-p))^0.5,0,lty=2)
pontos = ifelse(dfits>2*(p/(n-p))^0.5,labels(dfits)," ")
text(dfits,pontos,pos=4,xpd=TRUE)
#
C = abs(tsi)*(((n-p)*h)/(p*(1-h)))^0.5
plot(C, xlab="Índice", ylab=expression(plain(C)[i]), main="Pontos
Influentes",
ylim=range(C,na.rm=T,finite=T), pch=16)
abline(2,0,lty=2)
pontos = ifelse(C>2,labels(C)," ")
text(C,pontos,pos=4,xpd=TRUE)
par(mfrow=c(1,1))
dev.off()
}

```

```

graficos.diagnostico2(fit.modelo)

## Teste de Shapiro-Wilk para verificar a normalidade dos resíduos
shapiro.test(fit.modelo$res)

pdf('C:\\Documents and Settings\\User\\Meus documentos\\Gui
Ferraudo\\Estatística\\Mestrado 2007\\MLG_ajuste\\gráfico_pred x
obs_normal.pdf')
plot(VAL_TOT, (fit.modelo$fitted)^2, xlab="Valores Observados", ylab="Valores
Preditos^2", main="Valores Observados vs Valores Preditos^2")
abline(lsf(fit(VAL_TOT, (fit.modelo$fitted)^2), col='red')
dev.off()

#####

##PARTE 2. Validação do Modelo
valida = read.table('E:\\Mestrado 2007\\MLG_ajuste\\amostra_valid.txt',
head=T)
attach(valida)

a = 97.2 - 82.18*NUC_PRINC - 4.87*PLN_CENTRAL + 27.63*CONDO +
17.79*FERROVIA + 42.63*RODOVIA_WL - 89.86*FECHADO + 39.39*ESTRIT_RESID +
0.33*(NUC_PRINC*AREA) + 0.06*(PLN_CENTRAL*AREA) - 0.12*(ESTRIT_RESID*AREA)
+ 0.27*(FECHADO*AREA)
val_tot_imovel = a^2

## Diferença Relativa Média

DRM_sqrt_normal = ( ( sum( abs((VAL_TOT - val_tot_imovel)/VAL_TOT) ) ) /
length(val_tot_imovel) ) * 100
DRM_sqrt_normal
summary(VAL_TOT)
summary(val_tot_imovel)

#Fim

##Modelagem via Análise de Sobrevida considerando censura à esquerda

#entrada do dados
lotes = read.table("C:\\Documents and Settings\\User\\Meus documentos\\Gui
Ferraudo\\Estatística\\Mestrado
2007\\Neto\\modelagem_sobrevida\\lotes_SaoCarlos.txt", head=T)
attach(lotes)

##Consideraremos os lotes com área inferior ou igual a 800m.

val_lote = subset(lotes[c(-7,-9,-14,-41,-86,-158,-152,-40,-93,-137,-91,-
68,-34,-47,-48),], aarea < 801, select=val_unit) ## variável q representa o
valor do lote por metro quadrado. Como se fosse o tempo num modelo de
sobrevida.
censuras = subset(lotes[c(-7,-9,-14,-41,-86,-158,-152,-40,-93,-137,-91,-
68,-34,-47,-48),], aarea < 801, select=OF_TRANS)
AREA = subset(lotes[c(-7,-9,-14,-41,-86,-158,-152,-40,-93,-137,-91,-68,-
34,-47,-48),], aarea < 801, select=aarea)
NUC_PRINC = subset(lotes[c(-7,-9,-14,-41,-86,-158,-152,-40,-93,-137,-91,-
68,-34,-47,-48),], aarea < 801, select=NUC_PRINC)
PLN_CENTRAL = subset(lotes[c(-7,-9,-14,-41,-86,-158,-152,-40,-93,-137,-91,-
68,-34,-47,-48),], aarea < 801, select=PLN_CENTRAL)
FERROVIA = subset(lotes[c(-7,-9,-14,-41,-86,-158,-152,-40,-93,-137,-91,-
68,-34,-47,-48),], aarea < 801, select=FERROVIA)

```

```

RODOVIA_WL = subset(lotes[c(-7,-9,-14,-41,-86,-158,-152,-40,-93,-137,-91,-
68,-34,-47,-48),],aarea < 801, select=RODOVIA_WL)
ENCOSTA = subset(lotes[c(-7,-9,-14,-41,-86,-158,-152,-40,-93,-137,-91,-68,-
34,-47,-48),],aarea < 801, select=ENCOSTA)
CONDO = subset(lotes[c(-7,-9,-14,-41,-86,-158,-152,-40,-93,-137,-91,-68,-
34,-47,-48),],aarea < 801, select=CONDO)
FECHADO = subset(lotes[c(-7,-9,-14,-41,-86,-158,-152,-40,-93,-137,-91,-68,-
34,-47,-48),],aarea < 801, select=FECHADO)
ESTRIT_RESID = subset(lotes[c(-7,-9,-14,-41,-86,-158,-152,-40,-93,-137,-
91,-68,-34,-47,-48),],aarea < 801, select=ESTRIT_RESID)

val_lote = val_lote[,1]
censuras = censuras[,1]
AREA = AREA[,1]
NUC_PRINC = NUC_PRINC[,1]
PLN_CENTRAL = PLN_CENTRAL[,1]
FERROVIA = FERROVIA[,1]
RODOVIA_WL = RODOVIA_WL[,1]
ENCOSTA = ENCOSTA[,1]
CONDO = CONDO[,1]
FECHADO = FECHADO[,1]
ESTRIT_RESID =ESTRIT_RESID[,1]

#Procedimentos não-paramétricos
##Estimativa da Curva de Sobrevida: Kaplan-Meier
ekm = survfit(Surv(val_lote, censuras))
summary(ekm)

#Gráficos
pdf("C:\\Documents and Settings\\User\\Meus documentos\\Gui
Ferraud\\Estatística\\Mestrado
2007\\Neto\\modelagem_sobrevida\\KM_RiscoEmp.pdf", width=12, height=8)
par(mfrow = c(1,2))
plot(ekm, lty = c(2,1), conf.int=T, xlab = expression(paste("Valor unitário
do lote (R$/", m^2, ")")), ylab = expression(paste(hat(S),"(v)")), main =
"Kaplan-Meier")
title(sub = '(a)')
plot(ekm, lty = c(2,1), conf.int=T, fun = "cumhaz", xlab =
expression(paste("Valor unitário do lote (R$/", m^2, ")")), ylab =
expression(paste(hat(H),"(v)")), main = "Risco acumulado empírico")
title(sub = '(b)')
dev.off()

## Teste logrank e gráficos
pdf("C:\\Documents and Settings\\User\\Meus documentos\\Gui
Ferraud\\Estatística\\Mestrado
2007\\Neto\\modelagem_sobrevida\\KM_logank_FERROVIA.pdf", width=10,
height=8)
par(mfrow=c(1,2))
ekm_FERROVIA = survfit(Surv(val_lote, censuras) ~ FERROVIA, conf.type =
'plain')
summary(ekm_FERROVIA)
logrank_FERROVIA = survdiff(Surv(val_lote, censuras) ~ FERROVIA, rho=0)
RR_FERROVIA =
(logrank_FERROVIA[[2]][2]/logrank_FERROVIA[[3]][2])/(logrank_FERROVIA[[2]][
1]/logrank_FERROVIA[[3]][1])
plot(ekm_FERROVIA, conf.int = F, lty = c(2,1), xlab =
expression(paste("Valor unitário do lote (R$/", m^2, ")")), ylab =
expression(paste(hat(S),"(v)")), main="FERROVIA")
legend('topright', lty = c(2,1), c("Não Ferrovia", "Ferrovia"), lwd = 1,
bty = 'n', cex=0.7)

```

```

legend('right', paste(c("P-valor (logrank):", "RR_FERROVIA"), c(round(1 -
pchisq(logrank_FERROVIA[[5]], 1),2), round(RR_FERROVIA,2))), bty = 'n',
cex=0.7)
title(sub = '(a)')
plot(ekm_FERROVIA, lty = c(2,1), conf.int=F, fun = "cumhaz", xlab =
expression(paste("Valor unitário do lote (R$/", m^2, ")")), ylab =
expression(paste(hat(H), "(v)")), main = "Risco acumulado empírico")
legend('topleft', lty = c(2,1), c("Não Ferrovia", "Ferrovia"), lwd = 1, bty
= 'n', cex=0.7)
title(sub = '(b)')
dev.off()

pdf("C:\\Documents and Settings\\User\\Meus documentos\\Gui
Ferrauda\\Estatística\\Mestrado
2007\\Neto\\modelagem_sobrevivência\\KM_logank_FERRO_NUCPRINC_PLNCENTRAL_RO
DWL.pdf", width=10, height=8)
par(mfrow=c(2,2))
ekm_FERROVIA = survfit(Surv(val_lote, censuras) ~ FERROVIA, conf.type =
'plain')
summary(ekm_FERROVIA)
logrank_FERROVIA = survdiff(Surv(val_lote, censuras) ~ FERROVIA, rho=0)
RR_FERROVIA =
(logrank_FERROVIA[[2]][2]/logrank_FERROVIA[[3]][2])/(logrank_FERROVIA[[2]][
1]/logrank_FERROVIA[[3]][1])
plot(ekm_FERROVIA, conf.int = F, lty = c(2,1), xlab =
expression(paste("Valor unitário do lote (R$/", m^2, ")")), ylab = 'S(v)
estimada', main="FERROVIA")
legend('topright', lty = c(2,1), c("Não Ferrovia", "Ferrovia"), lwd = 1,
bty = 'n')
legend('right', paste(c("P-valor (logrank):", "RR_FERROVIA"), c(round(1 -
pchisq(logrank_FERROVIA[[5]], 1),2), round(RR_FERROVIA,2))), bty = 'n')
#dev.off()

#NÚCLEO PRINCIPAL. Kaplan-Meier e teste logrank
ekm_NUC_PRINC = survfit(Surv(val_lote, censuras) ~ NUC_PRINC, conf.type =
'plain')
summary(ekm_NUC_PRINC)
logrank_NUC_PRINC = survdiff(Surv(val_lote, censuras) ~ NUC_PRINC, rho=0)
RR_NUC_PRINC =
(logrank_NUC_PRINC[[2]][2]/logrank_NUC_PRINC[[3]][2])/(logrank_NUC_PRINC[[2]
][1]/logrank_NUC_PRINC[[3]][1])
plot(ekm_NUC_PRINC, conf.int = F, lty = c(2,1), xlab =
expression(paste("Valor unitário do lote (R$/", m^2, ")")), ylab = 'S(v)
estimada', main="NÚCLEO PRINCIPAL")
legend('topright', lty = c(2,1), c("Não Núcleo Principal", "Núcleo
Principal"), lwd = 1, bty = 'n')
legend('right', paste(c("P-valor (logrank):", "RR_NUC_PRINC"), c(round(1 -
pchisq(logrank_NUC_PRINC[[5]], 1),2), round(RR_NUC_PRINC,2))), bty = 'n')

#PLANÍCIE CENTRAL. Kaplan-Meier e teste logrank
ekm_PLN_CENTRAL = survfit(Surv(val_lote, censuras) ~ PLN_CENTRAL, conf.type
= 'plain')
summary(ekm_PLN_CENTRAL)
logrank_PLN_CENTRAL = survdiff(Surv(val_lote, censuras) ~ PLN_CENTRAL,
rho=0)
RR_PLN_CENTRAL =
(logrank_PLN_CENTRAL[[2]][2]/logrank_PLN_CENTRAL[[3]][2])/(logrank_PLN_CENT
RAL[[2]][1]/logrank_PLN_CENTRAL[[3]][1])
plot(ekm_PLN_CENTRAL, conf.int = F, lty = c(2,1), xlab =
expression(paste("Valor unitário do lote (R$/", m^2, ")")), ylab = 'S(v)
estimada', main="PLANÍCIE CENTRAL")

```

```

legend('topright', lty = c(2,1), c("Não Planície Central", "Planície
Central"), lwd = 1, bty = 'n')
legend('right', paste(c("P-valor (logrank):", "RR_PLN_CENTRAL"), c(round(1
- pchisq(logrank_PLN_CENTRAL[[5]], 1),2), round(RR_PLN_CENTRAL,2))), bty =
'n')

#RODOVIA WL. Kaplan-Meier e teste logrank
ekm_RODOVIA_WL = survfit(Surv(val_lote, censuras) ~ RODOVIA_WL, conf.type =
'plain')
summary(ekm_RODOVIA_WL)
logrank_RODOVIA_WL = survdiff(Surv(val_lote, censuras) ~ RODOVIA_WL, rho=0)
RR_RODOVIA_WL =
(logrank_RODOVIA_WL[[2]][2]/logrank_RODOVIA_WL[[3]][2])/(logrank_RODOVIA_WL
[[2]][1]/logrank_RODOVIA_WL[[3]][1])
plot(ekm_RODOVIA_WL, conf.int = F, lty = c(2,1), xlab =
expression(paste("Valor unitário do lote (R$/", m^2, ")")), ylab = 'S(v)
estimada', main="RODOVIA WL")
legend('topright', lty = c(2,1), c("Não Rodovia WL", "Rodovia WL"), lwd =
1, bty = 'n')
legend('right', paste(c("P-valor (logrank):", "RR_RODOVIA_WL"), c(round(1
- pchisq(logrank_RODOVIA_WL[[5]], 1),2), round(RR_RODOVIA_WL,2))), bty =
'n')
dev.off()

pdf("C:\\Documents and Settings\\User\\Meus documentos\\Gui
Ferrauo\\Estatística\\Mestrado
2007\\Neto\\modelagem_sobrevivência\\KM_logank_CONDO_FECHADO ESTRITRESID.pd
f", width=10, height=8)
par(mfrow=c(2,2))
#CONDOMINIO. Kaplan-Meier e teste logrank
ekm_CONDO = survfit(Surv(val_lote, censuras) ~ CONDO, conf.type = 'plain')
summary(ekm_CONDO)
logrank_CONDO = survdiff(Surv(val_lote, censuras) ~ CONDO, rho=0)
RR_CONDO =
(logrank_CONDO[[2]][2]/logrank_CONDO[[3]][2])/(logrank_CONDO[[2]][1]/logran
k_CONDO[[3]][1])
plot(ekm_CONDO, conf.int = F, lty = c(2,1), xlab = expression(paste("Valor
unitário do lote (R$/", m^2, ")")), ylab = 'S(v) estimada',
main="CONDOMÍNIO")
legend('topright', lty = c(2,1), c("Não Condominio", "Condominio"), lwd =
1, bty = 'n')
legend('right', paste(c("P-valor (logrank):", "RR_CONDO"), c(round(1 -
pchisq(logrank_CONDO[[5]], 1),2), round(RR_CONDO,2))), bty = 'n')

#FECHADO. Kaplan-Meier e teste logrank
ekm_FECHADO = survfit(Surv(val_lote, censuras) ~ FECHADO, conf.type =
'plain')
summary(ekm_FECHADO)
logrank_FECHADO = survdiff(Surv(val_lote, censuras) ~ FECHADO, rho=0)
RR_FECHADO =
(logrank_FECHADO[[2]][2]/logrank_FECHADO[[3]][2])/(logrank_FECHADO[[2]][1]/
logrank_FECHADO[[3]][1])
plot(ekm_FECHADO, conf.int = F, lty = c(2,1), xlab =
expression(paste("Valor unitário do lote (R$/", m^2, ")")), ylab = 'S(v)
estimada', main="FECHADO")
legend('topright', lty = c(2,1), c("Não Fechado", "Fechado"), lwd = 1, bty
= 'n')
legend('right', paste(c("P-valor (logrank):", "RR_FECHADO"), c(round(1 -
pchisq(logrank_FECHADO[[5]], 1),2), round(RR_FECHADO,2))), bty = 'n')

#ESTRITAMENTE RESIDENCIAL. Kaplan-Meier e teste logrank

```

```

ekm ESTRIT_RESID = survfit(Surv(val_lote, censuras) ~ ESTRIT_RESID,
conf.type = 'plain')
summary(ekm ESTRIT_RESID)
logrank ESTRIT_RESID = survdiff(Surv(val_lote, censuras) ~ ESTRIT_RESID,
rho=0)
RR ESTRIT_RESID =
(logrank ESTRIT_RESID[[2]][2]/logrank ESTRIT_RESID[[3]][2])/(logrank ESTRIT
_RESID[[2]][1]/logrank ESTRIT_RESID[[3]][1])
plot(ekm ESTRIT_RESID, conf.int = F, lty = c(2,1), xlab =
expression(paste("Valor unitário do lote (R$/", m^2, ")")), ylab = 'S(v)
estimada', main="ESTRIT_RESID")
legend('topright', lty = c(2,1), c("Não Estritamente Residencial",
"Estritamente Residencial"), lwd = 1, bty = 'n')
legend('right', paste(c("P-valor (logrank):", "RR ESTRIT_RESID"),
c(round(1 - pchisq(logrank ESTRIT_RESID[[5]], 1), 2),
round(RR ESTRIT_RESID, 2))), bty = 'n')
dev.off()

#Modelo de Regressão Weibull
require(survival)#carregando pacote
yfit = survreg(Surv((val_lote), censuras, type='left') ~ AREA + NUC_PRINC +
PLN_CENTRAL + FERROVIA + RODOVIA_WL + CONDO + FECHADO + ESTRIT_RESID,
dist='weibull')
gama = 1/yfit$scale #parâmetro de forma da distribuição Weibull
alpha = exp(yfit$coef) #parâmetro de escala da distribuição Weibull

#p-valor do TRV, onde p é o número de parâmetros no modelo sob
#investigação e trv é o valor da estatística do teste da razão de
#verossimilhança. O resultado do objeto "p_val_trv" é utilizado na
#estratégia de seleção de covariáveis de Collet.
p_val_trv = 1 - pchisq(trv, p)

##Construção dos Intervalos de Confiança para os parâmetros
IC = exp(confint(yfit))

#Fim

```



## APÊNDICE B – Código para o estudo de simulação

```
#Simulação da probabilidade de cobertura dos parâmetros de uma #regressão
linear usual
```

```
cobertura_beta = function(repli, n, alfa, beta0, beta1, beta2, beta3,
beta4, beta5, desvpad){
```

```
x1 = matrix(NA,repli,n)
x2 = matrix(NA,repli,n)
x3 = matrix(NA,repli,n)
x4 = matrix(NA,repli,n)
x5 = matrix(NA,repli,n)
y = matrix(NA,repli,n)
yfit = list()
```

```
IC_inf = matrix(NA, repli, 6)
IC_sup = matrix(NA, repli, 6)
```

```
contbeta0 = numeric()
contbeta1 = numeric()
contbeta2 = numeric()
contbeta3 = numeric()
contbeta4 = numeric()
contbeta5 = numeric()
```

```
contbeta0_inf = numeric()
contbeta1_inf = numeric()
contbeta2_inf = numeric()
contbeta3_inf = numeric()
contbeta4_inf = numeric()
contbeta5_inf = numeric()
```

```
contbeta0_sup = numeric()
contbeta1_sup = numeric()
contbeta2_sup = numeric()
contbeta3_sup = numeric()
contbeta4_sup = numeric()
contbeta5_sup = numeric()
```

```
cob_beta0 = numeric()
cob_beta1 = numeric()
cob_beta2 = numeric()
cob_beta3 = numeric()
cob_beta4 = numeric()
cob_beta5 = numeric()
```

```
cob_beta0_inf = numeric()
cob_beta1_inf = numeric()
cob_beta2_inf = numeric()
cob_beta3_inf = numeric()
cob_beta4_inf = numeric()
cob_beta5_inf = numeric()
```

```
cob_beta0_sup = numeric()
cob_beta1_sup = numeric()
cob_beta2_sup = numeric()
cob_beta3_sup = numeric()
cob_beta4_sup = numeric()
cob_beta5_sup = numeric()
```

```

for(i in 1:repli){

x1[i,] = rbinom(n,1,0.4)
x2[i,] = rbinom(n,1,0.7)
x3[i,] = rbinom(n,1,0.5)
x4[i,] = rbinom(n,1,0.3)
x5[i,] = rweibull(n, 8, 750)

y[i,] = rnorm(n,beta0 + beta1*x1[i,] + beta2*x2[i,] + beta3*x3[i,] +
beta4*x4[i,] + beta5*x5[i,], desvpad)

yfit[[i]] = glm( (y[i,]) ~ x1[i,] + x2[i,] + x3[i,] + x4[i,] + x5[i,] )
  for (j in 1:length(diag(vcov(yfit[[i]])))){
    IC_inf[i,j]<-(yfit[[i]]$coeff)[j] + qnorm(alfa/2) *
sqrt(diag(vcov(yfit[[i]]))[j])
    IC_sup[i,j]<-(yfit[[i]]$coeff)[j] + qnorm(1-(alfa/2)) *
sqrt(diag(vcov(yfit[[i]]))[j])
  }

## Esta condição testa se cada beta(i) estão dentro do IC para os
respectivos beta(i) estimados ##
contbeta0[i] = c (if ( beta0 >= IC_inf[i,1] & beta0 <= IC_sup[i,1] ) 1 else
0)
contbeta1[i] = c (if ( beta1 >= IC_inf[i,2] & beta1 <= IC_sup[i,2] ) 1 else
0)
contbeta2[i] = c (if ( beta2 >= IC_inf[i,3] & beta2 <= IC_sup[i,3] ) 1 else
0)
contbeta3[i] = c (if ( beta3 >= IC_inf[i,4] & beta3 <= IC_sup[i,4] ) 1 else
0)
contbeta4[i] = c (if ( beta4 >= IC_inf[i,5] & beta4 <= IC_sup[i,5] ) 1 else
0)
contbeta5[i] = c (if ( beta5 >= IC_inf[i,6] & beta5 <= IC_sup[i,6] ) 1 else
0)

## Sabendo que os parâmetros não pertencem ao IC, verificado na condição
acima. Esta condição testa se beta(i) saíram por "baixo" do IC para ##cada
beta(i) estimados, ##
contbeta0_inf[i] = c (if ( contbeta0[i] == 0 & beta0 <= IC_inf[i,1] ) 1
else 0)
contbeta1_inf[i] = c (if ( contbeta1[i] == 0 & beta1 <= IC_inf[i,2] ) 1
else 0)
contbeta2_inf[i] = c (if ( contbeta2[i] == 0 & beta2 <= IC_inf[i,3] ) 1
else 0)
contbeta3_inf[i] = c (if ( contbeta3[i] == 0 & beta3 <= IC_inf[i,4] ) 1
else 0)
contbeta4_inf[i] = c (if ( contbeta4[i] == 0 & beta4 <= IC_inf[i,5] ) 1
else 0)
contbeta5_inf[i] = c (if ( contbeta5[i] == 0 & beta5 <= IC_inf[i,6] ) 1
else 0)

## Sabendo que os parâmetros não pertencem ao IC, verificado na condição
acima. Esta condição testa se beta(i) saíram por "cima" do IC para ##cada
beta(i) estimados, ##
contbeta0_sup[i] = c (if ( contbeta0[i] == 0 & beta0 >= IC_sup[i,1] ) 1
else 0)
contbeta1_sup[i] = c (if ( contbeta1[i] == 0 & beta1 >= IC_sup[i,2] ) 1
else 0)
contbeta2_sup[i] = c (if ( contbeta2[i] == 0 & beta2 >= IC_sup[i,3] ) 1
else 0)

```

```

contbeta3_sup[i] = c (if ( contbeta3[i] == 0 & beta3 >= IC_sup[i,4] ) 1
else 0)
contbeta4_sup[i] = c (if ( contbeta4[i] == 0 & beta4 >= IC_sup[i,5] ) 1
else 0)
contbeta5_sup[i] = c (if ( contbeta5[i] == 0 & beta5 >= IC_sup[i,6] ) 1
else 0)

}

cob_beta0 = mean(contbeta0)
cob_beta1 = mean(contbeta1)
cob_beta2 = mean(contbeta2)
cob_beta3 = mean(contbeta3)
cob_beta4 = mean(contbeta4)
cob_beta5 = mean(contbeta5)

cob_beta0_inf = mean(contbeta0_inf)
cob_beta1_inf = mean(contbeta1_inf)
cob_beta2_inf = mean(contbeta2_inf)
cob_beta3_inf = mean(contbeta3_inf)
cob_beta4_inf = mean(contbeta4_inf)
cob_beta5_inf = mean(contbeta5_inf)

cob_beta0_sup = mean(contbeta0_sup)
cob_beta1_sup = mean(contbeta1_sup)
cob_beta2_sup = mean(contbeta2_sup)
cob_beta3_sup = mean(contbeta3_sup)
cob_beta4_sup = mean(contbeta4_sup)
cob_beta5_sup = mean(contbeta5_sup)

matrix(c(cob_beta0_inf,   cob_beta1_inf,   cob_beta2_inf,   cob_beta3_inf,
cob_beta4_inf, cob_beta5_inf,
cob_beta0, cob_beta1, cob_beta2, cob_beta3, cob_beta4, cob_beta5,
cob_beta0_sup, cob_beta1_sup, cob_beta2_sup, cob_beta3_sup, cob_beta4_sup,
cob_beta5_sup),
nrow=6, ncol=3, byrow = F,
dimnames = list(c('beta0', 'beta1', 'beta2', 'beta3', 'beta4', 'beta5'),
c('Limite Inferior', 'Centro', 'Limite Superior') ) )

}

##### Amostra 1: n = 25 #####
cobertura_25 = cobertura_beta(repli = 1000, n = 25, alfa = 0.05, beta0 =
9.7, beta1 = -8.1, beta2 = 3, beta3= -2.5, beta4 = 1.8, beta5 = 4.3,
desvpad = 10)
cobertura_25

##### Amostra 2: n = 50 #####
cobertura_50 = cobertura_beta(repli = 1000, n = 50, alfa = 0.05, beta0 =
9.7, beta1 = -8.1, beta2 = 3, beta3= -2.5, beta4 = 1.8, beta5 = 4.3,
desvpad = 10)
cobertura_50

##### Amostra 3: n = 150 #####
cobertura_150 = cobertura_beta(repli = 1000, n = 150, alfa = 0.05, beta0 =
9.7, beta1 = -8.1, beta2 = 3, beta3= -2.5, beta4 = 1.8, beta5 = 4.3,
desvpad = 10)
cobertura_150

##### Amostra 4: n = 500 #####

```

```

cobertura_500 = cobertura_beta(repli = 1000, n = 500, alfa = 0.05, beta0 =
9.7, beta1 = -8.1, beta2 = 3, beta3= -2.5, beta4 = 1.8, beta5 = 4.3,
desvpad = 10)
cobertura_500

##### Amostra 5: n = 1000 #####
cobertura_1000 = cobertura_beta(repli = 1000, n = 1000, alfa = 0.05, beta0 =
9.7, beta1 = -8.1, beta2 = 3, beta3= -2.5, beta4 = 1.8, beta5 = 4.3,
desvpad = 10)
cobertura_1000

##Salvando a matriz de saída em um arquivo texto para poder ser aberto no
Excel##
write.table(cobertura_25, "E:\\Mestrado 2007\\Neto\\cob_25.txt", sep = ';',
dec = ",")
write.table(cobertura_50, "E:\\Mestrado 2007\\Neto\\cob_50.txt", sep = ';',
dec = ",")
write.table(cobertura_150, "E:\\Mestrado 2007\\Neto\\cob_150.txt", sep =
';', dec = ",")
write.table(cobertura_500, "E:\\Mestrado 2007\\Neto\\cob_500.txt", sep =
';', dec = ",")
write.table(cobertura_1000, "E:\\Mestrado 2007\\Neto\\cob_1000.txt", sep =
';', dec = ",")

### Apresentando os resultados obtidos em forma de tabela ###

prob_cobertura=matrix(c(cobertura_25[1,2],cobertura_50[1,2],cobertura_150[1
,2],cobertura_500[1,2],cobertura_1000[1,2],
cobertura_25[2,2],cobertura_50[2,2],cobertura_150[2,2],cobertura_500[2,2],c
obertura_1000[2,2],
cobertura_25[3,2],cobertura_50[3,2],cobertura_150[3,2],cobertura_500[3,2],c
obertura_1000[3,2],
cobertura_25[4,2],cobertura_50[4,2],cobertura_150[4,2],cobertura_500[4,2],c
obertura_1000[4,2],
cobertura_25[5,2],cobertura_50[5,2],cobertura_150[5,2],cobertura_500[5,2],c
obertura_1000[5,2],
cobertura_25[6,2],cobertura_50[6,2],cobertura_150[6,2],cobertura_500[6,2],c
obertura_1000[6,2]), ncol=6,
dimnames = list( c( '25', '50', '150', '500', '1000'), c('Beta 0', 'Beta
1', 'Beta 2', 'Beta 3', 'Beta 4', 'Beta 5') ) )

prob_cobertura

write.table(prob_cobertura, "E:\\Mestrado 2007\\Neto\\prob_cobertura.txt",
sep = ';', dec = ",")

## Fazendo o gráfico para melhor visualização da tabela armazenada no
objeto "prob_cobertura" ##

fator = as.factor(c(25, 50, 150, 500, 1000))

pdf("C:\\Documents and Settings\\User\\Meus documentos\\Gui
Ferraudo\\Estatística\\Mestrado
2007\\Neto\\Correção_Simulação_10_06_2008\\graf_beta0_prob_cobertura.pdf")
plot(fator, prob_cobertura[,1], type = 'p', main=
expression(bold(beta[0])), xlab="Tamanhos de Amostra (n)",ylab=expression(
paste("Vício para ", beta[0]), sep=" " ), ylim = c(0.9, 1))
abline(h=0.9365) ## LI ##
abline(h=0.95) ## LC ##
abline(h=0.9635) ## LS ##
dev.off()

```

```

pdf("C:\\Documents and Settings\\User\\Meus documentos\\Gui
Ferraudo\\Estatística\\Mestrado
2007\\Neto\\Correção_Simulação_10_06_2008\\graf_beta1_prob_cobertura.pdf")
plot(fator,      prob_cobertura[,2],      type      =      'p',      main=
expression(bold(beta[1])), xlab="Tamanhos de Amostra (n)",ylab=expression(
paste("Vício para ", beta[1]), sep=" " ), ylim = c(0.9, 1))
abline(h=0.9365) ## LI ##
abline(h=0.95) ## LC ##
abline(h=0.9635) ## LS ##
dev.off()

pdf("C:\\Documents and Settings\\User\\Meus documentos\\Gui
Ferraudo\\Estatística\\Mestrado
2007\\Neto\\Correção_Simulação_10_06_2008\\graf_beta2_prob_cobertura.pdf")
plot(fator,      prob_cobertura[,3],      type      =      'p',      main=
expression(bold(beta[2])), xlab="Tamanhos de Amostra (n)",ylab=expression(
paste("Vício para ", beta[2]), sep=" " ), ylim = c(0.9, 1))
abline(h=0.9365) ## LI ##
abline(h=0.95) ## LC ##
abline(h=0.9635) ## LS ##
dev.off()

pdf("C:\\Documents and Settings\\User\\Meus documentos\\Gui
Ferraudo\\Estatística\\Mestrado
2007\\Neto\\Correção_Simulação_10_06_2008\\graf_beta3_prob_cobertura.pdf")
plot(fator,      prob_cobertura[,4],      type      =      'p',      main=
expression(bold(beta[3])), xlab="Tamanhos de Amostra (n)",ylab=expression(
paste("Vício para ", beta[3]), sep=" " ), ylim = c(0.9, 1))
abline(h=0.9365) ## LI ##
abline(h=0.95) ## LC ##
abline(h=0.9635) ## LS ##
dev.off()

pdf("C:\\Documents and Settings\\User\\Meus documentos\\Gui
Ferraudo\\Estatística\\Mestrado
2007\\Neto\\Correção_Simulação_10_06_2008\\graf_beta4_prob_cobertura.pdf")
plot(fator,      prob_cobertura[,5],      type      =      'p',      main=
expression(bold(beta[4])), xlab="Tamanhos de Amostra (n)",ylab=expression(
paste("Vício para ", beta[4]), sep=" " ), ylim = c(0.9, 1))
abline(h=0.9365) ## LI ##
abline(h=0.95) ## LC ##
abline(h=0.9635) ## LS ##
dev.off()

pdf("C:\\Documents and Settings\\User\\Meus documentos\\Gui
Ferraudo\\Estatística\\Mestrado
2007\\Neto\\Correção_Simulação_10_06_2008\\graf_beta5_prob_cobertura.pdf")
plot(fator,      prob_cobertura[,6],      type      =      'p',      main=
expression(bold(beta[5])), xlab="Tamanhos de Amostra (n)",ylab=expression(
paste("Vício para ", beta[5]), sep=" " ), ylim = c(0.9, 1))
abline(h=0.9365) ## LI ##
abline(h=0.95) ## LC ##
abline(h=0.9635) ## LS ##
dev.off()

##Fim

# Cálculo do vício das estimativas dos parâmetros de uma regressão #linear
usual

```

```

set.seed(354739)

vicio_beta = function(repli, n, beta0, beta1, beta2, beta3, beta4, beta5,
desvpad){

x1 = matrix(NA,repli,n)
x2 = matrix(NA,repli,n)
x3 = matrix(NA,repli,n)
x4 = matrix(NA,repli,n)
x5 = matrix(NA,repli,n)
y = matrix(NA,repli,n)
yfit = list()

vicio_beta0 = numeric()
vicio_beta1 = numeric()
vicio_beta2 = numeric()
vicio_beta3 = numeric()
vicio_beta4 = numeric()
vicio_beta5 = numeric()

coef = matrix(NA, repli, 6)
med_coef_b0 = numeric()
med_coef_b1 = numeric()
med_coef_b2 = numeric()
med_coef_b3 = numeric()
med_coef_b4 = numeric()
med_coef_b5 = numeric()

for(i in 1:repli){

x1[i,] = rbinom(n,1,0.4)
x2[i,] = rbinom(n,1,0.7)
x3[i,] = rbinom(n,1,0.5)
x4[i,] = rbinom(n,1,0.3)
x5[i,] = rweibull(n, 8, 750)

y[i,] = rnorm(n,beta0 + beta1*x1[i,] + beta2*x2[i,] + beta3*x3[i,] +
beta4*x4[i,] + beta5*x5[i,], desvpad)

yfit[[i]] = glm( y[i,] ~ x1[i,] + x2[i,] + x3[i,] + x4[i,] + x5[i,] )
coef[i,] = yfit[[i]]$coef
}

med_coef_b0 = apply(coef, 2, mean)[1]
med_coef_b1 = apply(coef, 2, mean)[2]
med_coef_b2 = apply(coef, 2, mean)[3]
med_coef_b3 = apply(coef, 2, mean)[4]
med_coef_b4 = apply(coef, 2, mean)[5]
med_coef_b5 = apply(coef, 2, mean)[6]

vicio_beta0 = abs(med_coef_b0 - beta0)
vicio_beta1 = abs(med_coef_b1 - beta1)
vicio_beta2 = abs(med_coef_b2 - beta2)
vicio_beta3 = abs(med_coef_b3 - beta3)
vicio_beta4 = abs(med_coef_b4 - beta4)
vicio_beta5 = abs(med_coef_b5 - beta5)

matrix(c(vicio_beta0, vicio_beta1, vicio_beta2, vicio_beta3, vicio_beta4,
vicio_beta5),
nrow=6, ncol=1,

```

```

dimnames = list(c('beta0', 'beta1', 'beta2', 'beta3', 'beta4', 'beta5'),
c('Vicio') ) )

}

##### Amostra 1: n = 25 #####
vicio_25 = vicio_beta(repli = 1000, n = 25, beta0 = 9.7, beta1 = -8.1,
beta2 = 3, beta3= -2.5, beta4 = 1.8, beta5 = 4.3, desvpad = 10)
vicio_25

##### Amostra 2: n = 50 #####
vicio_50 = vicio_beta(repli = 1000, n = 50, beta0 = 9.7, beta1 = -8.1,
beta2 = 2.7, beta3= -2.5, beta4 = 1.8, beta5 = 4.3, desvpad = 10)
vicio_50

##### Amostra 3: n = 150 #####
vicio_150 = vicio_beta(repli = 1000, n = 150, beta0 = 9.7, beta1 = -8.1,
beta2 = 2.7, beta3= -2.5, beta4 = 1.8, beta5 = 4.3, desvpad = 10)
vicio_150

##### Amostra 4: n = 500 #####
vicio_500 = vicio_beta(repli = 1000, n = 500, beta0 = 9.7, beta1 = -8.1,
beta2 = 2.7, beta3= -2.5, beta4 = 1.8, beta5 = 4.3, desvpad = 10)
vicio_500

##### Amostra 5: n = 1000 #####
vicio_1000 = vicio_beta(repli = 1000, n = 500, beta0 = 9.7, beta1 = -8.1,
beta2 = 2.7, beta3= -2.5, beta4 = 1.8, beta5 = 4.3, desvpad = 10)
vicio_1000

##Salvando a matriz de saída em um arquivo texto para poder ser aberto no
Excel##
write.table(vicio_25,      "C:\\Documents and Settings\\User\\Meus
documentos\\Gui          Ferraudo\\Estatística\\Mestrado
2007\\Neto\\Correção_Simulação_10_06_2008\\vicio_25.txt", sep = ';', dec =
".")
write.table(vicio_50,      "C:\\Documents and Settings\\User\\Meus
documentos\\Gui          Ferraudo\\Estatística\\Mestrado
2007\\Neto\\Correção_Simulação_10_06_2008\\vicio_50.txt", sep = ';', dec =
".")
write.table(vicio_150,      "C:\\Documents and Settings\\User\\Meus
documentos\\Gui          Ferraudo\\Estatística\\Mestrado
2007\\Neto\\Correção_Simulação_10_06_2008\\vicio_150.txt", sep = ';', dec =
".")
write.table(vicio_500,      "C:\\Documents and Settings\\User\\Meus
documentos\\Gui          Ferraudo\\Estatística\\Mestrado
2007\\Neto\\Correção_Simulação_10_06_2008\\vicio_500.txt", sep = ';', dec =
".")
write.table(vicio_1000,      "C:\\Documents and Settings\\User\\Meus
documentos\\Gui          Ferraudo\\Estatística\\Mestrado
2007\\Neto\\Correção_Simulação_10_06_2008\\vicio_1000.txt", sep = ';', dec
= ".")

### Apresentando os resultados obtidos em forma de tabela ###

vicio
matrix(c(vicio_25[1],vicio_50[1],vicio_150[1],vicio_500[1],vicio_1000[1],
vicio_25[2],vicio_50[2],vicio_150[2],vicio_500[2],vicio_1000[2],
vicio_25[3],vicio_50[3],vicio_150[3],vicio_500[3],vicio_1000[3],

```

```

vicio_25[4],vicio_50[4],vicio_150[4],vicio_500[4],vicio_1000[4],
vicio_25[5],vicio_50[5],vicio_150[5],vicio_500[5],vicio_1000[5],
vicio_25[6],vicio_50[6],vicio_150[6],vicio_500[6],vicio_1000[6]), ncol=6,
dimnames = list( c( '25', '50', '150', '500', '1000'), c('Beta 0', 'Beta
1', 'Beta 2', 'Beta 3', 'Beta 4', 'Beta 5') ) )

vicio

write.table(vicio, "C:\\Documents and Settings\\User\\Meus documentos\\Gui
Ferraudo\\Estatística\\Mestrado
2007\\Neto\\Correção_Simulação_10_06_2008\\vicio.txt", sep = ';', dec =
".")

## Fazendo o gráfico para melhor visualização da tabela armazenada no
#objeto "vicio"

fator = as.factor(c(25, 50, 150, 500, 1000))

pdf("C:\\Documents and Settings\\User\\Meus documentos\\Gui
Ferraudo\\Estatística\\Mestrado
2007\\Neto\\Correção_Simulação_10_06_2008\\graf_beta0_vicio.pdf")
plot(fator, vicio[,1], type = 'p', main= expression(bold(beta[0])),
xlab="Tamanhos de Amostra (n)",ylab=expression( paste("Vício para ",
beta[0]), sep=" " ), ylim = c(0, 70))
abline(h=0, col='red', lty=2)
dev.off()

pdf("C:\\Documents and Settings\\User\\Meus documentos\\Gui
Ferraudo\\Estatística\\Mestrado
2007\\Neto\\Correção_Simulação_10_06_2008\\graf_beta1_vicio.pdf")
plot(fator, vicio[,2], type = 'p', main= expression(bold(beta[1])),
xlab="Tamanhos de Amostra (n)",ylab=expression( paste("Vício para ",
beta[1]), sep=" " ), ylim = c(0, 70))
abline(h=0, col='red', lty=2)
dev.off()

pdf("C:\\Documents and Settings\\User\\Meus documentos\\Gui
Ferraudo\\Estatística\\Mestrado
2007\\Neto\\Correção_Simulação_10_06_2008\\graf_beta2_vicio.pdf")
plot(fator, vicio[,3], type = 'p', main= expression(bold(beta[2])),
xlab="Tamanhos de Amostra (n)",ylab=expression( paste("Vício para ",
beta[2]), sep=" " ), ylim = c(0, 70))
abline(h=0, col='red', lty=2)
dev.off()

pdf("C:\\Documents and Settings\\User\\Meus documentos\\Gui
Ferraudo\\Estatística\\Mestrado
2007\\Neto\\Correção_Simulação_10_06_2008\\graf_beta3_vicio.pdf")
plot(fator, vicio[,4], type = 'p', main= expression(bold(beta[3])),
xlab="Tamanhos de Amostra (n)",ylab=expression( paste("Vício para ",
beta[3]), sep=" " ), ylim = c(0, 70))
abline(h=0, col='red', lty=2)
dev.off()

pdf("C:\\Documents and Settings\\User\\Meus documentos\\Gui
Ferraudo\\Estatística\\Mestrado
2007\\Neto\\Correção_Simulação_10_06_2008\\graf_beta4_vicio.pdf")

```



```

plot(fator, vicio[,5], type = 'p', main= expression(bold(beta[4])),
xlab="Tamanhos de Amostra (n)",ylab=expression( paste("Vício para ",
beta[4]), sep=" " ), ylim = c(0, 70))
abline(h=0, col='red', lty=2)
dev.off()

pdf("C:\\Documents and Settings\\User\\Meus documentos\\Gui
Ferraudo\\Estatística\\Mestrado
2007\\Neto\\Correção_Simulação_10_06_2008\\graf_beta5_vicio.pdf")
plot(fator, vicio[,6], type = 'p', main= expression(bold(beta[5])),
xlab="Tamanhos de Amostra (n)",ylab=expression( paste("Vício para ",
beta[5]), sep=" " ), ylim = c(0, 70))
abline(h=0, col='red', lty=2)
dev.off()

##Fim

#Simulação da probabilidade de cobertura dos parâmetros de uma #regressão
Weibull considerando censura à esquerda

set.seed(354739)

cobertura_beta = function(repli, n, alfa, beta0, beta1, beta2, beta3,
beta4, beta5, gama, censura){

require(survival)

x1 = matrix(NA,repli,n)
x2 = matrix(NA,repli,n)
x3 = matrix(NA,repli,n)
x4 = matrix(NA,repli,n)
x5 = matrix(NA,repli,n)
cens = matrix(NA,repli,n)
y = matrix(NA,repli,n)
yfit = list()

IC_inf = matrix(NA, repli, 7)
IC_sup = matrix(NA, repli, 7)
IC = list()

contbeta0 = numeric()
contbeta1 = numeric()
contbeta2 = numeric()
contbeta3 = numeric()
contbeta4 = numeric()
contbeta5 = numeric()

contbeta0_inf = numeric()
contbeta1_inf = numeric()
contbeta2_inf = numeric()
contbeta3_inf = numeric()
contbeta4_inf = numeric()
contbeta5_inf = numeric()

contbeta0_sup = numeric()
contbeta1_sup = numeric()
contbeta2_sup = numeric()
contbeta3_sup = numeric()
contbeta4_sup = numeric()
contbeta5_sup = numeric()

```

```

cob_beta0 = numeric()
cob_beta1 = numeric()
cob_beta2 = numeric()
cob_beta3 = numeric()
cob_beta4 = numeric()
cob_beta5 = numeric()

cob_beta0_inf = numeric()
cob_beta1_inf = numeric()
cob_beta2_inf = numeric()
cob_beta3_inf = numeric()
cob_beta4_inf = numeric()
cob_beta5_inf = numeric()

cob_beta0_sup = numeric()
cob_beta1_sup = numeric()
cob_beta2_sup = numeric()
cob_beta3_sup = numeric()
cob_beta4_sup = numeric()
cob_beta5_sup = numeric()

for(i in 1:repli){

x1[i,] = rbinom(n,1,0.4)
x2[i,] = rbinom(n,1,0.7)
x3[i,] = rbinom(n,1,0.5)
x4[i,] = rbinom(n,1,0.3)
x5[i,] = rweibull(n, 8, 750)
cens[i,] = rbinom(n, 1, (1-censura))

y[i,] = rweibull(n, scale = exp(beta0 + beta1*x1[i,] + beta2*x2[i,] +
beta3*x3[i,] + beta4*x4[i,] + beta5*x5[i,]), shape = (1/gama))

yfit[[i]] = survreg(Surv(y[i,], cens[i,], type='left') ~ x1[i,] + x2[i,] +
x3[i,] + x4[i,] + x5[i,], scale = gama, dist='weibull')

##Construção dos Intervalos de Confiança
IC[[i]] = confint(yfit[[i]])

##Construção dos Intervalos de Confiança
for (j in 1:nrow(IC[[i]])){
IC_inf[i,j] = ((IC[[i]])[,1])[j]
IC_sup[i,j] = ((IC[[i]])[,2])[j]
}

## Esta condição testa se cada beta(i) estão dentro do IC para os
respectivos beta(i) estimados ##
contbeta0[i] = c (if ( beta0 >= IC_inf[i,1] & beta0 <= IC_sup[i,1] ) 1 else
0)
contbeta1[i] = c (if ( beta1 >= IC_inf[i,2] & beta1 <= IC_sup[i,2] ) 1 else
0)
contbeta2[i] = c (if ( beta2 >= IC_inf[i,3] & beta2 <= IC_sup[i,3] ) 1 else
0)
contbeta3[i] = c (if ( beta3 >= IC_inf[i,4] & beta3 <= IC_sup[i,4] ) 1 else
0)
contbeta4[i] = c (if ( beta4 >= IC_inf[i,5] & beta4 <= IC_sup[i,5] ) 1 else
0)
contbeta5[i] = c (if ( beta5 >= IC_inf[i,6] & beta5 <= IC_sup[i,6] ) 1 else
0)

```

```

## Sabendo que os parâmetros não pertencem ao IC, verificado na condição
acima. Esta condição testa se beta(i) saíram por "baixo" do IC para ##cada
beta(i) estimados, ##
contbeta0_inf[i] = c (if ( contbeta0[i] == 0 & beta0 <= IC_inf[i,1] ) 1
else 0)
contbeta1_inf[i] = c (if ( contbeta1[i] == 0 & beta1 <= IC_inf[i,2] ) 1
else 0)
contbeta2_inf[i] = c (if ( contbeta2[i] == 0 & beta2 <= IC_inf[i,3] ) 1
else 0)
contbeta3_inf[i] = c (if ( contbeta3[i] == 0 & beta3 <= IC_inf[i,4] ) 1
else 0)
contbeta4_inf[i] = c (if ( contbeta4[i] == 0 & beta4 <= IC_inf[i,5] ) 1
else 0)
contbeta5_inf[i] = c (if ( contbeta5[i] == 0 & beta5 <= IC_inf[i,6] ) 1
else 0)

## Sabendo que os parâmetros não pertencem ao IC, verificado na condição
acima. Esta condição testa se beta(i) saíram por "cima" do IC para ##cada
beta(i) estimados, ##
contbeta0_sup[i] = c (if ( contbeta0[i] == 0 & beta0 >= IC_sup[i,1] ) 1
else 0)
contbeta1_sup[i] = c (if ( contbeta1[i] == 0 & beta1 >= IC_sup[i,2] ) 1
else 0)
contbeta2_sup[i] = c (if ( contbeta2[i] == 0 & beta2 >= IC_sup[i,3] ) 1
else 0)
contbeta3_sup[i] = c (if ( contbeta3[i] == 0 & beta3 >= IC_sup[i,4] ) 1
else 0)
contbeta4_sup[i] = c (if ( contbeta4[i] == 0 & beta4 >= IC_sup[i,5] ) 1
else 0)
contbeta5_sup[i] = c (if ( contbeta5[i] == 0 & beta5 >= IC_sup[i,6] ) 1
else 0)

}

cob_beta0 = mean(contbeta0)
cob_beta1 = mean(contbeta1)
cob_beta2 = mean(contbeta2)
cob_beta3 = mean(contbeta3)
cob_beta4 = mean(contbeta4)
cob_beta5 = mean(contbeta5)

cob_beta0_inf = mean(contbeta0_inf)
cob_beta1_inf = mean(contbeta1_inf)
cob_beta2_inf = mean(contbeta2_inf)
cob_beta3_inf = mean(contbeta3_inf)
cob_beta4_inf = mean(contbeta4_inf)
cob_beta5_inf = mean(contbeta5_inf)

cob_beta0_sup = mean(contbeta0_sup)
cob_beta1_sup = mean(contbeta1_sup)
cob_beta2_sup = mean(contbeta2_sup)
cob_beta3_sup = mean(contbeta3_sup)
cob_beta4_sup = mean(contbeta4_sup)
cob_beta5_sup = mean(contbeta5_sup)

matrix(c(cob_beta0_inf,   cob_beta1_inf,   cob_beta2_inf,   cob_beta3_inf,
cob_beta4_inf, cob_beta5_inf,
cob_beta0, cob_beta1, cob_beta2, cob_beta3, cob_beta4, cob_beta5,

```

```

cob_beta0_sup, cob_beta1_sup, cob_beta2_sup, cob_beta3_sup, cob_beta4_sup,
cob_beta5_sup),
nrow=6, ncol=3, byrow = F,
dimnames = list(c('beta0', 'beta1', 'beta2', 'beta3', 'beta4', 'beta5'),
c('Limite Inferior', 'Centro', 'Limite Superior') ) )

}

##### Amostra 1: n = 25 #####
cobertura_25 = cobertura_beta(repli = 1000, n = 25, beta0 = 2, beta1 = 1,
beta2 = 3, beta3= 5, beta4 = 7, beta5 = 0.0008, gama = 2, censura = 0.01)
cobertura_25

##### Amostra 2: n = 50 #####
cobertura_50 = cobertura_beta(repli = 1000, n = 50, beta0 = 2, beta1 = 1,
beta2 = 3, beta3= 5, beta4 = 7, beta5 = 0.0008, gama = 2, censura = 0.01)
cobertura_50

##### Amostra 3: n = 150 #####
cobertura_150 = cobertura_beta(repli = 1000, n = 150, beta0 = 2, beta1 = 1,
beta2 = 3, beta3= 5, beta4 = 7, beta5 = 0.0008, gama = 2, censura = 0.01)
cobertura_150

##### Amostra 4: n = 500 #####
cobertura_500 = cobertura_beta(repli = 1000, n = 500, beta0 = 2, beta1 = 1,
beta2 = 3, beta3= 5, beta4 = 7, beta5 = 0.0008, gama = 2, censura = 0.01)
cobertura_500

##### Amostra 5: n = 1000 #####
cobertura_1000 = cobertura_beta(repli = 1000, n = 1000, beta0 = 2, beta1 =
1, beta2 = 3, beta3= 5, beta4 = 7, beta5 = 0.0008, gama = 2, censura =
0.01)
cobertura_1000

##Salvando a matriz de saída em um arquivo texto para poder ser aberto no
Excel##
write.table(cobertura_25, "C:\\Documents and Settings\\User\\Meus
documentos\\Gui Ferraudo\\Estatística\\Mestrado
2007\\Neto\\Simulação_Sobrevivência_2008\\crescente\\cob_25.txt", sep =
';', dec = ".")
write.table(cobertura_50, "C:\\Documents and Settings\\User\\Meus
documentos\\Gui Ferraudo\\Estatística\\Mestrado
2007\\Neto\\Simulação_Sobrevivência_2008\\crescente\\cob_50.txt", sep =
';', dec = ".")
write.table(cobertura_150, "C:\\Documents and Settings\\User\\Meus
documentos\\Gui Ferraudo\\Estatística\\Mestrado
2007\\Neto\\Simulação_Sobrevivência_2008\\crescente\\cob_150.txt", sep =
';', dec = ".")
write.table(cobertura_500, "C:\\Documents and Settings\\User\\Meus
documentos\\Gui Ferraudo\\Estatística\\Mestrado
2007\\Neto\\Simulação_Sobrevivência_2008\\crescente\\cob_500.txt", sep =
';', dec = ".")
write.table(cobertura_1000, "C:\\Documents and Settings\\User\\Meus
documentos\\Gui Ferraudo\\Estatística\\Mestrado
2007\\Neto\\Simulação_Sobrevivência_2008\\crescente\\cob_1000.txt", sep =
';', dec = ".")

### Apresentando os resultados obtidos em forma de tabela ###

```

```

prob_cobertura=matrix(c(cobertura_25[1,2],cobertura_50[1,2],cobertura_150[1
,2],cobertura_500[1,2],cobertura_1000[1,2],

cobertura_25[2,2],cobertura_50[2,2],cobertura_150[2,2],cobertura_500[2,2],c
obertura_1000[2,2],

cobertura_25[3,2],cobertura_50[3,2],cobertura_150[3,2],cobertura_500[3,2],c
obertura_1000[3,2],

cobertura_25[4,2],cobertura_50[4,2],cobertura_150[4,2],cobertura_500[4,2],c
obertura_1000[4,2],

cobertura_25[5,2],cobertura_50[5,2],cobertura_150[5,2],cobertura_500[5,2],c
obertura_1000[5,2],

cobertura_25[6,2],cobertura_50[6,2],cobertura_150[6,2],cobertura_500[6,2],c
obertura_1000[6,2]), ncol=6,
dimnames = list( c( '25', '50', '150', '500', '1000'), c('Beta 0', 'Beta
1', 'Beta 2', 'Beta 3', 'Beta 4', 'Beta 5') ) )

prob_cobertura

write.table(prob_cobertura,      "C:\\Documents and Settings\\User\\Meus
documentos\\Gui
Ferraud\\Estatística\\Mestrado
2007\\Neto\\Simulação_Sobrevivência_2008\\crescente\\prob_cobertura.txt",
sep = ';', dec = ".")

## Fazendo o gráfico para melhor visualização da tabela armazenada no
objeto "prob_cobertura" ##

fator = as.factor(c(25, 50, 150, 500, 1000))

pdf("C:\\Documents and Settings\\User\\Meus documentos\\Gui
Ferraud\\Estatística\\Mestrado
2007\\Neto\\Simulação_Sobrevivência_2008\\crescente\\graf_beta0.pdf")
plot(fator,      prob_cobertura[,1],      type      =      'p',      main=
expression(bold(beta[0])), xlab="Tamanhos de Amostra (n)",ylab=expression(
paste("Probabilidade de Cobertura para ", beta[0]), sep="  " ), ylim =
c(0.8,1))
abline(h=0.9365) ## LI ##
abline(h=0.95) ## LC ##
abline(h=0.9635) ## LS ##
dev.off()

pdf("C:\\Documents and Settings\\User\\Meus documentos\\Gui
Ferraud\\Estatística\\Mestrado
2007\\Neto\\Simulação_Sobrevivência_2008\\crescente\\graf_beta1.pdf")
plot(fator,      prob_cobertura[,2],      type      =      'p',      main=
expression(bold(beta[1])), xlab="Tamanhos de Amostra (n)",ylab=expression(
paste("Probabilidade de Cobertura para ", beta[1]), sep="  " ), ylim =
c(0.8,1))
abline(h=0.9365) ## LI ##
abline(h=0.95) ## LC ##
abline(h=0.9635) ## LS ##
dev.off()

pdf("C:\\Documents and Settings\\User\\Meus documentos\\Gui
Ferraud\\Estatística\\Mestrado
2007\\Neto\\Simulação_Sobrevivência_2008\\crescente\\graf_beta2.pdf")
plot(fator,      prob_cobertura[,3],      type      =      'p',      main=
expression(bold(beta[2])), xlab="Tamanhos de Amostra (n)",ylab=expression(

```

```

paste("Probabilidade de Cobertura para ", beta[2]), sep=" " ), ylim =
c(0.8,1))
abline(h=0.9365) ## LI ##
abline(h=0.95) ## LC ##
abline(h=0.9635) ## LS ##
dev.off()

pdf("C:\\Documents and Settings\\User\\Meus documentos\\Gui
Ferraudo\\Estatística\\Mestrado
2007\\Neto\\Simulação_Sobrevivência_2008\\crescente\\graf_beta3.pdf")
plot(fator, prob_cobertura[,4], type = 'p', main=
expression(bold(beta[3])), xlab="Tamanhos de Amostra (n)",ylab=expression(
paste("Probabilidade de Cobertura para ", beta[3]), sep=" " ), ylim =
c(0.8,1))
abline(h=0.9365) ## LI ##
abline(h=0.95) ## LC ##
abline(h=0.9635) ## LS ##
dev.off()

pdf("C:\\Documents and Settings\\User\\Meus documentos\\Gui
Ferraudo\\Estatística\\Mestrado
2007\\Neto\\Simulação_Sobrevivência_2008\\crescente\\graf_beta4.pdf")
plot(fator, prob_cobertura[,5], type = 'p', main=
expression(bold(beta[4])), xlab="Tamanhos de Amostra (n)",ylab=expression(
paste("Probabilidade de Cobertura para ", beta[4]), sep=" " ), ylim =
c(0.8,1))
abline(h=0.9365) ## LI ##
abline(h=0.95) ## LC ##
abline(h=0.9635) ## LS ##
dev.off()

pdf("C:\\Documents and Settings\\User\\Meus documentos\\Gui
Ferraudo\\Estatística\\Mestrado
2007\\Neto\\Simulação_Sobrevivência_2008\\crescente\\graf_beta5.pdf")
plot(fator, prob_cobertura[,6], type = 'p', main=
expression(bold(beta[5])), xlab="Tamanhos de Amostra (n)",ylab=expression(
paste("Probabilidade de Cobertura para ", beta[5]), sep=" " ), ylim =
c(0.8,1))
abline(h=0.9365) ## LI ##
abline(h=0.95) ## LC ##
abline(h=0.9635) ## LS ##
dev.off()

##Fim
# Cálculo do vício das estimativas dos parâmetros de uma regressão #Weibull
considerando censura à esquerda

set.seed(354739)

vicio_beta = function(repli, n, beta0, beta1, beta2, beta3, beta4, beta5,
gama, censura){

require(survival)

x1 = matrix(NA,repli,n)
x2 = matrix(NA,repli,n)
x3 = matrix(NA,repli,n)
x4 = matrix(NA,repli,n)
x5 = matrix(NA,repli,n)
cens = matrix(NA,repli,n)
y = matrix(NA,repli,n)

```

```

yfit = list()

vicio_beta0 = numeric()
vicio_beta1 = numeric()
vicio_beta2 = numeric()
vicio_beta3 = numeric()
vicio_beta4 = numeric()
vicio_beta5 = numeric()

coef = matrix(NA, repli, 6)
med_coef_b0 = numeric()
med_coef_b1 = numeric()
med_coef_b2 = numeric()
med_coef_b3 = numeric()
med_coef_b4 = numeric()
med_coef_b5 = numeric()

for(i in 1:repli){

x1[i,] = rbinom(n,1,0.4)
x2[i,] = rbinom(n,1,0.7)
x3[i,] = rbinom(n,1,0.5)
x4[i,] = rbinom(n,1,0.3)
x5[i,] = rweibull(n, 8, 750)
cens[i,] = rbinom(n, 1, (1-censura))

y[i,] = rweibull(n, scale = exp(beta0 + beta1*x1[i,] + beta2*x2[i,] +
beta3*x3[i,] + beta4*x4[i,] + beta5*x5[i,]), shape = (1/gama))

yfit[[i]] = survreg(Surv(y[i,], cens[i,], type='left') ~ x1[i,] + x2[i,] +
x3[i,] + x4[i,] + x5[i,], scale = gama, dist='weibull')

coef[i,] = yfit[[i]]$coef }

med_coef_b0 = apply(coef, 2, mean)[1]
med_coef_b1 = apply(coef, 2, mean)[2]
med_coef_b2 = apply(coef, 2, mean)[3]
med_coef_b3 = apply(coef, 2, mean)[4]
med_coef_b4 = apply(coef, 2, mean)[5]
med_coef_b5 = apply(coef, 2, mean)[6]

vicio_beta0 = abs(med_coef_b0 - beta0)
vicio_beta1 = abs(med_coef_b1 - beta1)
vicio_beta2 = abs(med_coef_b2 - beta2)
vicio_beta3 = abs(med_coef_b3 - beta3)
vicio_beta4 = abs(med_coef_b4 - beta4)
vicio_beta5 = abs(med_coef_b5 - beta5)

matrix(c(vicio_beta0, vicio_beta1, vicio_beta2, vicio_beta3, vicio_beta4,
vicio_beta5),
nrow=6, ncol=1,
dimnames = list(c('beta0', 'beta1', 'beta2', 'beta3', 'beta4', 'beta5'),
c('Vicio') ) )

}

##### Amostra 1: n = 25 #####
vicio_25 = vicio_beta(repli = 1000, n = 25, beta0 = 2, beta1 = 1, beta2 =
3, beta3= 5, beta4 = 7, beta5 = 0.0008, gama = 2, censura = 0)

```

```

vicio_25

##### Amostra 2: n = 50 #####
vicio_50 = vicio_beta(repli = 1000, n = 50, beta0 = 2, beta1 = 1, beta2 =
3, beta3= 5, beta4 = 7, beta5 = 0.0008, gama = 2, censura = 0)
vicio_50

##### Amostra 3: n = 150 #####
vicio_150 = vicio_beta(repli = 1000, n = 150, beta0 = 2, beta1 = 1, beta2 =
3, beta3= 5, beta4 = 7, beta5 = 0.0008, gama = 2, censura = 0)
vicio_150

##### Amostra 4: n = 500 #####
vicio_500 = vicio_beta(repli = 1000, n = 500, beta0 = 2, beta1 = 1, beta2 =
3, beta3= 5, beta4 = 7, beta5 = 0.0008, gama = 2, censura = 0)
vicio_500

##### Amostra 5: n = 1000 #####
vicio_1000 = vicio_beta(repli = 1000, n = 1000, beta0 = 2, beta1 = 1, beta2
= 3, beta3= 5, beta4 = 7, beta5 = 0.0008, gama = 2, censura = 0)
vicio_1000

##Salvando a matriz de saída em um arquivo texto para poder ser aberto no
Excel##
write.table(vicio_25,      "C:\\Documents and Settings\\User\\Meus
documentos\\Gui          Ferraudo\\Estatística\\Mestrado
2007\\Neto\\Simulação_Sobrevivência_2008\\crescente\\vicio_25.txt", sep =
';', dec = ".")
write.table(vicio_50,      "C:\\Documents and Settings\\User\\Meus
documentos\\Gui          Ferraudo\\Estatística\\Mestrado
2007\\Neto\\Simulação_Sobrevivência_2008\\crescente\\vicio_50.txt", sep =
';', dec = ".")
write.table(vicio_150,      "C:\\Documents and Settings\\User\\Meus
documentos\\Gui          Ferraudo\\Estatística\\Mestrado
2007\\Neto\\Simulação_Sobrevivência_2008\\crescente\\vicio_150.txt", sep =
';', dec = ".")
write.table(vicio_500,      "C:\\Documents and Settings\\User\\Meus
documentos\\Gui          Ferraudo\\Estatística\\Mestrado
2007\\Neto\\Simulação_Sobrevivência_2008\\crescente\\vicio_500.txt", sep =
';', dec = ".")
write.table(vicio_1000,      "C:\\Documents and Settings\\User\\Meus
documentos\\Gui          Ferraudo\\Estatística\\Mestrado
2007\\Neto\\Simulação_Sobrevivência_2008\\crescente\\vicio_1000.txt", sep =
';', dec = ".")

### Apresentando os resultados obtidos em forma de tabela ###

vicio
matrix(c(vicio_25[1],vicio_50[1],vicio_150[1],vicio_500[1],vicio_1000[1],
vicio_25[2],vicio_50[2],vicio_150[2],vicio_500[2],vicio_1000[2],
vicio_25[3],vicio_50[3],vicio_150[3],vicio_500[3],vicio_1000[3],
vicio_25[4],vicio_50[4],vicio_150[4],vicio_500[4],vicio_1000[4],
vicio_25[5],vicio_50[5],vicio_150[5],vicio_500[5],vicio_1000[5],
vicio_25[6],vicio_50[6],vicio_150[6],vicio_500[6],vicio_1000[6]), ncol=6,
dimnames = list( c( '25', '50', '150', '500', '1000'), c('Beta 0', 'Beta
1', 'Beta 2', 'Beta 3', 'Beta 4', 'Beta 5') ) )

```



```

vicio

write.table(vicio, "C:\\Documents and Settings\\User\\Meus documentos\\Gui
Ferrauo\\Estatística\\Mestrado
2007\\Neto\\Simulação_Sobrevivência_2008\\crescente\\vicio.txt", sep = ';',
dec = ".")

## Fazendo o gráfico para melhor visualização da tabela armazenada no
objeto "prob_vicio" ##

fator = as.factor(c(25, 50, 150, 500, 1000))

pdf("C:\\Documents and Settings\\User\\Meus documentos\\Gui
Ferrauo\\Estatística\\Mestrado
2007\\Neto\\Simulação_Sobrevivência_2008\\crescente\\graf_beta0_vicio.pdf")
plot(fator, vicio[,1], type = 'p', main= expression(bold(beta[0])),
xlab="Tamanhos de Amostra (n)",ylab=expression( paste("Vício para ",
beta[0]), sep=" " ), ylim = c(-0.01, 3))
abline(h=0, col='red', lty=2)
dev.off()

pdf("C:\\Documents and Settings\\User\\Meus documentos\\Gui
Ferrauo\\Estatística\\Mestrado
2007\\Neto\\Simulação_Sobrevivência_2008\\crescente\\graf_beta1_vicio.pdf")
plot(fator, vicio[,2], type = 'p', main= expression(bold(beta[1])),
xlab="Tamanhos de Amostra (n)",ylab=expression( paste("Vício para ",
beta[1]), sep=" " ), ylim = c(-0.01, 3))
abline(h=0, col='red', lty=2)
dev.off()

pdf("C:\\Documents and Settings\\User\\Meus documentos\\Gui
Ferrauo\\Estatística\\Mestrado
2007\\Neto\\Simulação_Sobrevivência_2008\\crescente\\graf_beta2_vicio.pdf")
plot(fator, vicio[,3], type = 'p', main= expression(bold(beta[2])),
xlab="Tamanhos de Amostra (n)",ylab=expression( paste("Vício para ",
beta[2]), sep=" " ), ylim = c(-0.01, 3))
abline(h=0, col='red', lty=2)
dev.off()

pdf("C:\\Documents and Settings\\User\\Meus documentos\\Gui
Ferrauo\\Estatística\\Mestrado
2007\\Neto\\Simulação_Sobrevivência_2008\\crescente\\graf_beta3_vicio.pdf")
plot(fator, vicio[,4], type = 'p', main= expression(bold(beta[3])),
xlab="Tamanhos de Amostra (n)",ylab=expression( paste("Vício para ",
beta[3]), sep=" " ), ylim = c(-0.01, 3))
abline(h=0, col='red', lty=2)
dev.off()

pdf("C:\\Documents and Settings\\User\\Meus documentos\\Gui
Ferrauo\\Estatística\\Mestrado
2007\\Neto\\Simulação_Sobrevivência_2008\\crescente\\graf_beta4_vicio.pdf")
plot(fator, vicio[,5], type = 'p', main= expression(bold(beta[4])),
xlab="Tamanhos de Amostra (n)",ylab=expression( paste("Vício para ",
beta[4]), sep=" " ), ylim = c(-0.01, 3))
abline(h=0, col='red', lty=2)
dev.off()

pdf("C:\\Documents and Settings\\User\\Meus documentos\\Gui
Ferrauo\\Estatística\\Mestrado
2007\\Neto\\Simulação_Sobrevivência_2008\\crescente\\graf_beta5_vicio.pdf")

```

```
plot(fator, vicio[,6], type = 'p', main= expression(bold(beta[5])),
xlab="Tamanhos de Amostra (n)",ylab=expression( paste("Vício para ",
beta[5]), sep=" " ), ylim = c(-0.01, 3))
abline(h=0, col='red', lty=2)
dev.off()

#Fim
```

## APÊNDICE C – Código para gerar TTT-Plot no SAS

```

/* TTT PLOT */
data falha;          /* Separação das falhas do banco de dados */
set work.Gui;
where censura EQ 1;
run;

%let dados=falha;          /* nome do conjunto de dados */
%let tempo=tempo;        /* nome da variável tempo no conjunto de dados*/

proc iml;              /* criação do conjunto graft3plot */
sort &dados by &tempo;
use &dados var{&tempo};
read all into ti;
n=nrow(ti);
nn=repeat(n,n);
t=repeat(0,n);
do i=1 to n;
    t[i]=i;
end;
aux=0;
do i=1 to n;
    t[i]=ti[i]+aux;
    aux=t[i];
end;
r=repeat(0,n);
do i=1 to n;
    r[i]=i;
end;
f=t+(nn-r)#ti;
s=sum(ti);
f=f/s;
razao=r/n;
grafico=shape(0,n,4);
grafico[,1]=razao;
grafico[,2]=f;
grafico[1,3]=0;
grafico[1,4]=0;
grafico[2:n,3]=repeat(1,n-1);
grafico[2:n,4]=repeat(1,n-1);
create graft3plot FROM grafico[colname={razao f diagx diagy}];
append from grafico;
close graft3plot;
quit;

proc gplot data=graft3plot;          /* gráfico TTT - construção */
plot f *razao ;
label f="G(r/n)"
      razao="r/n";
symbol V=POINT W=1 LINE=1 I=JOIN;
plot2 razao*razao /overlay NOAXIS ;
symbol V=POINT W=1 LINE=1 I=JOIN;
run;
quit;

```