UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ
COORDENAÇÃO DO CURSO DE LICENCIATURA EM MATEMÁTICA

RAPHAEL FERNANDO DE MELO

MODELO DE REGRESSÃO LINEAR MÚLTIPLA: UMA APLICAÇÃO EM DADOS DE IMÓVEIS DA CIDADE DE TOLEDO-PR

TRABALHO DE CONCLUSÃO DE CURSO

TOLEDO
2019

UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ
COORDENAÇÃO DO CURSO DE LICENCIATURA EM MATEMÁTICA

RAPHAEL FERNANDO DE MELO

MODELO DE REGRESSÃO LINEAR MÚLTIPLA: UMA APLICAÇÃO EM DADOS DE IMÓVEIS DA CIDADE DE TOLEDO-PR

Trabalho de Conclusão de Curso apresentado ao Curso de Licenciatura em Matemática da Universidade Tecnológica Federal do Paraná, Câmpus Toledo, como requisito parcial à obtenção do título de Licenciado em Matemática.

Orientador(a): Dra. Suellen Ribeiro Pardo Garcia

TOLEDO
2019

UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ
COORDENAÇÃO DO CURSO DE LICENCIATURA EM MATEMÁTICA

TERMO DE APROVAÇÃO

O Trabalho de Conclusão de Curso intitulado "Modelo de Regressão Linear Múltipla: Uma Aplicação em dados de imóveis da Cidade de Toledo-PR"foi considerado **APROVADO** de acordo com a ata nº __ de --/--/----

Fizeram parte da banca examinadora os professores:

Dra. Suellen Ribeiro Pardo Garcia

Dr. Gustavo Henrique Dalposso

Dra. Rosângela Aparecida Botinha Assumpção

TOLEDO
2019

AGRADECIMENTOS

Primeiramente agradeço a Deus, por ter me dado todos os conhecimentos possíveis para progredir a cada dia e ser a luz em meu caminho.

À minha família, em especial a minha mãe Cremilda por ter sido uma guerreira todos esses anos, sem ela nada disso seria possível. À minha vó Maria pelo exemplo de mulher, e uma segunda mãe para mim. Obrigado por todo o apoio e paciência.

À minha família de coração Evanir, Stefanie e em memória de Josué. Obrigado por todos os ensinamentos e acolhimento.

Aos meus amigos Stefanie, Kevin, Felipe e Natalia que estiveram comigo nessa caminhada, me incentivando e estando ao meu lado sempre que precisei. Em especial à minhas colegas e amigas de turma Giovanna e Vitória, por sempre me apoiarem e me confortar nesses anos de curso. À todos vocês meu sincero obrigado.

À professora Suellen, minha querida orientadora, um agradecimento enorme por toda a compreensão, paciência e dedicação na realização deste trabalho, uma excelente pessoa e extraordinária profissional.

A banca examinadora, professor Gustavo e professora Rosângela, pela disposição, sugestões e contribuições acerca do trabalho.

Agradeço a todos que de certa forma contribuiram pela realização do presente trabalho, em especial aos professores e colegas do Curso.

Desejo a todos nós sucesso sempre!

RESUMO

A estatística é uma forte ferramenta de gestão de dados relacionados a situações cotidianas e tem contribuído de forma significativa para o processo de tomada de decisão nas mais diferentes áreas como engenharia, economia, medicina, entre outras. Nas ramificações da estatística o estudo da regressão linear é uma das mais amplas técnicas utilizadas que expressa a relação entre duas ou mais variáveis, de tal forma que uma variável possa ser predita a partir de outras. O presente trabalho apresenta um modelo de regressão linear múltipla para um conjunto de dados sobre preço e características de imóveis situados no município de Toledo no Oeste do Paraná. Os dados foram obtidos por meio de websites de diversas imobiliárias do município. Considerou-se como variáveis que influenciam no preço da casa: área construída, tamanho do lote, número de banheiros, número de vagas na garagem e o bairro em que se encontra o imóvel (variável *dummy*). A modelagem foi realizada por meio do *software R* e o método *stepwise* aplicado para selecionar o melhor conjunto de variáveis que descreva a variável resposta, ou seja, o preço da casa. Foi realizada a validação do modelo ótimo e o mesmo consegue explicar 83% da variabilidade do valor da casa. Assim, o presente trabalho é uma contribuição para o mercado imobiliário, pois o preço de uma casa, na cidade de Toledo-PR, pode ser estimado com base em suas características.

Palavras-chave: Modelagem Estatística. Avaliação de Imóveis. Stepwise. Software R

ABSTRACT

Statistics is a strong data management tool related to everyday situations and has contributed significantly to the decision-making process in the most different areas as engineering, economics, medicine, among others. In the ramifications of statistics, the study of Linear regression is one of the broadest techniques used that expresses the relationship between two or more variables, such that one variable can be predicted from others. The present work presents a multiple linear regression model for a dataset about price and characteristics of real estate located in the city of Toledo in Paraná State. The data were obtained through websites of several real estate agencies in the city. We consider as variables that influence the price of the house: built area, lot size, number of bathrooms, number of parking spaces and the neighborhood where the property is located (dummy variable). The modeling was performed using the software R and the stepwise method applied to select the best set of variables that describes the response variable, ie the price of the house. The validation of the optimal model was performed and it can explain 83% of the house value variability. Thus, the present work is a contribution to the real estate market, since the price of a house in the city of Toledo-PR can be estimated based on its characteristics.

Keywords: Statistical modeling. Property Appraisal. Stepwise. R software.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

O estudo da estatística é uma forte ferramenta de gestão de dados relacionados a situações cotidianas e tem contribuído de forma significativa para o processo de tomada de decisão, de maneira que, por meio de métodos matemáticos é possível relacionar dados com modelos algébricos, podendo assim interpretar e até mesmo prever determinadas situações.

Nas ramificações da estatística o estudo da regressão linear é uma das duas mais amplas técnicas estatísticas utilizadas no mundo que é caracterizada pela relação entre duas ou mais variáveis, a análise de variância é a outra Junior et al (2004).

A finalidade da análise de regressão é estudar a relação funcional entre duas ou mais variáveis, as quais assumem valores quantitativos ou qualitativos, de tal forma que elas possam ser preditas a partir de outras, ou seja, podem projetar ou estimar uma nova observação (MARTINS, 2005). Há diversos exemplos que se pode mostrar onde uma variável é predita por meio de outra ou de outras, onde existe esta relação funcional entre as variáveis.

A) Investimento em propaganda e retorno de vendas;

B) Tempo de entrega de produtos agrícolas e porcentagem de produtos estragados;

C) Número de pessoas internadas com sintomas da Dengue em relação ao índice pluviométrico da região estudada.

D) Influencia de alguns parâmetros físico-químicos que visam uma melhoria no processo de cultivo da produção de camarão;

E) Análise de desenvolvimento de produtos que conservam suas propriedades sensoriais e nutritivas.

O município de Toledo, localizado no oeste paranaense vem passando por um grande desenvolvimento na última década. Sendo considera a 7ª cidade com maior crescimento do Brasil (GAZETA DE TOLEDO, 2018), tal fato reflete diretamente no crescimento imobiliário do município, no qual estima-se que apresente um aumento entre 15% a 20% dos lançamentos de ímoveis até o final do ano de 2019, seguindo a tendência do país (JORNAL DO OESTE, 2019).

Diante de tais circunstâncias o município apresenta demasiadas imobiliárias, nas quais todas apresentam várias características sobre os imóveis a venda. Sendo assim, o valor de um imóvel sofre influência de algumas variáveis, como por exemplo: a área de construção do imóvel, a área do lote,a idade do imóvel, quantidade de quartos, banheiros, suítes, vagas de garagem, sendo essas quantitativas e também as variáveis qualitativas como, a localização do imóvel, estado de conservação, entre outras.

Uma imobiliária é um investimento financeiro, e a construção de um modelo que possa predizer o seu valor passa a ser necessária para a observação de sua volatilidade, para a estimativa dos retornos esperados e para sua avaliação. Assim, os compradores poderão medir o retorno de seus investimentos e os gerentes poderão estimar seus preços conforme parâmetros valorizados pelo mercado (ROZENBAUM; MACEDO-SOARES, 2007).

Neste contexto, apresenta-se um modelo de regressão linear múltipla para os dados dos valores das casa à venda na cidade de Toledo, afim de, poder realizar previsões para tais valores conforme algumas características do imóvel.

.

2 OBJETIVOS

OBJETIVO GERAL

Esse trabalho tem como objetivo apresentar um modelo de regressão linear múltipla para valor de imóveis à venda na cidade de Toledo-PR. Considerando como variáveis independentes, área do imóvel construído, área do lote, quantidade de banheiros, quantidade de quartos, quantidade de vagas na garagem e o bairro em que o imóvel está situado.

OBJETIVOS ESPECÍFICOS

Para atingir o objetivo geral do trabalho, será preciso:

- Estudar a teoria do modelo de regressão linear simples e múltipla;
- Encontrar e analisar bibliografias que trazem aplicações do modelo de regressão linear no contexto do mercado imobiliário;
- Elaborar e apresentar um texto com a teoria dos modelos de regressão e estatísticas utilizadas para avaliação e validação do modelo escolhido;
- Apresentar a construção do modelo para previsão do valor do imóvel (casa) na cidade de Toledo, Paraná.

3 JUSTIFICATIVA

A importância deste trabalho se dá pelo fato que de que, o estudo dos modelos de regressão podem ser utilizados para resolver problemas em diferentes contextos e áreas como na Física, Biologia, Química, Ciências Sociais, Ciências da Saúde, etc, ou seja, este estudo abre possibilidades para compreender diversas aplicações da matemática com o uso da modelagem sw dados reais.

Durante a graduação, na disciplina de Probabilidade e Estatística, é estudado brevemente sobre a análise de regressão e o enfoque é na regressão linear simples, ou seja, não é trabalhado com a regressão linear múltipla e nota-se que as aplicações desta teoria encontradas na literatura, são em geral de regressão linear múltipla.

Segundo Gazola (2002), a possibilidade de contribuir com uma metodologia científica acessível de Regressão Linear Múltipla e Inferência Estatística, no que se refere a dar subsídios para a melhoria da qualidade das avaliações dos imóveis, é a principal justificativa para o desenvolvimento de trabalhos como este. Outra importância, segundo o autor, é a utilidade destes modelos para a avaliação de imóveis na realização de laudos, relativos a programas habitacionais, a patrimônios da União, seguros, entre outros. O autor ainda destaca que a regressão linear múltipla é de fácil interpretação e, principalmente, de simples aplicabilidade .

Pereira et al (2012) justificam sua proposta de modelo de regressão linear múltipla que auxilie na estimação dos preços de venda de imóveis da cidade, a partir de suas características físicas e de sua localização, diante de um intenso aquecimento do mercado imobiliário e pela dificuldade de predição e avaliação dos preços de venda de imóveis.

No âmbito público cita-se o uso da avaliação de imóveis para fins de compra e privatização e no cálculo de valores para lançamentos de impostos. Já no âmbito judicial é utilizada nas discussões entre pessoas físicas ou jurídicas que envolvam valores de imóveis, frequentes em ações demarcatórias, possessórias e indenizatórias e, também, nas discussões acerca de indenizações por desapropriações ou servidões de passagem . Os métodos estatísticos são os procedimentos de modelagem matemática mais utilizados para a avaliação imobiliária, destacando o modelo de regressão linear múltipla como o preferido dos avaliadores, pela eficiência (Baptistella et al, 2006).

Diante disso, o presente trabalho apresenta um modelo de regressão linear múltipla para um conjunto de dados sobre preço e características de imóveis situados no município de Toledo no oste do Paraná.

4 REFERENCIAL TEÓRICO

4.1 REGRESSÃO LINEAR SIMPLES

A regressão linear simples é um modelo matemático para duas variáveis, a variável resposta ou variável dependente. Y é a variável preditora ou variável independente, X. Apenas a variável dependente é tratada como uma variável aleatória (DOANNE & SEWARD, 2014). Nesta situação, se existir uma relação linear entre ambas as variáveis, teria o comportamento da variável dependente em função do comportamento da variável independente.

4.1.1 Terminologia da regressão

Esta seção baseia-se fortemente no livro Estatística Aplicada à Administração e Economia de Doanne e Seward (2014).

Modelos e parâmetros

Os parâmetros da população desconhecidos do modelo de regressão são denotados por β_0 e β_1 . O modelo da população para uma relação linear é:

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i, i = 1, 2, ...n \tag{4.1}$$

De modo que:

- Y_i corresponde ao valor da variável dependente, Y, na observação i, com $i = 1, ...n$
- x_i corresponde ao valor da variável independente, X, na observação i, com $i = 1, ...n$
- ϵ_i,$i = 1, ..., n$ correspondem aos erros aleatórios, de forma que é possivel explicar a variabilidade existente em Y e que não é explicada por X;
- β_0 e β_1 correspondem aos parâmetros do modelo.
- O β_0 representa o ponto em que a reta regressora intersecta o eixo Y, de modo que isso ocorre quando $X = 0$ e é chamado de intercepto Y ou coeficiente linear.
- O parâmetro β_1 representa a inclinação da reta regressora, expressando a taxa de mudança em Y, ou seja, indica a mudança na média da distribuição de probabilidade de Y para um aumento de uma unidade na variável X.

A partir da amostra, estima-se a equação de regressão e a utiliza para prever o valor esperado de Y para um determinado valor de X:

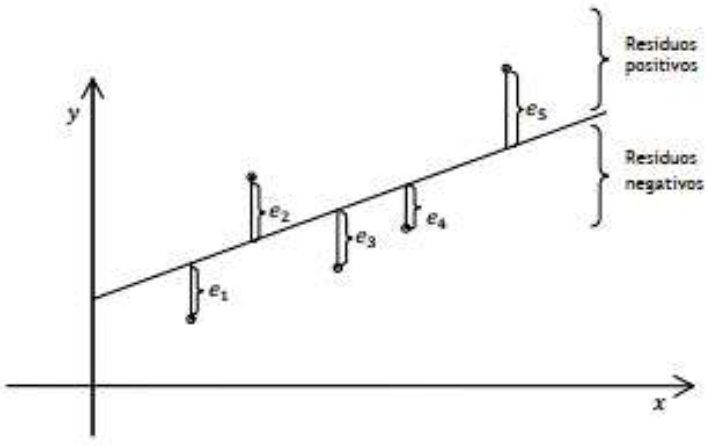
$$\hat{y} = b_0 + b_1 x \tag{4.2}$$

A equação (4.2) é chamada de equação da regressão estimada.

4.1.2 Método dos mínimos quadrados

O método dos mínimos quadrados (MMQ) é usado para estimar os parâmetros na equação de regressão. Na reta os pontos que estão acima dão erros positivos e os que estão abaixo dão erros negativos. A figura (4.1) mostra como é dado a representação gráfica dos resíduos e disso vêm o objetivo do MMQ, que é minimizar a soma dos quadrados dos desvios das observações em relação à linha real de regressão (MONTGOMERY & RUNGER, 2003).

Figura 4.1: Representação Gráfica dos Resíduos



Fonte: RODRIGUES (2012).

Inclinação e intercepto

Encontrar o melhor ajuste, significa que o coeficiente angular e o intercepto são de tal forma que os resíduos sejam os menores possíveis(DOANE & SEWARD,2014), sendo os resíduos a diferença entre o y observado e o y estimado ($\hat{y}$).

Os coeficientes ajustados b_0 e b_1 são calculados de maneira que o modelo linear ajustado $\hat{y} = b_0 + b_1 x$ tenha a menor soma de resíduos ao quadrado possível ($SQerro$):

$$SQerro = \sum_{i=1}^{n} {e_1}^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2 \tag{4.3}$$

Estimador da MQO da inclinação:

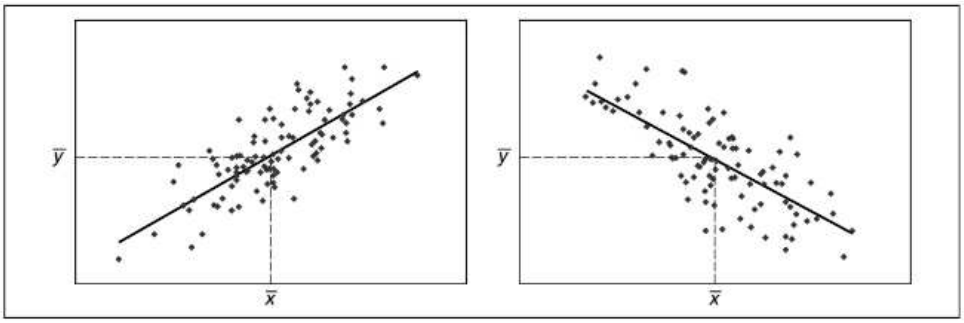
$$b_1 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \tag{4.4}$$

Estimador do Intercepto:

$$b_0 = \overline{y} - b_1\overline{x} \tag{4.5}$$

As fórmulas dos MQO produzem estimativas não viciadas e consistentes [1] de β_0 e β_1. A linha de regressão MQO passa sempre pelo ponto $(\overline{x}, \overline{y})$ para quaisquer dados, como ilustrado na figura (4.2).

Figura 4.2: Linha de Regressão



Fonte: DOANE & SEWARD (2014).

Fontes de variação em Y

Numa regressão, busca-se explicar a variação na variável dependente em torno de sua média. Expressa-se a variação total como uma soma de quadrados (denotada por SQTot):

$$SQTot = \sum_{i=1}^{n}(y_i - \overline{y}) \text{ (soma total de quadrados)}$$

Pode-se dividir a variação total em duas partes:

$$SQTot = SQReg + SQErro \tag{4.6}$$

[1] Para mais informações sobre estimativas não viciadas e consistentes leia capítulo 8 do Livro: Estatística Aplicada à Administração e Economia (Doanne e Seward, 2014).

Sendo, $SQReg$ a variação explicada pela regressão que é a soma de quadrados das diferenças entre as médias condicionais $\hat{y}_i$ (condicionada a um dado valor x_i) e a média incondicional $\overline{y}$ (a mesma para todos os x_i):

$$SQReg = \sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2$$

$SQErro$ a variação inexplicada em Y é a soma de quadrados dos resíduos, algumas vezes chamada soma de quadrados do erro denotada por:

$$SQErro = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Avaliação do ajuste: coeficiente de determinação

A magnitude da $SQErro$ é dependente do tamanho da amostra e das unidades mensuradas, desse modo, precisa-se uma referência que seja adimensional para o ajuste da equação da regressão. Pode-se obter uma medida de ajuste relativa comparando $SQTot$ com o $SQReg$. Dividindo a equação (4.6) por $SQTot$ tem-se:

$$\frac{SQTot}{SQTot} = \frac{SQReg}{SQTot} + \frac{SQErro}{SQTot} \text{ ou } 1 = \frac{SQReg}{SQTot} + \frac{SQErro}{SQTot}$$

A proporçãoo $\frac{SQReg}{SQTot}$ tem um nome especial: coeficiente de determinação ou R^2. Ela pode ser calculada de duas maneiras:

$$R^2 = 1 - \frac{SQErro}{SQTot} \tag{4.7}$$

$$R^2 = \frac{SQReg}{SQTot} \tag{4.8}$$

O coeficiente de determinação varia entre $0 \leq R^2 \leq 1$. Uma regressão de ajuste perfeito teria $R^2 = 1$ o que implica em $SQErro = 0$.

4.1.3 Teste de significância

Erro padrão da regressão

Uma medida de ajuste total é o erro padrão da regressão, denotado por S_e:

$$S_e = \sqrt{\frac{SQErro}{n-2}} \tag{4.9}$$

Na equação (4.9) se as previsões do modelo ajustado fossem perfeitas ($SQErro = 0$), o erro padrão S_e seria 0, isso implica que um valor menor de S_e indica um ajuste melhor.

O erro padrão S_e é um estimador de σ (o desvio padrão dos erros não observáveis). Por medir o ajuste total, S_e tem uma função semelhante com a do coeficiente de determinação. Porém,a magnitude de S_e se diferencia do R^2 pela unidade de mensuração da variável dependente e da ordem de magnitude dos dados. Por essa razão, o R^2 é frequentemente a medida preferida de ajuste total porque sua escala está sempre entre 0 e 1. A finalidade principal do erro padrão S_e é construir intervalos de confiança.

Intervalos de confiança para o coeficiente angular e o intercepto

Uma vez que se tem o erro padrão S_e, pode-se construir intervalos de confiança (IC) para os coeficientes. Seguem as fórmulas para o erro padrão do coeficiente angular e o erro padrão do intercepto, respectivamente:

$$S_{b_1} = \frac{S_e}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}} \tag{4.10}$$

$$S_{b_0} = S_e \sqrt{\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}} \tag{4.11}$$

Esses erros padrões são utilizados para construir intervalos de confiança para os verdadeiros valores do coeficiente angular e do intercepto, usando a distribuição t de Student com g.l. = n – 2 graus de liberdade e nível de significância. Os intervalos de confiança são calculados e dados por:

$$b_1 - t_{\frac{\alpha}{2}} S_{b_1} \leq \beta_1 \leq b_1 + t_{\frac{\alpha}{2}} S_{b_1} \tag{4.12}$$

$$b_0 - t_{\frac{\alpha}{2}} S_{b_0} \leq \beta_0 \leq b_0 + t_{\frac{\alpha}{2}} S_{b_0} \tag{4.13}$$

Na equação (4.12), tem-se o intervalo de confiança para o verdadeiro coeficiente angular e na equação (4.13), tem-se o intervalo de confiança para o verdadeiro intercepto.

Inferências sobre o coeficiente β

Após o ajustamento da reta e o cálculo de S_e, pode-se avaliar a qualidade do modelo pela realização de inferências estatísticas sobre seus parâmetros. Realiza-se os testes de hipótese isto é: verifica-se a existência, ou não, de regressão linear entre as variáveis X e Y. Tendo como hipótese nula, $H_0 : \beta_1 = 0$ e a hipótese alternativa $H_1 : \beta \neq 0$. Se o teste indicar a rejeição de H_0, pode-se concluir, com o erro estipulado, que há regressão de X sobre Y.

Procedimento para a realização do teste:

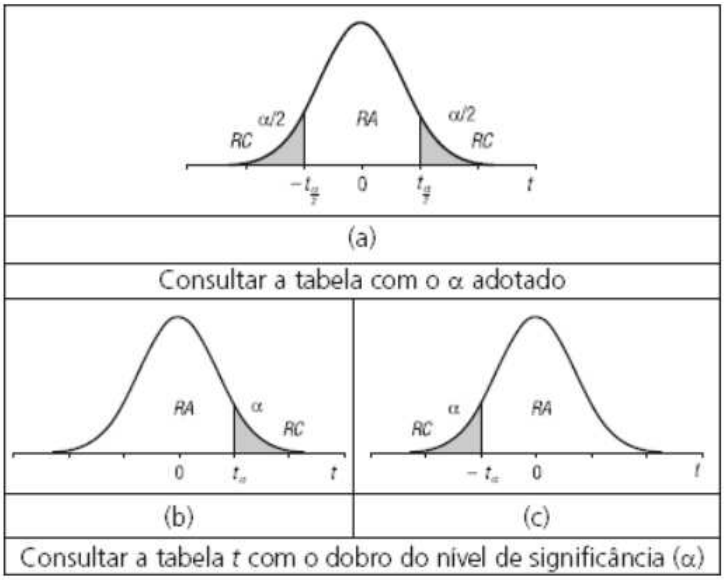
1. $H_0 : \beta = 0 \rightarrow$ Não existe regressão da variável Y sobre a variável X.

   $H_1 : \beta \neq 0$ $(a) \rightarrow$ Existe regressão da variável Y sobre a variável X.

   $H_1 : \beta > 0$ (b) $\rightarrow$ Existe regressão da variável Y sobre a variável X.

   $H_1 : \beta < 0$ (c) $\rightarrow$ Existe regressão da variável Y sobre a variável X.

Figura 4.3: Regiões Críticas Sobre a Distribuição t students



Fonte: MARTINS & DOMINGUES (2017).

2. Fixar α (probabilidade do erro) e escolher a variável do teste, no caso, a distribuição t de Student com $\phi = n - 2$.

3. Com auxílio da *tabela t*, construir as regiões de rejeição e aceitação para H_0.

4. Com os dados amostrais, calcular o valor da variável:

Coeficientes	Hipóteses	Estatística de Teste
Inclinação	$\frac{H_0:\beta_1=0}{H_1:\beta_1\neq 0}$	$t_{calc} = \frac{\text{inclinação estimada - inclinação hipotética}}{\text{erro padrão da inclinação}} = \frac{b_1-0}{S_{b_1}}$
Intercepto	$\frac{H_0:\beta_0=0}{H_1:\beta_0\neq 0}$	$t_{calc} = \frac{\text{intercepto estimado - intercepto hipotética}}{\text{erro padrão da inclinação}} = \frac{b_0-0}{S_{0_1}}$

5. Conclusão para teste:
   Caso(a): se $t_{cal} > t_{\frac{\alpha}{2}}$, ou $t_{cal} < -t_{\frac{\alpha}{2}}$, rejeita-se H_0, com risco α, ou seja,existe uma regressão linear entre as variáveis.

   Caso(b): se $-t_{\frac{\alpha}{2}} \leq t_{cal} \leq t_{-\frac{\alpha}{2}}$, rejeita-se H_0, com risco α, ou seja, não existe uma regressão linear entre as variáveis.

4.1.4 Estatística F para o ajuste geral

Para testar se uma regressão é significante, compara-se as somas dos quadrados explicadas ($SQReg$) e inexplicados ($SQErro$) utilizando um teste F. Divide-se cada soma pelos seus respectivos graus de liberdade para obter os quadrados médios ($QMReg$ e $QMErro$). A estatística F é a razão desses dois quadrados médios. Os cálculos da estatística F são apresentados em uma tabela chamada análise de variância ou tabela ANOVA (4.1)

Tabela 4.1: Tabela ANOVA

Fonte de Variação	**Soma de Quadrados**	**g.l**	**Quadrado médio**	F
Regressão (explicada)	SQReg=$\sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2$	1	$QMReg = \frac{SQReg}{1}$	$F_{cal} = \frac{QMReg}{QMErro}$
Resíduo (Inexplicado)	SQErro=$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$	$n-2$	$QMErro = \frac{SQErro}{n-2}$	
Total	SQTot=$\sum_{i=1}^{n}(y_i - \overline{y})^2$	$n-1$		

FOnte: DOANNE & SEWARD (2014).

Conclusão: Obtem-se o F_{tab} por $F(\alpha, glreg, glres)$ e se $F_{cal} > F_{tab}$ rejeita-se H_0, concluindo, com o risco α, que existe regressão linear simples, isto é, o modelo pode explicar e prever a variável Y.

4.1.5 Intervalos de confiança e de predição para Y

Como construir uma estimativa intervalar para Y

A reta de regressão é um estimador da média condicional de Y, mas as estimativas podem ser discrepantes. Para uma estimativa pontual, precisa-se de uma estimativa intervalar que mostre um intervalo de possíveis valores. Para isso, inseri-se o valor de x_i na equação de regressão ajustada, calcula-se a estimativa $\hat{y}_i$ e usa-se as fórmulas a seguir:

$$\hat{y}_i \pm t_{\frac{\alpha}{2}} S_e = \sqrt{\frac{1}{n} + \frac{(x_i - \overline{x}^2)}{\sum_{i=1}^{n}(x_i - \overline{x})^2}} \tag{4.14}$$

$$\hat{y}_i \pm t_{\frac{\alpha}{2}} S_e = \sqrt{1 + \frac{1}{n} + \frac{(x_i - \overline{x}^2)}{\sum_{i=1}^{n}(x_i - \overline{x})^2}} \tag{4.15}$$

A fórmula (4.14) produz um intervalo de confiança para a média condicional de Y, enquanto a fórmula (4.15) é um intervalo de predição para valores individuais de Y. As fórmulas são similares, exceto pelo fato de os intervalos de predição serem mais largos porque valores individuais de Y variam mais do que a média de Y.

4.1.6 Análise de resíduos

Hair et.al (2009), diz que a principal medida usada na avaliação da variável estatística de regressão é o resíduo (a diferença entre o valor real da variável dependente e o seu valor previsto). Sendo assim, é feita uma análise sobre os resíduos seguindo as suposições de regressão, sendo elas:

- Suposição 1: Variância constante.

  É avaliada por meio de análise de resíduos e gráficos de regressão parcial.

- Suposição 2: Independência dos resíduos.

  Tal suposição lida com o efeito de envolvimento de uma observação com a outra, tornando assim o resíduo não-independente.

- Suposição 3: Normalidade.

  A suposição final é a normalidade do termo do erro da variável estatística com o auxilio de gráficos de probabilidade normal dos resíduos.

4.1.7 Observações incomuns

Resíduos incomuns

Como toda regressão pode ter unidade diferente em Y, diante disso, é útil padronizar os resíduos, dividindo cada resíduo, e_i, pelo seu erro padrão individual S_{ei^*}.

$$e_i{}^* = \frac{e_i}{S_{e_i}} \tag{4.16}$$

A equação (4.16) diz respeito ao resíduo padronizado na observação i, em que:

$$S_{e_i} = S_e\sqrt{1-h_i}$$
$$h_i = \frac{1}{n} + \frac{(x_i-\overline{x})^2}{\sum(x_i-\overline{x})^2}$$

e,

$$h_i = \frac{1}{n} + \frac{(x_i-\overline{x})^2}{\sum(x_i-\overline{x})^2}$$

Observa-se que esse cálculo exige um único ajuste para cada resíduo, baseado na distância da observação com relação à média.

Alavancagem

Uma estatística de alavancagem indica que a observação está muito afastada da média de X. A alavancagem da observação i é denotada por h_i e é calculada por:

$$h_i = \frac{1}{n} + \frac{(x_i - \overline{x})^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \tag{4.17}$$

Uma regra simples para tal caso, uma estatística de alavancagem que exceda $\frac{4}{n}$ é um valor incomum.

Outliers

Barnett & Lewis (1995) definiram *outlier* em um conjunto de dados como sendo uma observação que parece ser discrepante com o conjunto de dados em análise. Uma observação é chamada de *outlier* quando ela apresenta um desvio da tendência linear apresentada pelas demais observações na direção Y.

4.2 REGRESSÃO LINEAR MÚLTIPLA

Existem situações em que apenas duas variáveis não conseguem explicar significativamente o relacionamento entre ela e a variável resposta. Nesse âmbito surgiu o estudo da regressão linear múltipla que é semelhante à regressão linear simples. Porém é definida por Tabachnick e Fidell (1996) como um conjunto de técnicas estatísticas que possibilitam a avaliação do relacionamento de uma variável dependente com várias variáveis independentes.

Diante de tal situação, existem algumas limitações nas quais abrangem a regressão linear simples, sendo elas:

- Estimativas viesadas, se preditores relavantes ao modelo são excluídos.
- Falta de ajuste não indica que X não está relacionada a Y se o modelo verdadeiro for múltiplo.

A finalidade das variáveis independentes adicionais é melhorar a capacidade de predição em confronto com a regressão linear simples. Isto é, como múltiplos preditores são geralmente relevantes, a regressão linear simples acaba servindo apenas para situações nas quais se requer um modelo mais simples contrariando o principio da regressão linear múltipla.

Terminologia da regressão

O modelo de regressão linear múltipla com k variáveis explicativas é por meio de uma equação linear chamada de modelo de regressão populacional, como a seguir:

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_k x_{ki} + \epsilon_i, i = 1, 2, ...n \tag{4.18}$$

onde,

- Y_i representa o valor da variável resposta Y na observação i, $i = 1, ..., n$;
- x_{1i}, x_{2i},..., x_{ki}, $i = 1, ..., n$ da iésima observação das variáveis $(X_1, X_2, \cdots, X_K$ explicativas, os preditores;
- $\beta_0, \beta_1, \beta_2, ..., \beta_k$ são os parâmetros ou coeficientes de regressão;
- ϵ_i, $i = 1, ..., n$ correspondem aos erros aleatórios.

Formatos dos dados

Os dados trabalhados na regressão linear múltipla podem ser representados da seguinte maneira:

Tabela 4.2: Dados Utilizados para a Regressão Linear Múltipla

Y	X_1	X_2	...	X_k
Y_1	x_{11}	x_{21}	...	x_{1k}
Y_2	x_{21}	x_{22}	...	x_{2k}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
Y_n	x_{n1}	x_{n2}	...	x_{nk}

Fonte: Dos Autores (2019).

A Tabela (4.2) representa os dados coletados de uma experiência qualquer, de modo que, os valores de k são independentes e o tamanho da amostra é n.

O modelo apresentado na equação (4.18) é um sistema de n equações e pode ser representado matricialmente por:

$$Y = X\beta + \epsilon \tag{4.19}$$

$$\text{onde, } \mathbf{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \text{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix}, \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} e \epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

1. De forma que $\mathbf{Y}$: é o vetor coluna $(n \times 1)$ constituído pelas observações da variável resposta.

2. Matriz $\mathbf{X}$ $(n \times p)$, onde $p = k+1$ no qual as linhas são constituídas pelos valores das variáveis independentes. Na primeira coluna todos os valores são iguais a 1, pois é o coeficiente de β_0.

3. Matriz β é um vetor coluna $(p \times 1)$ dos coeficientes de regressão.

4. Matriz ϵ, é um vetor coluna $(n \times 1)$ dos erros aleatórios.

Para encontrar o vetor de estimadores dos mínimos quadrados $\hat{\beta} = \beta$ que minimize a soma de quadrados do erro, tem-se $\epsilon = Y - X\beta$ e, assim:

$$SQErro = L = \sum_{i=1}^{n} {\epsilon_i}^2 = \epsilon^T \epsilon = (Y - X\beta)^T (Y - X\beta) = Y^T Y - 2\beta^T X^T Y + \beta^T X^T X\beta \tag{4.20}$$

sendo $\beta^T X^T Y$ do tipo (1×1), a sua trasposta $Y^T X\beta$ tem o mesmo valor. O estimador dos mínimos quadrados $\hat{\beta}$ será a solução das sequintes equações:

$$\frac{\partial L}{\partial \hat{\beta}} = 0 \Leftrightarrow -2X^T Y + 2X^T X\hat{\beta} = 0 \Leftrightarrow X^T X\hat{\beta} = X^T Y \tag{4.21}$$

Em 4.21 multiplicando ambos os membros, à esquerda por $(X^T X)^{-1}$ obtem-se o estimador:

$$\hat{\beta} = B = (X^T X)^{-1}\ X^T Y \tag{4.22}$$

No qual, em 4.22 as matrizes $X^T X$ e $X^T Y$ são:

$$X^T X = \begin{bmatrix} n & \sum_{i=1}^{n} x_{i1} & \sum_{i=1}^{n} x_{i2} & \cdots & \sum_{i=1}^{n} x_{ik} \\ \sum_{i=1}^{n} x_{i1} & \sum_{i=1}^{n} x_{i1}^2 & \sum_{i=1}^{n} x_{i1}x_{i2} & \cdots & \sum_{i=1}^{n} x_{i1}x_{ik} \\ \sum_{i=1}^{n} x_{i2} & \sum_{i=1}^{n} x_{i2}x_{i1} & \sum_{i=1}^{n} x_{i2}^2 & \cdots & \sum_{i=1}^{n} x_{i2}x_{ik} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^{n} x_{ik} & \sum_{i=1}^{n} x_{ik}x_{i1} & \sum_{i=1}^{n} x_{ik}x_{i2} & \cdots & \sum_{i=1}^{n} x_{ik}^2 \end{bmatrix}$$

$$X^TY = \begin{bmatrix} 1 & 1 & 1 & \cdots & 1 \\ x_{11} & x_{21} & x_{31} & \cdots & x_{n1} \\ x_{12} & x_{22} & x_{32} & \cdots & x_{n2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{1k} & x_{2k} & x_{3k} & \cdots & x_{nk} \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_{i1}y_i \\ \sum_{i=1}^{n} x_{i2}y_i \\ \vdots \\ \sum_{i=1}^{n} x_{ik}y_i \end{bmatrix}$$

então,

$$\hat{\beta} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_k \end{bmatrix} = \begin{bmatrix} n & \sum_{i=1}^{n} x_{i1} & \sum_{i=1}^{n} x_{i2} & \cdots & \sum_{i=1}^{n} x_{ik} \\ \sum_{i=1}^{n} x_{i1} & \sum_{i=1}^{n} x_{i1}^2 & \sum_{i=1}^{n} x_{i1}x_{i2} & \cdots & \sum_{i=1}^{n} x_{i1}x_{ik} \\ \sum_{i=1}^{n} x_{i2} & \sum_{i=1}^{n} x_{i2}x_{i1} & \sum_{i=1}^{n} x_{i2}^2 & \cdots & \sum_{i=1}^{n} x_{i2}x_{ik} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^{n} x_{ik} & \sum_{i=1}^{n} x_{ik}x_{i1} & \sum_{i=1}^{n} x_{ik}x_{i2} & \cdots & \sum_{i=1}^{n} x_{ik}^2 \end{bmatrix}^{-1} \begin{bmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_{i1}y_i \\ \sum_{i=1}^{n} x_{i2}y_i \\ \vdots \\ \sum_{i=1}^{n} x_{ik}y_i \end{bmatrix}$$

A matriz X^TX é uma matriz simétrica $(p \times p)$ e X^TY é um vetor coluna $(p \times 1)$, ou seja, a matriz $\hat{\beta}$ é um vetor coluna $(p \times 1)$.

Coeficientes padronizados

Deve-se ter atenção nas unidades métricas das variáveis independentes, sendo assim, não é possível interpretar os valores dos seus parâmetros como uma medida de contribuição de cada regressor para a explicação da variável resposta. Pode-se padronizar a equação de regressão fazendo a seguinte transformação:

$$\beta_j^{'} = \beta_j \frac{S_{x_j}}{S_y} (4.23)$$

sendo,

- $\beta_j^{'}$: Com $j = 1, 2 \cdots k$ são os coeficiente de regressão padronizado e estão relacionados com os coeficientes de regressão convencional.
- β_j: Com $j = 1, 2 \cdots k$ são os coeficientes de regressão convencional.
- S_{x_j}: Desvio padrão amostral das variáveis x_j, com $j = 1, 2 \cdots k$.
- S_y: Desvio padrão da variável resposta Y.

Os coeficientes padronizados, expressam a taxa de variação em unidades de desvio padrão para y por cada variação de uma unidade de desvio padrão para x. A principal vantagem da utilização dos coeficientes padronizados se dá pelo fato de que seus valores podem ser diretamente comparados, sendo assim, é possível classificar qual variável independente que mais contribui para a explicação da variação da variável dependente.

4.2.1 Avaliando o Ajuste Geral do Modelo

Para validar os resultados obtidos pelo Método dos Mínimos Quadrados (MMQ), bem como a significância das inferências sobre o modelo de regressão linear múltipla, são necessárias as mesmas hipóteses que foram feitas para o modelo de regressão linear simples (MARTINS, 2005).

Por trabalhar com mais variáveis que a regressão linear simples, a tabela ANOVA para a regressão linear múltipla apresenta algumas modificações, de acordo com a 4.3.

Tabela 4.3: Tabela ANOVA para Regressão Linear Múltipla

Fonte de Variação	**Soma de Quadrados**	**g.l**	**Quadrado médio**	F
Regressão (explicada)	SQReg=$\sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2$	k	QMReg=$\frac{SQReg}{k}$	$F_{cal} = \frac{QMReg}{QMErro}$
Resíduo (Inexplicado)	SQErro=$\sum_{i=1}^{n}(y_i - \hat{y})^2$	$n-k-1$	QMErro=$\frac{SQErro}{n-k-1}$	
Total	SQTot=$\sum_{i=1}^{n}(y_i - \overline{y})^2$	$n-1$		

FOnte:DOANNE & SEWARD (2014).

4.2.2 Coeficiente de determinação (R^2)

Coeficiente de determinação ajustado

O coeficiente de determinação ajustado tem por finalidade caracterizar a redução da variabilidade total de Y com o conjunto de variáveis X_i , onde $1 \leq i \leq n$ e $1 \leq j \leq k$. É tentador incluir muitos preditores para obter um "melhor ajuste", o que é denominado de sobreajuste. Para não ocorrer isso, um ajustamento é feito na estatística R^2 para penalizar a inclusão de preditores inúteis. O coeficiente de determinação ajustado considerando n observações e k preditores é:

$$R^2{}_{\text{ajustado}} = 1 - \frac{(\frac{SQErro}{n-k-1})}{(\frac{SQTot}{n-1})} \tag{4.24}$$

O $R^2{}_{\text{ajustado}}$ é sempre menor que R^2. À medida que se inclui preditores, R^2 não diminuirá. Mas $R^2{}_{\text{ajustado}}$ pode aumentar, permanecer estável ou diminuir, dependendo se os preditores incluídos aumentarem R^2 suficientemente para compensar a penalidade. Se $R^2{}_{\text{ajustado}}$ é substancialmente menor que R^2, sugere-se que o modelo contém preditores inúteis (DOANE & SEWARD, 2014).

Existência de Regressão Linear Múltipla

Para o caso geral, sinteticamente, o teste seria:

1. H_0: $\beta_1 = \beta_2 = \cdots, \beta_k = 0$.
   H_1: $\beta_j \neq 0$ (para no mínimo um j)

2. Fixar α (probabilidade do erro) e escolher uma variável $F(K; glreg, glres)$.

3. Com o auxílio da tabela de distribuição F, determinar a região de aceitação e a região critica.

4. Calcular a estatística do teste:

$$F_{cal} = \left(\frac{R^2}{1-R^2}\right)\left[\frac{n-(K-1)}{k}\right] = \frac{QMreg}{QMres}$$

5. Caso $F_{cal} > F_{tab}$ rejeita-se H_0, concluindo-se, com risco α, que existe regressão, isto é, o modelo é capaz de explixar e prever Y.

4.2.3 Estimação Stepwise

Por meio de métodos computacionais é possível otimizar o ajuste do melhor modelo usando $1, 2 \cdots, k$ preditores. Na ausência de um modelo teórico, no método em questão cada variável é considerada para inclusão antes do desenvolvimento da equação. De acordo com Hair et. al, 2009 o procedimento *stepwise* pode ser realizado por meio dos seguintes estágios:

1. Começar com o modelo de regressão simples selecionando a variável independente que é a mais fortemente correlacionada com a variável dependente

2. Examinar os coeficientes de correlação parcial para encontrar uma variável independente adicional que explique a maior parte estatisticamente significante do erro remanescente.

3. Recalcular a equação de regressão usando duas variáveis independentes e examinar o valor parcial de F para ver se ainda há uma contributo significante. Caso contrario, é feita a eliminação de tal variável.

4. Continuar o procedimento examinando todas as variáveis independentes não presentes no modelo para determinar se alguma faria uma adição significativa para equação. Se uma nova for adicionada, examinar todas as variáveis independentes de modo a julgar se elas devem ser mantidas.

5. Adicionar variáveis independentes ate que nenhuma das candidatas contribua numa melhora estatisticamente significante.

4.2.4 Preditores Binários

Não é possível inserir uma variável qualitativa em um modelo como preditor porque a regressão requer dados numéricos, então é preciso fazer uma codificação numérica, de forma a transformar dados categóricos em preditores úteis.

Considera-se aqui o caso de dados categóricos que têm apenas dois níveis. Estes são relacionados a uma variável binária, ou seja, relaciona-se os dados a dois valores, denotando a presença ou a ausência de uma condição (normalmente codificada como 0 e 1). Ao se codificar uma categoria cria-se então um preditor binário.

Preditores binários são de fácil criação e de grande utilidade, pois permite a utilização de variáveis qualitativas(categóricas) dentro de um modelo numérico. Tais variáveis também são chamadas de *variável dummy*, *dicotômicas* ou *variáveis indicadoras* (DOANE & SEWARD, 2014).

O teste t é utilizado para testar se o coeficiente da variável preditora binária é igual a zero, da mesma forma que o coeficiente de uma variével preditora quantitativa é testado quanto à sua significância. Se o coeficiente do preditor binário for considerado significativamente diferente de zero, então conclui-se que o preditor binário é um preditor significante de Y. O coeficiente do preditor binário contribui para o valor previsto de Y quando o valor da variável binária é 1, mas não tem efeito sobre Y quando a variável binária é 0.

Se na pesquisa na qual está sendo desenvolvida existe um conjunto com c variáveis qualitativas é necessário criar categorias com essas variáveis de forma a atribuir para cada categoria os valores 0 e 1. Mas se há c categorias (assumindo que sejam mutualmente exclusivas) precisa-se apenas de $c - 1$ variáveis binárias para codificar cada observação (DOANE & SEWARD, 2014).

4.2.5 Multicolinearidade

A multicolinearidade ocorre quando qualquer variável independente é altamente correlacionada com um conjunto de outras variáveis independentes. Um caso extremo de colinearidade/multicolinearidade é a singularidade, na qual uma variável independente é perfeitamente prevista (ou seja, correlação de $1,0$) por uma outra variável independente (ou mais de uma). Vale ressaltar que existe uma distinção entre colinearidade e multicolinearidade, de forma que, colinearidade é a associação medida como a correlação entre duas variáveis independentes. A multicolinearidade refere-se à correlação entre três ou mais variáveis independentes (HAIR et. al 2009).

Correlação entre as variáveis independentes pode ter um forte impacto sobre o modelo de regressão. Para maximizar a previsão a partir de um dado número de variáveis independentes, deve-se procurar variáveis independentes que tenham baixa multicolinearidade com as outras variáveis independentes, mas que apresentem correlações elevadas com a variável dependente.

Tolerância (TOL) e Fator de inflação da variância (VIF)

A multicolinearidade é analisada pelas estatísticas Tolerância (TOL) e Fator de Inflação da Variância (VIF). Essas medidas indicam o grau em que cada variável independente é explicada pelas demais variáveis independentes. O cálculo é realizado por meio de várias regressões colocando uma variável independente em função das outras. A TOL é a porção da variabilidade da variável dependente selecionada não explicada pelas demais. Assim, valores muito pequenos denotam colinearidade elevada e vice-versa. Para dado preditor X_j o VIF é definido por:

$$VIF_j = \frac{1}{1 - R_j^2} = \frac{1}{TOL} \tag{4.25}$$

em que $R^2{}_j$ é o coeficiente de determinação quando fazemos a regressão do preditor X_j contra todos os outros preditores (excluindo Y). Adota-se como referência o valor deve ser maior que $0,10$, e para o VIF (que é o inverso da tolerância) o valor deve ser menor que 10 (HAIR et. al 2009).

5 MATERIAL E MÉTODOS

O presente trabalho é uma pesquisa de natureza aplicada, com abordagem quantitativa e com objetivo exploratório e explicativo, pois visa a proporcionar maior familiaridade com o assunto, aprofundando o conhecimento da realidade (PRODOVANI & FREITAS, 2013).

Primeiramente, uma amostra de 159 observações foi obtida a partir de pesquisas em *websites* de diversas imobiliárias da cidade de Toledo-PR, abrangendo todas as regiões da cidade. Dentre as informações disponíveis nos *websites*, foram consideradas as variáveis: o preço de venda da casa (Y), tamanho do lote em m^2 (X_1), área construída em m^2 (X_2), quantidade de banheiros (X_3), quantidade de quartos (X_4), número de vagas na garagem (X_5) e a localização do imóvel indicada pelo bairro (b). A variável b por ser categórica, foi construída por características semelhantes.

Os dados foram organizados em um documento de texto e à partir disso iniciou-se todo o processo de modelagem utilizando a teoria dos modelos de regressão linear múltipla. Um modelo para o valor do imóvel em função das informações coletadas foi construído, de modo a destacar as variáveis que mais influenciam o preço dos imóveis em Toledo-PR.

O *software* R foi utilizado em todo o processo da modelagem, pois é um *software* livre sob os termos da GNU (General Public License, http://www.gnu.org/), que compila e roda em várias plataformas UNIX e sistemas semelhantes (FreeBSD e Linux), Windows e MacOS (R Core Team , 2019).

6 ANÁLISE DOS DADOS E RESULTADOS

No processo da regressão linear múltipla não é sempre que todas as variáveis independentes do modelo são úteis para predizer a variável resposta. Existem métodos que auxiliam na seleção das variáveis que melhor explicam no modelo. O objetivo ao se utilizar esses métodos é encontrar um modelo de regressão linear que contenha o melhor subconjunto de regressores, de modo a desempenhar sua função de forma satisfatória (HAIR et. al, 2009).

Os cálculos aqui presentes foram realizados no *software R* e os *scripts* utilizados encontram-se no Apêndice A.

6.1 Correlação Entre as Variáveis

Antes de iniciar o processo de ajuste do modelo e seleção de variáveis foi realizada uma analise da correlação entre as variáveis quantitativas preditoras candidatas a regressores, sendo elas o preço de venda da casa (Y), tamanho do lote em m^2 (x_1), área construída em m^2 (x_2), quantidade de banheiros (x_3), quantidade de quartos (x_4) e número de vagas na garagem (x_5). As correlações são apresentadas na Tabela (6.1).

Tabela 6.1: Matriz de Correlação entre as Variáveis

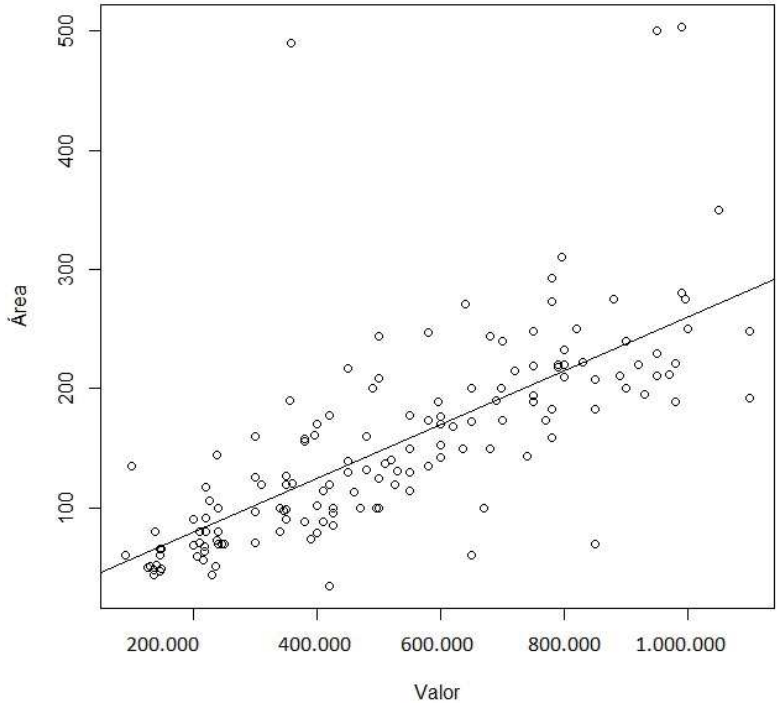
	Y	X_1	X_2	X_3	X_4	X_5
Y	1					
X_1	**0,71**	1				
X_2	**0,72**	0,61	1			
X_3	**0,70**	0,47	0,61	1		
X_4	0,55	0,36	0,46	0,65	1	
X_5	0,52	0,37	0,49	0,50	0,30	1

Fonte: Dos Autores (2019).

Percebe-se que os valores destacados das variáveis tamanho do lote (X_1), área construída (X_2) e quantidade de banheiros (X_3) apresentam correlações mais altas com a variável Y do que as variáveis quantidade de quartos (X_4) e quantidades de vagas na garagem (X_5), ou seja, o tamanho do lote, da área construída e a quantidade de banheiros apresentam uma forte tendência linear com o preço das casas à venda no município de Toledo.

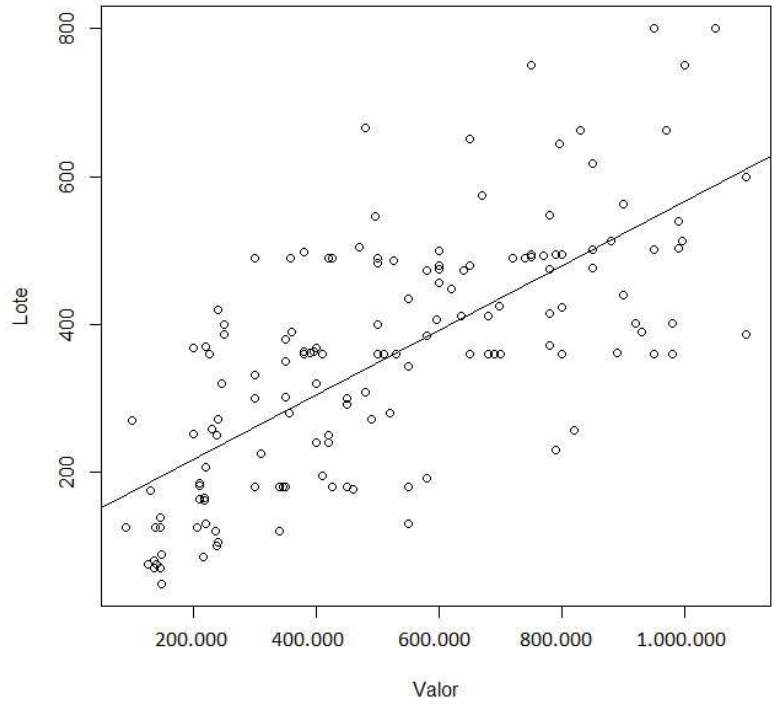
As figuras (6.1) e (6.2) trazem respectivamente gráficos de dispersão entre o valor do imóvel com a área e o valor do imóvel com o lote.

Figura 6.1: Diagrama de Dispersão entre a Área Construída (m^2) e o Valor do Imóvel



Fonte: Dos Autores(2019).

Figura 6.2: Diagrama de Dispersão entre o Tamanho do Lote (m^2) e o Valor do Imóvel



Fonte: Dos Autores (2019).

Observando ambos os gráficos é possível visualizar alguns pontos discrepantes em relação ao comportamento dos demais. Foi utilizado então, o método para detectar esses pontos por meio da função *outlier* no *software R.*Assim, algumas observações foram retiradas da análise, pois pelo teste foram julgadas como atípicas.Assim, o conjunto de dados que possuía 159 observações, passou a ter 133 observações. Essas observações retiradas foram consideradas atípicas, pois não foi considerado no modelo alguma variável para representar o padrão da construção, por exemplo, um dos *outliers* era uma casa de madeira bem antiga com grande área construída, no entanto o valor da casa era muito inferior quando comparado às casas de mesma área construída.

Ao retirar esses pontos, observou-se a mudança na correlação das variáveis independentes em relação a variável resposta. A Tabela (6.2) apresenta as correlações entre as variáveis no conjunto de dados modificado.

Tabela 6.2: Matriz de Correlação entre as Variáveis após a Remoção de Algumas Observações

	Y	X_1	X_2	X_3	X_4	X_5
Y	1					
X_1	0,69	1				
X_2	0,86	0,61	1			
X_3	0,67	0,38	0,68	1		
X_4	0,60	0,41	0,62	0,69	1	
X_5	0,56	0,39	0,47	0,54	0,38	1

Fonte: Dos Autores (2019).

A nova matriz de correlação apresenta que o preço da casa tem alta correlação com a área construída e, em seguida, com o tamanho do lote, a quantidade de banheiros, quantidade de quartos e uma menor correlação com a quantidade de vagas de garagem, mas mesmo assim uma correlação significativa.

6.2 SELEÇÃO DAS VARIÁVEIS PELO MÉTODO *STEPWISE*

Para selecionar as variáveis que melhor irão representar o modelo, utiliza-se o método *Stepwise*. Observa-se pela Tabela (6.2) que mostra todas as correlações entre as 5 variáveis independentes e suas correlações com a dependente (Y, valor do imóvel), que a variável **área** (X_2) tem a mais elevada correlação com a variável dependente $(0,86)$, sendo assim tal variável é a primeira a ser adicionada no modelo.

A Tabela (6.3) mostra as 4 variáveis que foram adicionadas respectivamente no modelo utilizando o *stepwise* no *software R*, seguindo os seguintes passos de adição:

- Passo 1: X_2 Área;
- Passo 2: X_2 Área + X_1 Lote;
- Passo 3: X_2 Área + X_1 Lote + X_5 Vagas;
- Passo 4: X_2 Área + X_1 Lote + X_5 Vagas + X_3 Banheiros.

O método inicia com o modelo de regressão linear simples selecionando a variável independente com maior correlação com a variável dependente, no caso X_2. Depois calcula os coeficientes de correlação parcial para encontrar uma variável independente adicional que explique a maior parte estatisticamente significante da variância não explicada. Calcula novamente a equação de regressão usando duas variáveis independentes e examina o valor da estatística F (da Tabela ANOVA[2]) para ver se ainda há uma contribuição significante. Caso contrário, elimina a variável. Continua o processo examinando todas as variáveis independentes não presentes no modelo para determinar se alguma faria uma adição significativa para equação. Se uma nova for adicionada, o algoritmo verifica todas as variáveis independentes novamente para determinar se elas devem ser mantidas. O procedimento finaliza quando nenhuma das candidatas para inclusão contribuam signicativamente para o modelo (HAI et al, 2009).

Tabela 6.3: Ajuste Geral do Modelo Seguindo os Passos do Método de Seleção de Variáveis Stepwise

Passo	**R**	R^2	$R^2{}_{ajustado}$	Erro Padrão	**Mudança no R²**	**F**	**Valor-p** [3]
1	0,86	0,74	0,74	$1,25 \cdot 10^{11}$	0,74	376,38	0,00
2	0,88	0,78	0,78	$1,16 \cdot 10^{11}$	0,04	233,31	0,00
3	0,89	0,80	0,80	$1,10 \cdot 10^{11}$	0,02	176,90	0,00
4	0,90	0,81	0,80	$1,09 \cdot 10^{11}$	0,00	137,25	0,00

Fonte: Dos Autores (2019).

Pela tabela anterior pode-se notar que cada uma das duas primeiras variáveis adicionadas à equação fez contribuições substanciais ao ajuste geral do modelo, com significativos aumentos no R^2 e no $R^2{}_{ajustado}$, ao mesmo tempo que diminui o erro padrão da estimativa.

Observa-se que no passo 4 o teste de significância global, que sob H_0 temos $F \sim F_{4,128;0,05}$. Como $F = 137,25 > 2,44$ rejeita-se H_0 (valor-p $< 0,05$). Com 5% de significância, afirma-se que pelo menos uma variável do modelo é estatisticamente aceita para explicar o preço do imóvel.

[2]Ver Tabela 4.3

[3]Todos os valores-p são significativo ao nível de 5%.

Com apenas as duas primeiras variáveis, 78% do valor do imóvel é explicada com tais variáveis. Duas variáveis adicionais foram acrescentadas para chegar no modelo final, mas essas variáveis apesar de estatisticamente significantes, fazem contribuições menores.

A Tabela (6.4) traz os coeficientes tanto o real quanto o padronizado do modelo construído, juntamente com a significância estatística (estatística t e valor-p) e a estatística de multicolinearidade (tolerância e VIF).

Tabela 6.4: Coeficientes e Testes Estatísticos das Variáveis Quantitativas

Variáveis	β	**Erro Padrão**	β'	**Estatística t**	**valor-P**	**VIF**
(Constante)	$-76821,20$	$26736,15$	-	$-2,87$	$0,00$	-
x_2 Área	$2166,48$	$84,69$	$0,57$	$9,33$	$0,00$	$2,54$
x_1 Lote	$408,39$	$232,15$	$0,24$	$4,82$	$0,00$	$1,65$
x_5 Vaga	$30408,11$	$12854,41$	$0,14$	$2,91$	$0,00$	$1,49$
x_3 Banheiro	$26902,39$	$10443,28$	$0,12$	$2,09$	$0,03$	$2,11$

Fonte: Dos Autores(2019).

Por β é possível observar os valores de cada coeficiente atribuído ao modelo, e por meio de β' pode-se classificar qual coeficiente apresenta o maior grau de explicação no modelo, sendo este o coeficiente da variável X_2, ou seja, a área do imóvel. Pode-se observar que todas as variáveis são significativas para o modelo, pois apresentam o valor-p inferior a $0,05$, ou seja, o teste t rejeitou a hipótese nula de que o coeficiente $\beta_j = 0, j = 1, \ldots, 4$.

A presença de multicolinearidade foi avaliada por meio do fator de inflação da variância (VIF). A coluna do VIF na Tabela (6.4) mostra que os mesmos foram inferiores a 10 para as quatro variáveis independentes, ou seja, a multicolinearidade não influenciou nas estimativas dos parâmetros do modelo.

Assim, a equação que representa o modelo final dado pelo *Stepwise* é dada por:

$$Y = -76821,20 + 408,39x_1 + 2166,48x_2 + 26902,39x_3 + 30408,11x_5 \qquad (6.1)$$

em que x_1= Tamanho do Lote (m^2),x_2=Área Construída (m^2),x_3= Quantidade de Banheiros e x_5= Número de Vagas.

6.3 VARIÁVEL BAIRRO

Na amostra coletada, existiam casas de 20 bairros diferentes, assim, para incluí-los no modelo seria necessário a utilização de 20 variáveis *Dummy*. A inclusão de tantas variáveis não seria vantajoso, visto que o objetivo é um modelo mais simples para explicar a variação dos preços das casas. Assim, foi discutido uma forma de agrupar os bairros e

chegou-se a conclusão que dois grupos seria uma alternativa. Os bairros foram agrupados da seguinte forma:

- Grupo 1: Bairros com imóveis, geralmente, de maior valor comercial. Contém os bairros nobres da cidade.
- Grupo 2: Bairros com imóveis, geralmente, de menor valor comercial. Bairros mais populares e antigos.

A Tabela (6.5) apresenta os dois grupos definidos.

Tabela 6.5: Grupos Criados nos Agrupamentos dos Bairros

Grupo 1	Grupo 2
Jardim La Salle, Vila Becker Centro, Jardim Gisela, Jardim Pancera Jardim Porto Alegre, Vila Industrial	Jardim Europa, Vila Boa Esperança Jardim Bressan, Vila Panorama, Jardim Pinheirinho Vila São Francisco, Jardim Coopagro Vila Santa Clara IV, Jardim Parizotto Jardim Concórdia, Jardim Paulista, Vila Pioneiro

Fonte: Dos Autores (2019).

Assim, o número de variáveis *Dummy* no modelo foi reduzido de 20 variáveis para apenas 2, sendo o grupo 1 (b_1) e o grupo 2 (b_2). Dessa forma, para realizar a estimação do valor de uma casa por meio do modelo final, deve-se utilizar a Tabela (6.5) para saber em qual grupo a casa em questão está inserida.

Para a adição da variável bairro, foram criados mais 3 modelos. Na Tabela (6.6) abaixo foi analisado a mudança estatística da adição das variáveis *dummy* no modelo já criado pelo método *stepwise*.

Tabela 6.6: Ajuste Geral do Modelo com a Adição das Variáveis *Dummy*

Variáveis	**R**	R^2	$R^2_{ajustado}$	**Erro Padrão**	**Mudança no R^2**	**F**	**Sig.**
Stepwise	$0,9005$	$0,8109$	$0,8050$	$1,08674 \cdot 10^{11}$	-	$137,25$	$0,00$
b_1	$0,9129$	$0,8334$	$0,8269$	$102390,484$	$0,0220$	$127,13$	$0,00$
b_2	$0,9149$	$0,8371$	$0,8307$	$101252,5834$	$0,0271$	$130,5$	$0,00$
b_1 e b_2	$0,9149$	$0,8371$	$0,8294$	$101645,628$	$0,0271$	$107,98$	$0,00$

Fonte: Dos Autores (2019).

Analisando os dados da Tabela (6.6), o melhor modelo seria com a adição da variável b_2, pois, não houve alteração no valor de R^2 e o $R^2_{ajustado}$ com a adição da variável b_2 é o maior entre os modelos.

Tabela 6.7: Coeficientes e Testes Estatísticos Parciais do Modelo Final

Variáveis	β	**Erro Padrão**	β'	**Estatística t**	**valor-P** [3]	**VIF**
(Constante)	$39435,21$	$35796,39$	-	$1,10$	$0,27$	-
Área x_2	$1969,19$	$81,43$	$0,52$	$8,92$	$0,00$	$2,64$
Lote x_1	$317,50$	$220,66$	$0,19$	$3,89$	$0,00$	$1,76$
Vaga x_5	$23309,13$	$11982,88$	$0,10$	$2,36$	$0,02$	$1,53$
Banheiro x_3	$25143,96$	$9855,91$	$0,11$	$2,09$	$0,04$	$2,11$
Grupo 2 b_2	$-100105,40$	$22135,77$	$-0,20$	$-4,52$	$0,00$	$1,53$

Fonte: Dos Autores (2019).

Com a adição da variável *dummy* b_2 a tabela (6.7) acima, apresenta os coeficientes do novo modelo e os testes estatísticos.

Por meio da Tabela (6.7) pode-se identificar os coeficientes do modelo final e também analisar qual apresentou mais representatividade no modelo, permanecendo assim a variável x_2 com $0,52$ de coeficiente padronizado. É interessante observar que o coeficiente da variável *dummy* b_2 é negativo, sendo o padronizado igual a $-0,2$, ou seja, o preço do imóvel é impactado negativamente em 20% se o imóvel estiver em um dos bairros do grupo 2.

Analisando os valores estatísticos realizados pelo teste t, pode-se concluir que todas as variáveis são significativas para o modelo, exceto a constante, pois todas apresentaram valor-p $< 0,05$.

Avaliando a multicolinearidade, os valores VIF foram inferiores a 10 nas cinco variáveis explicativas avaliadas, comprovando que a multicolinearidade não influenciou nas estimativas dos parâmetros do modelo.

Após acrescentar a variável b_2 realizou-se o teste global de significância do modelo final, como pode-se observar na tabela ANOVA a seguir:

Tabela 6.8: Tabela ANOVA para o Modelo Final

Fonte de variação	**gl**	**SQ**	**MQ**	**F**	**F de significação**
Regressão	5	$6,69 \cdot 10^{12}$	$1,34 \cdot 10^{12}$	$130,56$	$0,00$
Resíduo	127	$1,3 \cdot 10^{12}$	$1,03 \cdot 10^{10}$		
Total	132	$8 \cdot 10^{12}$			

Fonte: Dos Autores (2019).

Analisando a tabela 6.8 sob H_0 temos F $F_{5,127;0,05}$. Como $F = 130,56 > 2,29$ rejeita-se H_0 (valor-p $< 0,05$). Com 5% de significância, afirma-se que pelo menos uma variável do modelo é estatisticamente significante para explicar o preço do imóvel.

6.4 MODELO FINAL

Após as realizações dos testes para selecionar as variáveis presentes no modelo que explicam a variável Y (valor do imóvel), chegou-se ao conjunto de regressores composto pelas variáveis preditoras X_1, X_2, X_3, X_5 e b_2 que representam respectivamente a área do imóvel, a área do lote, a quantidade de banheiros, quantidade de vagas na garagem e pertencer a um grupo com alguns bairros do município. O modelo de regressão final é apresentado pela equação

$$Y = 39435,21 + 317,50x_1 + 1969,19x_2 + 25143,96x_3 + 23309,13x_5 - 100105,40b_2 \quad (6.2)$$

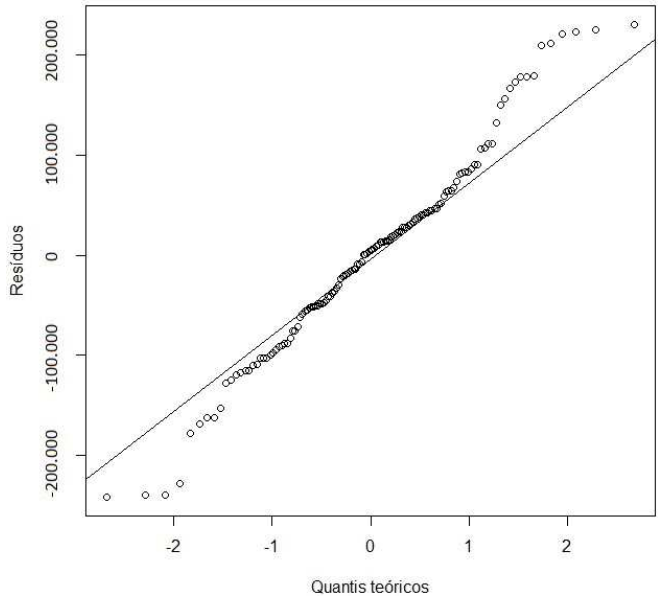
Para estimar o preço de venda do imóvel utilizando o modelo, deve-se substituir as variáveis x_1, x_2, x_3, x_5 e b_2 na equação, pelo tamanho da área construída, tamanho do lote, quantidade de banheiros, vagas de garagem do imóvel e para a variável *dummy* é preciso especificar o bairro do imóvel nos grupos expostos pela Tabela (6.5), da seguinte forma: se o imóvel a ser previsto está localizado no grupo 2 o valor de $b_2 = 1$, caso contrário, $b_2 = 0$.

6.4.1 Validação do Modelo

Nesta seção é apresentada a validação das suposições feitas sobre os resíduos, que são: os erros têm distribuição normal, os erros têm variância constante (isto é, eles são homocedásticos) e os erros são independentes (isto é, eles não são autocorrelacionados).

O gráfico q-q plot na Figura (6.3) verifica a hipótese de normalidade dos resíduos. Se os valores estão ao longo da diagonal sem desvios evidentes, afirma-se que os resíduos tem distribuição normal. Como nos extremos da reta, nota-se um desvio, foi realizado o teste de normalidade Shapiro-Wilk para verificar a normalidade. A estatística W = 0,98, com valor-p = 0,09, logo aceita-se a hipótese de normalidade dos resíduos, ao nível de 5% de significância.

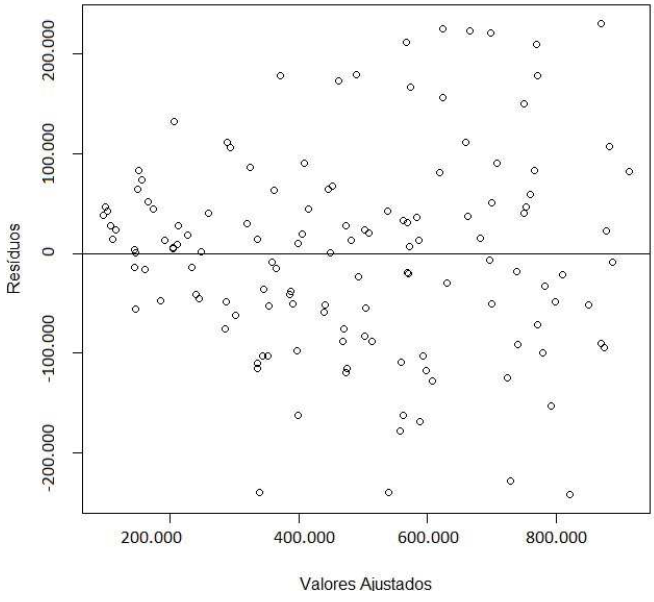
Figura 6.3: Resíduos x Quantis teóricos



Fonte: Dos Autores (2019).

A verificação se os resíduos têm variância constante é denominada de homocedasticidade do modelo. É a constância dos resíduos ao longo dos valores das variáveis independentes. A análise é realizada por meio do gráfico dos valores ajustados *versus* resíduos, no gráfico da Figura (6.4), que não mostra padrão de resíduos crescentes ou decrescentes. Pode-se afirmar que a variância dos erros é constante.

Figura 6.4: Resíduos x Valores Ajustados



Fonte:Dos Autores (2019).

Para verificar a hipótese de independência dos erros, ou seja, não existe autocorrelação, pode-se verificar por meio do teste Durbin-Watson. A estatística foi de DW=1,73 com valor-p de 0,06, ou seja, aceita-se a hipótese nula de que a correlação entre os resíduos é zero.

6.4.2 Avaliação prática do modelo construído

Segundo Gazola (2012), a construção de um modelo que satisfaz todas as suposições teóricas não é o suficiente para garantir a qualidade das predições. Uma avaliação prática do modelo mostrará a sua qualidade de ajuste e capacidade preditiva. O autor afirma que nos estudos encontrados em seu referencial teórico, os pesquisadores não apresentam dados que mostram a qualidade do ajuste.

A prática que o autor propõe consiste em avaliar os erros percentuais dentro das faixas de erros de 5%, 10%, 15%, 20%, 30% 40% e 50%, no estudo aqui, entre o preço observado do imóvel e o preço predito pelo modelo. Os resultados seguem na Tabela (6.9).

Tabela 6.9: Percentual de Valores Preditos na Faixa de 5% a 50%

Faixas de Erro	**Nº de preços preditos**	**% dos valores preditos**
5%	30	22,6
10%	52	39,1
15%	70	52,6
20%	86	64,7
30%	108	81,2
40%	123	92,5
50%	133	100,0

Fonte: Dos Autores (2019).

Nota-se na Tabela (6.9) que 10 imóveis tiveram valor de predição com erro superior a 40%, sendo que o maior erro individual foi de 48,5%

Para melhor clareza e entendimento do que significam os resultados da Tabela 6.9, tomou-se um imóvel de valor $R\$40.000,00$ e os valores preditos nas respectivas faixas de erro são apresentadas na Tabela (6.10).

Tabela 6.10: Exemplo de Variação do Preço Predito de um Imóvel de Valor de $R\$40.000,00$

Faixas de Erro	**Valor de variação**	**Variação do valor predito**
5%	$2000,00$	$(38000,00\ ;\ 42000,00)$
10%	$4000,00$	$(36000,00\ ;\ 44000,00)$
15%	$6000,00$	$(34000,00\ ;\ 46000,00)$
20%	$8000,00$	$(32000,00\ ;\ 48000,00)$
30%	$12000,00$	$(28000,00\ ;\ 52000,00)$

Fonte: Dos Autores (2019).

Nos estudos realizados por Worzala et al. (1995), são citados resultados entre 24% e 37% de imóveis dentro da faixa de 5%. Assim, os resultados citados pelo apoiam a conclusão de que o modelo encontrado neste trabalho, com 22,6% dos imóveis na faixa de 5%, 92,5% na faixa de 40% e maior erro individual de 48,5%, é considerado satisfatório.

7 CONSIDERAÇÕES FINAIS

A partir de dados reais cadastrados em *websites* de imobiliárias da cidade de Toledo-PR, foi possível estudar e aplicar a teoria do modelo de regressão linear múltipla e por meio da estatística, entender um pouco mais sobre o relacionamento entre estas variáveis.

Quando trabalha-se com modelagem, em específico a regressão linear múltipla, almeja-se um modelo eficaz e eficiente, ou seja, um modelo que descreva o comportamento de uma variável específica, e ao mesmo tempo que o resultado seja simples e parcimonioso. Por isso a escolha do método *stepwise* para seleção de variáveis, assim, variáveis redundantes seriam retiradas do modelo, como aconteceu com a variável quantidade de quartos (X_4), que parecia ser importante para determinar o valor do preço do imóvel, mas não entrou no modelo.

Foi possível apresentar um modelo que consegue prever 83% da variabilidade do valor de um imóvel situado no município de Toledo-PR, onde a variável (X_2) que representa a área construída foi a mais representativa. Com o modelo, situações de casas superestimadas ou subestimadas seriam facilmente detectadas e, ao mesmo tempo, o modelo proporciona facilidade para prever o preço do imóvel quando deseja-se fazer uma especulação.

O trabalho é uma contribuição para o mercado imobiliário, pois o preço de uma casa, na cidade de Toledo-PR, pode ser estimado com base em suas características. Porém, vale ressaltar que outros modelos podem ser gerados com uma quantidade maior de variáveis explicativas, ou com uma amostra contendo vários tipos de imóveis, como apartamentos, imóveis comerciais, trazendo assim outros benefícios para o setor imobiliário.

O trabalho também foi uma oportunidade de vivenciar a aplicabilidade da matemática na situação cotidiana, estudar conceitos não vistos na graduação como a teoria dos modelos de regressão linear múltipla, a maior utilização do *software R* na modelagem, melhorar o aprendizado da escrita de um trabalho científico e melhorar a oratória por meio de vários seminários sobre o assunto apresentados.

Que o trabalho possa ser referencial para futuros estudos sobre os modelos de regressão linear múltipla e sua aplicação a dados reais, principalmente para alunos de graduação em seus trabalhos de conclusão de curso. Como trabalho futuro, pretende-se fazer previsões fora da amostra para os valores dos preços das casas, desenvolver os intervalos de confiança, e com essas informações, elaborar um artigo para publicação em evento científico ou revista.

REFERÊNCIAS

BAPTISTELLA, M. STEINER, M, T,A. NETO, A, C **O uso de redes neurais e regressão linear múltipla na engenharia de avaliações: determinação dos valores venais de imóveis urbanos.** Goiânia, 2006.

BARNETT, V.; LEWIS, T. **Outliers in statistical data**. Chichester: John Wiley, 1996

DOANE, D, P. SEWARD, L.E **Estatística aplicada à administração e economia**. 4. ed. Porto Alegre: AMGH, 2014.

GAZOLA, S. **Construção de um modelo de regressão para avaliação de imóveis**. Florianópolis, 2002.

HAIR Jr., J.F.; BLACK, W.C.; BABIN, B.J.; ANDERSON, R.E. TATHAM, R.L. **Análise multivariada de dados**. 6.ed. Porto Alegre, Bookman, 2009. 688p.

JÚNIOR, I. OLIVEIRA, V, C. ARAÚJO, L,A,S. RAMOS, P,C,F **Aplicação da ferramenta estatística de analise de regressão numa fazenda de cultivo de camarão marinho no estado do Rio Grande do Norte.** Florianópolis, 2004.

MARTINS, G. **A Estatística Geral e Aplicada** 3 ed.- São Paulo : Atlas,2005.

Mercado imobiliário em Toledo espera crescimento. **Jornal do Oeste**. Toledo, 22 de outubro 2019. Disponível em: https://www.jornaldooeste.com.br/noticia/mercado-imobiliario-em-toledo-espera-crescimento-neste-semestre-de-2019. Acesso em: 22 de nov 2019.

MONTGOMERY, D. C.; RUNGER, G. C.**Estatística e Probabilidade para Engenheiros** 2 ed. Rio de Janeiro: LTC, 2003.

PEREIRA, J. C.; GARSON, S.; ARAÚJO, E. G. **Construção de um modelo para o preço de venda de casas residenciais na cidade de Sorocaba-SP**.GEPROS. Gestão da Produção, Operações e Sistemas, Ano 7, nº 4, out-dez/2012, p. 153-167.

PRODOVANI, C. C; FREITAS, E. C. **Metodologia do trabalho científico: Métodos e Técnicas da pesquisa e do trabalho acadêmico**. Novo Hamburgo, 2013.

R Core Team (2019). R: **A language and environment for statistical computing**. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

ROZENBAUN, S.; MACEDO-SOARES, T.D.L.V.A. **Proposta para Construção de Um Índice Local de Preços de Imóveis a Partir dos Lançamentos Imobiliários**

de Condomínios Residenciais. Rev. Adm. Pública. Rio de Janeiro, v. 41, n. 6, p. 1069-1094, 2007.

RODRIGUES, SANDRA C. **Modelo de Regressão linear e suas aplicações**.Govilhã, 2012.

SEWARD, LORI E.; DOANE, DAVID P. **Estatística Aplicada à Administração e Economia-4**. AMGH Editora, 2014.

TABACHNICK, B. FIDELL, L. S. **Using multivariate statistics** 3a ed.). New York: Harper Collins. 1996

Toledo é a 7ª cidade com maior crescimento do Brasil. **Gazeta Toledo**. 29 de jun de 2018. Disponível em: https://www.gazetatoledo.com.br/NOTICIA/38779. Acesso em: 21 de nov de 2019

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

8 APÊNDICE

8.1 Apêndice A: Script do software R

```
1
2  #Leitura dos Dados
3  preco = read.table("C:\\Users\\rapha\\Documents\\TCC\\Dados\\preco.txt",
       header=T)
4  attach(preco)
5  preco
6  cor(preco)
7
8
9  #Modelos
10 Modelo1 = lm(valor ~ area, data = preco)
11 summary(Modelo1)$r.squared
12 summary(Modelo1)$adj.r.squared
13 preco.anova1 <- aov(preco$valor ~ preco$area, data=preco)
14 summary(preco.anova1)
15 require(faraway)
16 vif(Modelo1)
17 cor(modelo1)
18
19
20 Modelo2 = lm(valor ~ lote, data = preco)
21 summary(Modelo2)$r.squared
22 summary(Modelo2)$adj.r.squared
23 preco.anova2 <- aov(preco$valor ~ preco$lote, data=preco)
24 summary(preco.anova2)
25 require(faraway)
26 vif(Modelo2)
27 cor(Modelo2)
28
29 Modelo3 = lm(valor ~ ban , data = preco)
30 summary(Modelo3)$r.squared
31 summary(Modelo3)$adj.r.squared
32 preco.anova3 <- aov(preco$valor ~ preco$ban, data=preco)
33 summary(preco.anova3)
34
35 Modelo4 = lm(valor ~ quar , data = preco)
36 summary(Modelo4)$r.squared
37 summary(Modelo4)$adj.r.squared
```

```
38 preco.anova4 <- aov(preco$valor ~ preco$quar, data=preco)
39 summary(preco.anova4)
40
41 Modelo5 = lm(valor ~ vaga , data = preco)
42 summary(Modelo5)$r.squared
43 summary(Modelo5)$adj.r.squared
44
45 Modelo6 = lm(valor ~ ba , data = preco)
46 summary(Modelo6)$r.squared
47 summary(Modelo6)$adj.r.squared
48
49 Modelo7= lm(valor ~ bb , data = preco)
50 summary(Modelo7)$r.squared
51 summary(Modelo7)$adj.r.squared
52
53 Modelo67= lm(valor ~ ba + bb , data = preco)
54 summary(Modelo67)$r.squared
55 summary(Modelo67)$adj.r.squared
56
57 Modelo12= lm(valor ~ area + lote , data = preco)
58 summary(Modelo12)$r.squared
59 summary(Modelo12)$adj.r.squared
60 preco.anova12 <- aov(preco$valor ~ preco$area + preco$lote, data=preco)
61 summary(preco.anova12)
62 require(faraway)
63 vif(Modelo12)
64 cor(Modelo12)
65
66 Modelo13= lm(valor ~ area + ban , data = preco)
67 summary(Modelo13)$r.squared
68 summary(Modelo13)$adj.r.squared
69
70 Modelo14= lm(valor ~ area + quar , data = preco)
71 summary(Modelo14)$r.squared
72 summary(Modelo14)$adj.r.squared
73
74 Modelo15= lm(valor ~ area + vaga , data = preco)
75 summary(Modelo15)$r.squared
76 summary(Modelo15)$adj.r.squared
77
78 Modelo23= lm(valor ~ lote + ban , data = preco)
79 summary(Modelo23)$r.squared
80 summary(Modelo23)$adj.r.squared
81
82 Modelo24= lm(valor ~ lote + quar , data = preco)
83 summary(Modelo24)$r.squared
84 summary(Modelo24)$adj.r.squared
```

```
85
86 Modelo25= lm(valor ~ lote + vaga , data = preco)
87 summary(Modelo25)$r.squared
88 summary(Modelo25)$adj.r.squared
89
90 Modelo34= lm(valor ~ ban + quar , data = preco)
91 summary(Modelo34)$r.squared
92 summary(Modelo34)$adj.r.squared
93
94 Modelo35= lm(valor ~ ban + vaga , data = preco)
95 summary(Modelo35)$r.squared
96 summary(Modelo35)$adj.r.squared
97
98 Modelo45= lm(valor ~ quar + vaga , data = preco)
99 summary(Modelo45)$r.squared
100 summary(Modelo45)$adj.r.squared
101
102 Modelo123= lm(valor ~ area + lote + ban , data = preco)
103 summary(Modelo123)$r.squared
104 summary(Modelo123)$adj.r.squared
105 preco.anova123 <- aov(preco$valor ~ preco$area + preco$lote + preco$ban,
      data=preco)
106 summary(preco.anova123)
107
108
109 Modelo124= lm(valor ~ area + lote + quar , data = preco)
110 summary(Modelo124)$r.squared
111 summary(Modelo124)$adj.r.squared
112
113 Modelo125= lm(valor ~ area + lote + vaga , data = preco)
114 summary(Modelo125)$r.squared
115 summary(Modelo125)$adj.r.squared
116
117 Modelo134= lm(valor ~ area + ban + quar , data = preco)
118 summary(Modelo134)$r.squared
119 summary(Modelo134)$adj.r.squared
120
121 Modelo135= lm(valor ~ area + ban + vaga , data = preco)
122 summary(Modelo135)$r.squared
123 summary(Modelo135)$adj.r.squared
124
125 Modelo145= lm(valor ~ area + quar + vaga , data = preco)
126 summary(Modelo145)$r.squared
127 summary(Modelo145)$adj.r.squared
128
129 Modelo234= lm(valor ~ lote + ban + quar , data = preco)
130 summary(Modelo234)$r.squared
```

```
131 summary(Modelo234)$adj.r.squared
132
133 Modelo235= lm(valor ~ lote + ban + vaga , data = preco)
134 summary(Modelo235)$r.squared
135 summary(Modelo235)$adj.r.squared
136
137 Modelo345= lm(valor ~ ban + quar + vaga , data = preco)
138 summary(Modelo345)$r.squared
139 summary(Modelo345)$adj.r.squared
140
141 Modelo1234= lm(valor ~ area + lote + ban + quar , data = preco)
142 summary(Modelo1234)$r.squared
143 summary(Modelo1234)$adj.r.squared
144
145 Modelo1235= lm(valor ~ area + lote + ban + vaga , data = preco)
146 summary(Modelo1235)$r.squared
147 summary(Modelo1235)$adj.r.squared
148
149 Modelo1245= lm(valor ~ area + lote + quar + vaga , data = preco)
150 summary(Modelo1245)$r.squared
151 summary(Modelo1245)$adj.r.squared
152
153 Modelo1345= lm(valor ~ area + ban + quar + vaga , data = preco)
154 summary(Modelo1345)$r.squared
155 summary(Modelo1345)$adj.r.squared
156
157 Modelo2345= lm(valor ~ lote + ban + quar + vaga , data = preco)
158 summary(Modelo2345)$r.squared
159 summary(Modelo2345)$adj.r.squared
160
161 Modelo12345= lm(valor ~ area + lote + ban + quar + vaga , data = preco)
162 summary(Modelo12345)$r.squared
163 summary(Modelo12345)$adj.r.squared
164
165 Modelo123456= lm(valor ~ area + lote + ban + quar + vaga + ba, data = preco
      )
166 summary(Modelo123456)$r.squared
167 summary(Modelo123456)$adj.r.squared
168 preco.anova123456 <- aov(preco$valor ~ preco$area + preco$lote + preco$ban
      + preco$quar + preco$vaga + preco$ba, data=preco)
169 summary(preco.anova123456)
170
171 which(rstudent(Modelo1234567)>2)
172
173 Modelo12357= lm(valor ~ area + lote + ban + vaga + bb, data = preco)
174 summary(Modelo123457)$r.squared
175 summary(Modelo123457)$adj.r.squared
```

```
176 preco.anova12357 <- aov(preco$valor ~  preco$area + preco$lote + preco$ban
      + preco$vaga  + preco$bb, data=preco)
177 summary(preco.anova12357)
178 cor(Modelo12357)
179
180 Modelo12367= lm(valor ~ area + lote + ban + ba + bb , data = preco)
181 summary(Modelo12367)$r.squared
182 cor(Modelo12367)
183 summary(Modelo12367)$adj.r.squared
184 preco.anova12367 <- aov(preco$valor ~  preco$area + preco$lote + preco$ban
      + preco$ba + preco$bb, data=preco)
185 summary(preco.anova123)
186
187 Modelo12357= lm(valor ~ area + lote + ban + vaga + bb, data = preco)
188 summary(Modelo12357)$r.squared
189 summary(Modelo12357)$adj.r.squared
190
191
192 #Outlier
193
194 which(rstudent(lm1) > 2)
195
196 #Stepwise
197 nulo = lm(valor ~ 1, data=preco)
198 completo = lm(valor ~ . ,data=preco)
199
200 step(completo, data=preco, direction="backward",trace=FALSE)
201
202 step(nulo, scope=list(lower=nulo, upper=completo), data=preco, + direction
      ="forward",trace=FALSE)
203
204 #Multicolinearidade
205
206 library(car)
207
208 vif(lm1)
209
210
211 #Gr ficos
212
213 windows()
214 model <-  lm( rea ~Valor)
215 plot( rea ~Valor)
216 abline(model)
217
218 plot(Lote~Valor)
219 plot( rea ~Valor)
```

```
220
221 windows()
222 modell <-  lm(Lote~Valor)
223 plot(Lote~Valor)
224 abline(modell)
225
226
227 influence.measures(preco)
228 lmstep <- lm(valor ~ area + lote + ban + vaga +bb , data = preco)
229
230 summary(lmstep)
231
232 plot(lmstep, which = 1)
233
234 plot(lmstep, which = 2)
235
236 windows()
237 plot(fitted(Modelo12357),residuals(Modelo12357),xlab="Valores Ajustados",
      ylab="Res duos")
238 abline(h=0)
239
240 shapiro.test(residuals(Modelo12357))
241 Shapiro-Wilk normality test
242 data: residuals(Modelo12357)
243
244 windows()
245 qqnorm(residuals(Modelo12357), ylab="Res duos",xlab="Quantis te ricos",
      main="")
246 qqline(residuals(Modelo12357))
247
248 eruption.lm = lm(valor ~ area + lote + ban + vaga +bb , data = preco)
249 eruption.stdres = rstandard(eruption.lm)
250
251 plot(fitted(eruption.lm), eruption.stdres,
252      ylab="Standardized Residuals",
253      xlab="Waiting Time",
254      main="Old Faithful Eruptions")
255 abline(0, 0)
256
257 #Durbin- Watson
258 require(lmtest)
259
260 dwtest(Modelo12357)
261
262 #Breush- Pagan
263 bptest(Modelo12357)
264
```

```
265 #Anderson- Daring
266 shapiro.test(residuals(Modelo12357))
```