



DOI: 10.5281/zenodo.18929531

Analizando o Mercado Imobiliário Através Da Regressão Linear Múltipla

João Victor Oliveira Honorato¹

Lívia Santiago Pelegrino²

Romário Júnior Coimbra Paulino³

Patrícia Lusié Velozo da Costa⁴

Resumo

Este estudo investiga os fatores que influenciam os preços dos imóveis residenciais por meio de um Modelo de Regressão Linear Múltipla, analisando a relação entre o logaritmo do preço e características estruturais, como metragem, ano de construção e número de vagas de garagem. Utilizou-se a base de dados *Ames Housing*, que reúne informações detalhadas sobre imóveis vendidos na cidade de Ames, Iowa, nos Estados Unidos da América (EUA), ao longo de vários anos, contendo variáveis relacionadas a características físicas, localização e condições de venda das propriedades. A modelagem estatística permitiu identificar padrões e quantificar o impacto dessas variáveis no mercado imobiliário, com todos os coeficientes do modelo se mostrando estatisticamente significativos. Apesar de algumas limitações associadas aos pressupostos clássicos, o modelo apresentou boa capacidade preditiva, com os intervalos de previsão contemplando a maior parte dos valores observados. Os resultados obtidos auxiliam na compreensão dos principais determinantes do valor dos imóveis e oferecem suporte para decisões mais informadas por parte de compradores, vendedores e investidores. Além disso, o estudo reforça a importância da análise de dados no setor, evidenciando o potencial das técnicas estatísticas para a obtenção de estimativas consistentes e informativas. **Palavras-chave:** Regressão Linear Múltipla, Mercado Imobiliário, Modelagem Estatística.

Abstract

This study investigates the factors that influence residential property prices through a Multiple Linear Regression Model, analyzing the relationship between the logarithm of the price and structural characteristics such as floor area, year of construction, and number of garage spaces. The Ames Housing dataset was used, which contains detailed information on properties sold in the city of Ames, Iowa, United

¹Universidade Federal Fluminense (UFF), joaohonorato@id.uff.br

²Universidade Federal Fluminense (UFF), liviapelegrino@id.uff.br

³Universidade Federal Fluminense (UFF), romariojunior@id.uff.br

⁴Universidade Federal Fluminense (UFF), patricialusie@id.uff.br

States of America (USA) over several years, including variables related to physical attributes, location, and sales conditions of the properties. The statistical modeling allowed the identification of patterns and the quantification of the impact of these variables on the real estate market, with all model coefficients proving to be statistically significant. Despite some limitations associated with classical assumptions, the model showed good predictive capability, with the prediction intervals covering most of the observed values. These findings contribute to a deeper understanding of price formation in the real estate market and can support strategic decision-making by buyers, sellers, and investors. Furthermore, the study highlights the importance of data analysis in the sector, demonstrating the potential of statistical techniques to obtain consistent and informative estimates. **Keywords:** Multiple Linear Regression, Real Estate Market, Statistical Modeling.

Introdução

O mercado imobiliário residencial é um setor dinâmico, influenciado por uma ampla variedade de fatores, incluindo características estruturais dos imóveis, localização e condições econômicas. A previsão dos preços das propriedades desempenha um papel fundamental para compradores, vendedores, investidores e formuladores de políticas, ao subsidiar decisões estratégicas e promover a alocação eficiente de recursos. Para compreender essa complexidade, diversas abordagens estatísticas podem ser aplicadas, destacando-se o modelo de regressão linear múltipla. Sua principal vantagem reside na capacidade de quantificar a relação entre múltiplas variáveis explicativas e o valor dos imóveis, além de proporcionar interpretações claras sobre o impacto individual de cada fator no preço final.

Chen (2022), Mao (2023) e Santos et al. (2021) modelaram os preços de imóveis utilizando técnicas de regressão linear múltipla, mas com enfoques distintos. Chen (2022) propôs um modelo de previsão de preços de imóveis na cidade de Zhaoqing, combinando análise de correlação e regressão linear múltipla para identificar os fatores mais influentes e gerar previsões precisas. Mao (2023), por sua vez, utilizou a regressão linear em conjunto com técnicas de aprendizado de máquina para explorar como essas abordagens podem melhorar a previsão dos preços imobiliários em diversos cenários. Já Santos et al. (2021) aplicaram a regressão linear múltipla no município de São Paulo, focando em variáveis como área construída e número de quartos para modelar o preço dos imóveis residenciais. Apesar das diferenças de contexto e abordagem, os três estudos destacaram a importância da regressão linear como ferramenta para prever preços de imóveis e auxiliar na tomada de decisões no mercado imobiliário.

Este estudo empregou um Modelo de Regressão Linear Múltipla para investigar de que forma características como metragem, número de vagas de garagem e ano de construção influenciam os preços das propriedades. Os dados analisados foram obtidos a partir da base *Ames Housing*, que reúne informações detalhadas sobre imóveis vendidos na cidade

de Ames, Iowa, nos Estados Unidos da América (EUA), ao longo de vários anos. O logaritmo do valor de venda das propriedades foi utilizado como variável resposta, enquanto três variáveis explicativas foram selecionadas com base em sua relevância para os objetivos da análise. O modelo ajustado apresentou coeficientes estatisticamente significativos e boa capacidade preditiva, com os intervalos de previsão abrangendo a maior parte dos valores observados. Essa abordagem permitiu identificar padrões e mensurar o impacto de diferentes atributos na formação dos preços, evidenciando o potencial da modelagem estatística para gerar previsões fundamentadas e apoiar uma análise consistente do mercado imobiliário.

Objetivo

Este trabalho originou-se como parte da disciplina de *Modelos Lineares 1* e tem como objetivo a aplicação de um Modelo de Regressão Linear Múltipla em um conjunto de dados reais. Utilizou-se dados sobre vendas de imóveis, com o intuito de analisar a influência de características estruturais das propriedades, como metragem, número de vagas de garagem e ano de construção, sobre os preços do mercado imobiliário. Buscou-se quantificar o impacto desses fatores no valor das propriedades e oferecer uma compreensão aprofundada dos determinantes da precificação. Além disso, a análise visou demonstrar o potencial das técnicas de modelagem estatística na previsão de preços, contribuindo para decisões mais informadas por parte de compradores, vendedores e investidores.

Material e Método

A base de dados utilizada neste estudo foi extraída da Ames Housing, que reúne informações detalhadas sobre imóveis vendidos na cidade de Ames, Iowa, nos Estados Unidos da América (EUA). O conjunto foi carregado no R por meio do pacote *modeldata*, que disponibiliza a base para análise. Inicialmente, o conjunto contava com 2.930 observações, cada uma representando um imóvel distinto. Para adequação aos objetivos da modelagem, foi realizado um pré-processamento: primeiramente, selecionaram-se apenas os imóveis classificados como "OneFam" e com zoneamento "Residential_Low_Density", o que resultou em 2.009 observações. Em seguida, algumas variáveis foram renomeadas para maior clareza e consistência — *Sale_Price* passou a ser *preco_venda*, *Gr_Liv_Area* tornou-se *area_construida*, *Year_Built* foi renomeada para *ano_construcao* e *Garage_Cars* para *vagas_garagem*. Posteriormente, a base foi dividida aleatoriamente em dois subconjuntos: um conjunto de treinamento, com 1.909 observações, utilizado para o ajuste do modelo, e um conjunto de teste, com 100 observações, reservado para a avaliação das previsões. As variáveis selecionadas para a modelagem foram o logaritmo do preço de venda, a metragem construída, o ano de construção e o número de vagas de garagem. A variável *ano_construcao* apresenta imóveis construídos entre 1875 e 2010, enquanto a

area_construida possui uma média de aproximadamente 1.543 pés quadrados. Já o número de *vagas_garagem* varia de 0 a 4, com mediana de 2 e uma média de cerca de 1,86 vaga por imóvel. Para a análise exploratória e a modelagem estatística, foi utilizado o software R (R CORE TEAM, 2024).

Resultados e Discussão

A Figura 1 apresenta as correlações amostrais de Pearson entre as variáveis da base de dados e a distribuição dos logaritmos dos preços de vendas das casas. Analisando as correlações, observa-se que o logaritmo do preço de venda apresenta forte associação com a área útil construída e moderada associação com as variáveis ano de construção e número de vagas de garagem.

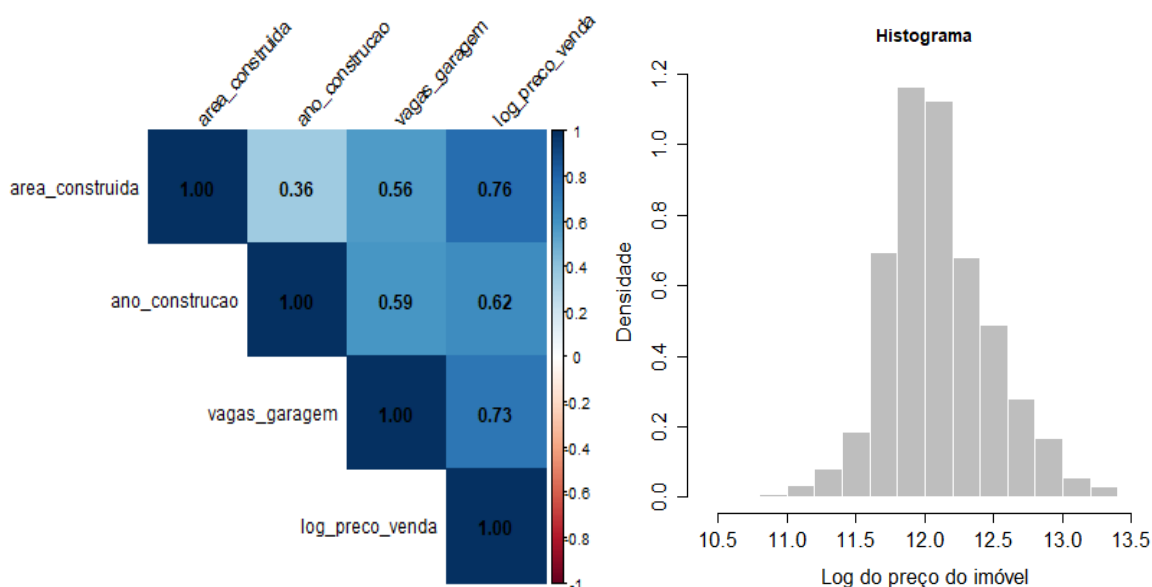


Figura 1: Correlações de Pearson entre as variáveis do conjunto de dados e distribuição dos logaritmos dos preços de venda dos imóveis.

Fonte: Elaboração Própria, 2025

Em seguida, modelou-se o logaritmo do preço de venda das casas com o objetivo de quantificar os efeitos das variáveis explicativas e verificar sua significância estatística conjuntamente. Para isso, ajustou-se um modelo de regressão linear múltipla com as três covariáveis. A variável área útil construída foi padronizada (subtraindo os valores observados da sua média e dividindo pelo desvio padrão).

Com base no Critério de Informação de Akaike (AIC) (Akaike, 1974) e nos testes t de Student aplicados aos coeficientes de regressão, concluiu-se que todas as variáveis são relevantes para o modelo, não havendo necessidade de descarte. Dessa forma, o modelo final pode ser descrito pela seguinte equação:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + e_i$$

sendo Y_i o logaritmo do preço de venda da i -ésima casa, β_0 o intercepto (que é o preço médio de venda quando todas as covariáveis forem nulas), os coeficientes β_1 , β_2 , β_3 são os efeitos das covariáveis `area_construida` (após a padronização), `ano_construcao` e `vagas_garagem` e, por fim, e_i são os erros aleatórios, assumidos como independentes e identicamente distribuídos segundo uma distribuição normal com média zero e variância constante σ^2 . A partir da estimação dos coeficientes, é possível entender o comportamento dos valores e prever o logaritmo do preço de venda de um imóvel com base nas características dos imóveis.

A Tabela 1 apresenta as estimativas pontuais dos coeficientes obtidas pelo método dos mínimos quadrados ordinários que, no modelo adotado, coincidem com as estimativas de máxima verossimilhança, bem como os respectivos intervalos de confiança (ICs) a 95%.

Observa-se que nenhum dos intervalos de confiança inclui o valor zero, o que confirma a significância estatística dos coeficientes. Dessa forma, rejeita-se a hipótese nula $H_0 : \beta_j = 0$ em favor da hipótese alternativa $H_1 : \beta_j \neq 0$ para $j = 0, 1, 2, 3$. Além disso, os p-valores calculados para esses testes foram todos inferiores a 2×10^{-16} , reforçando a evidência de que os coeficientes são estatisticamente significativos.

Tabela 1: Estimativas pontuais e intervalares dos coeficientes

Coeficientes	Estimativa
Intercepto	3,4962 (2,7119 ; 4,2804)
<code>area_construida</code>	0,1938 (0,1838 ; 0,2038)
<code>ano_construcao</code>	0,0042 (0,0038 ; 0,0046)
<code>vagas_garagem</code>	0,1452 (0,1290 ; 0,1615)

Fonte: Autores, 2025

A Figura 2 apresenta a comparação entre previsões (linha azul) e valores observados (pontos vermelhos) do logaritmo do preço de venda ao longo do tempo, com os intervalos de das previsões representados pela faixa sombreada azul claro. Observa-se que, apesar de variações nos dados reais, a série prevista acompanha razoavelmente bem a tendência geral, mantendo a maior parte dos valores observados dentro dos limites de incerteza. A visualização evidencia a capacidade do modelo de capturar os padrões, embora haja discrepâncias pontuais.

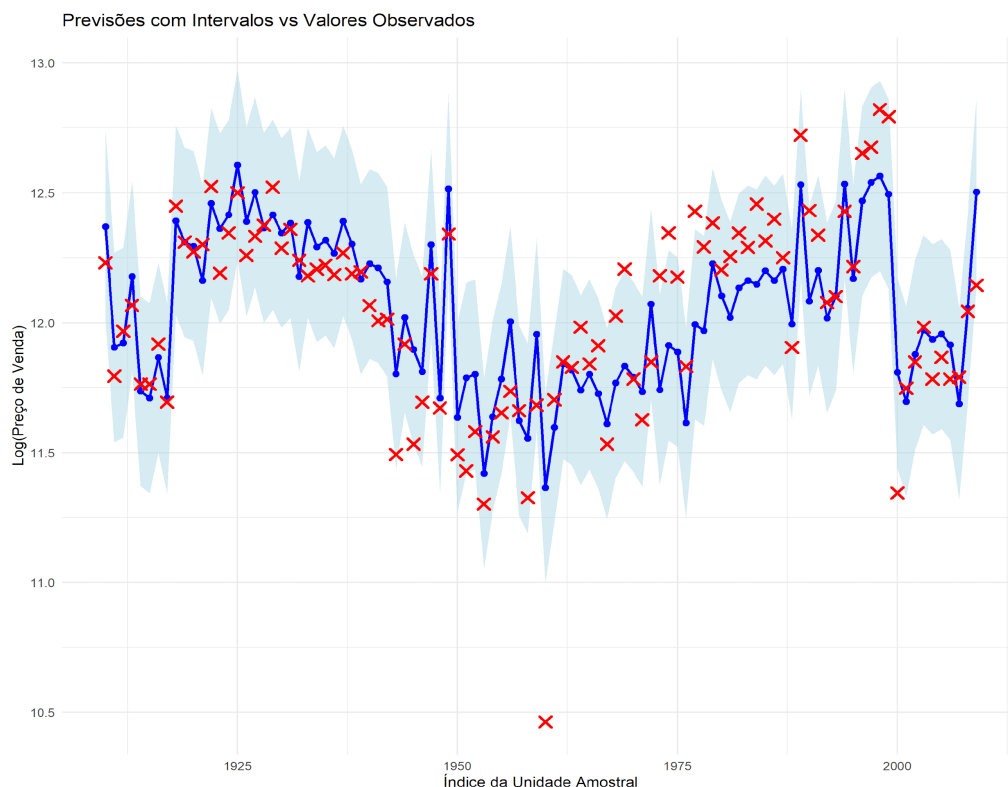


Figura 2: Previsões do logaritmo do preço de venda com respectivos intervalos de confiança (faixa azul clara) ao longo do índice das unidades amostrais. A linha azul representa os valores previstos pelo modelo, enquanto os pontos vermelhos indicam os valores observados.

Fonte: Elaboração Própria, 2025

Conclusão

Este estudo aplicou um modelo de Regressão Linear Múltipla para analisar os fatores que influenciam os preços das residências. Os resultados mostraram que variáveis como tamanho do imóvel, ano de construção e número de garagens são significativos no valor das propriedades, com destaque para o tamanho do imóvel.

A análise reforça a importância do uso de métodos estatísticos no mercado imobiliário, proporcionando insights valiosos para compradores, vendedores e investidores. Sugere-se que futuras pesquisas explorem técnicas mais avançadas e incluam variáveis econômicas e sociais, aprimorando ainda mais as previsões do mercado.

A análise reforça a importância do uso de métodos estatísticos no mercado imobiliário, proporcionando insights valiosos para compradores, vendedores e investidores. Apesar de o modelo apresentar um bom poder preditivo, as suposições de normalidade, homocedasticidade, independência e linearidade dos resíduos não foram completamente atendidas. Assim, sugere-se que futuras pesquisas explorem técnicas mais avançadas, além de considerar a inclusão de variáveis econômicas e sociais, com o objetivo de aprimorar ainda mais as previsões do mercado e a adequação do modelo.

Referências

AKAIKE, Hirotugu. **A new look at the statistical model identification**. IEEE Transactions on Automatic Control, v. 19, n. 6, p. 716–723, 1974.

CHEN, Ningyan. **House Price Prediction Model of Zhaoqing City Based on Correlation Analysis and Multiple Linear Regression Analysis**. Wireless Communications and Mobile Computing, 2022, 1, 9590704. Disponível em: <https://doi.org/10.1155/2022/9590704>.

MAO, Tingjun. **Real Estate Price Prediction Based on Linear Regression and Machine Learning Scenarios**. BCP Business & Management, 38, 400-409, 2023.

R CORE TEAM. **R: A Language and Environment for Statistical Computing**. R Foundation for Statistical Computing, Vienna, Austria, 2024. Disponível em <https://www.R-project.org>.

SANTOS, Rafael C.; SILVA, André F. de A.; PEREIRA, Roberto S. **Modelagem dos preços de imóveis residenciais no município de São Paulo utilizando regressão linear**. Revista Brasileira de Gestão Urbana, 2021, 13(2), 85-98.

Anexo

<https://rpubs.com/jvhonorato/1309267>