Recebido: 02/03/2012
Aprovado: 25/09/2012

Construção de um modelo para o preço de venda de casas residenciais na cidade de Sorocaba-SP

Júlio César Pereira (UFSCar-SP/Brasil) - julio.pereira.ufscar@gmail.com,
• Rodovia João Leme dos Santos (SP-264), Km 110, 18052-780, Sorocaba-SP
Salomão Garson (UFSCar-SP/Brasil) - salomaogars@hotmail.com
Elton Gean de Araújo (UFMS-MS/Brasil) - egarauj@yahoo.com.br

RESUMO O presente artigo foi desenvolvido com base em um estudo estatístico realizado no mercado imobiliário de Sorocaba, cidade localizada no interior de São Paulo. Neste estudo, objetivou-se a utilização de métodos de regressão linear múltipla na modelagem do preço de venda de casas da cidade de Sorocaba-SP com base em suas características. Visando obter maior qualidade na predição dos preços de venda, foram utilizados métodos de seleção de variáveis, que garantem a utilização apenas das variáveis relevantes ao problema, implicando em uma estimação mais fiel dos preços de venda. Além disso, uma das variáveis contidas era qualitativa, o que demandou o uso de variáveis dummy. Através dos métodos citados, chegou-se à conclusão de que algumas variáveis coletadas não deveriam fazer parte do modelo. Obteve-se assim, as variáveis importantes para a construção de um modelo de regressão linear múltipla adequado, que pode auxiliar de maneira eficiente na avaliação e estimação do preço de venda de imóveis situados em Sorocaba.

Palavras-chave Regressão Linear Múltipla; Seleção de Variáveis; Estimação do Preço de Venda de Casas.

ABSTRACT *This article is based on a survey about the real estate market in Sorocaba, a city in the interior of Sao Paulo. The main objective of this study was to use the multiple linear regression modeling method for gauging housing prices in Sorocaba. In order to obtain higher quality in the prediction of sales prices, various methods were used for selecting variables which can then guarantee the use of only using variables that are relevant to the problem, therefore resulting in a more accurate estimation of sale prices. In addition, a qualitative variable was included, which required the use of dummy variables. Using the methods mentioned we came to the conclusion that some of the variables collected should not be part of the model. Thus through obtaining the important variables for the construction of an appropriate multiple linear regression model, we can effectively assist in the evaluation and estimation of the sales prices of real estate located in Sorocaba.*

Keywords *Multiple Linear Regression; Selection of Variables; Estimation of the Sales Prices of Houses.*

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.

1. INTRODUÇÃO

O município de Sorocaba, localizado no interior do estado de São Paulo, tem passado por um grande desenvolvimento nas últimas décadas. Sendo considerado o terceiro município mais populoso do interior paulista e o quarto mercado consumidor do estado, com exceção da região metropolitana, Sorocaba recebe grandes investimentos nos mais diversos setores, como industrial e educacional (PORTAL SOROCABA, 2010). Consequentemente, a cidade tem tido um grande crescimento populacional, levando a uma movimentação estrondosa no mercado imobiliário, que segundo Steiner *et al.* (2007), é uma das áreas mais dinâmicas do setor terciário da economia e a maior dificuldade em desenvolver estudos acerca do mesmo é a grande heterogeneidade das características (atributos e variáveis) de cada imóvel, bem como as relações que elas podem guardar entre si.

Além da utilidade como moradia, uma unidade imobiliária é também um investimento financeiro, e a construção de um modelo para predição de seu preço 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).

Dada a grande quantidade de variáveis que podem ser utilizadas para explicar o preço de imóveis, é necessário que haja uma seleção do conjunto de variáveis independentes a ser usado no modelo. Algumas vezes, muitas das variáveis envolvidas não são importantes para modelar adequadamente o preço do imóvel. Nessas situações tem-se interesse em filtrar as variáveis candidatas para obter um modelo que contenha o melhor conjunto possível de variáveis regressoras que expliquem a variável preço (Y). Dessa forma, espera-se obter um modelo final que contenha variáveis regressoras suficientes, de modo a obter desempenho satisfatório do modelo na descrição, bem como previsão, da variável Y. Por outro lado, para manter os custos mínimos de manutenção e tornar um modelo de fácil utilização, é desejável que o modelo use o menor número possível de variáveis regressoras. Diante desse conflito entre usar uma quantidade suficiente de vaiáveis que descreva bem a variável Y e o menor número possível de variáveis para que o modelo seja de fácil interpretação, é necessária a utilização dos métodos de seleção de variáveis (MONTGOMERY; RUNGER, 2008).

Pelo intenso aquecimento do mercado imobiliário de Sorocaba e pela dificuldade de predição e avaliação dos preços de venda de imóveis, o presente artigo tem por objetivo propor um modelo de regressão linear múltipla que auxilie na estimação dos preços de venda de imóveis da cidade de Sorocaba, a partir de suas características físicas e de sua localização. Para isso, foi considerada uma amostra de casas residenciais à venda na cidade. Além disso, foram utilizados métodos de seleção de variáveis, com o intuito de se obter um modelo de regressão linear adequado, que contemple apenas as variáveis relevantes ao estudo em questão e que estime de maneira fiel os preços de venda de uma casa situada em Sorocaba.

Este artigo está estruturado da seguinte forma: na seção 2 é feita uma revisão de literatura, em que alguns estudos semelhantes ao assunto abordado são apresentados, de forma a ilustrar que tipos de modelos são desenvolvidos para avaliação do mercado imobiliário; a seção 3 apresenta os materiais e métodos utilizados; na seção 4 são apresentados os resultados obtidos e discussões. Nas seções 5 e 6 são apresentadas as considerações finais e as referências bibliográficas, respectivamente.

2. REVISÃO BIBLIOGRÁFICA

Em estudos do mercado imobiliário, é comum utilizar modelos de regressão linear múltipla, a fim de analisar uma variável de interesse (Y) em função de diversas outras variáveis (x_j). Por exemplo, Nadal *et al.* (2003) fez uso de uma amostra de 20 imóveis na cidade de Curitiba para o desenvolvimento de um modelo de regressão linear múltipla que auxiliasse na predição do preço de venda de um imóvel da cidade que havia sido desapropriado por fatores ambientais e turísticos. Foram utilizadas as variáveis idade aparente do imóvel, área equivalente, padrão da construção, número de vagas na garagem e preço de venda do imóvel para chegar ao modelo de regressão linear, através do método dos mínimos quadrados. Além disso, foram realizados alguns testes para validação do modelo adotado, com o intuito de garantir a qualidade do modelo. Este é apenas um exemplo, entre diversos outros que empregam modelos de regressão para a modelagem do preço de venda de imóveis, como é o caso de Steiner *et al.* (2007), Rozenbaum e Macedo-Soares (2007), Gazola (2002), Alves (2005), Couto (2007) e Braulio (2005).

A regressão linear é um método estatístico que estabelece uma relação entre uma variável resposta Y e outras variáveis independentes x. A regressão linear simples considera um único regressor ou preditor x e uma variável dependente Y, enquanto a regressão linear múltipla relaciona Y com n outras variáveis, como apresentado a seguir.

Considerando-se k regressores o modelo de regressão linear múltipla pode ser expresso pela Equação 1, podendo ser escrito na forma reduzida, como na Equação 2 (MONTGOMERY; RUNGER, 2008).

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \varepsilon \tag{1}$$

$$y = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} + \varepsilon_i \tag{2}$$

Em que, β_j, $j = 0, 1, \ldots, k$ são os coeficientes de regressão, sendo parâmetros que representam a variação esperada em y por unidade de variação em x_j quando todos os outros regressores são mantidos constantes. Além disso, a regressão linear múltipla também pode ser trabalhada na forma matricial, como ilustra a Equação 3, em que n representa o número de observações utilizadas na amostra (MONTGOMERY; RUNGER, 2008).

$$\begin{bmatrix} n & \sum_{i=1}^{n} x_{i1} & \sum_{i=1}^{n} x_{i2} & \cdots & \sum_{i=1}^{n} x_{ik} \\ \sum_{i=1}^{n} x_{i1} & \sum_{i=1}^{n} x_{i1}^2 & \sum_{i=1}^{n} x_{i1}x_{i2} & \cdots & \sum_{i=1}^{n} x_{i1}x_{ik} \\ \vdots & \vdots & \vdots & & \vdots \\ \sum_{i=1}^{n} x_{ik} & \sum_{i=1}^{n} x_{ik}x_{i1} & \sum_{i=1}^{n} x_{i1}x_{i2} & \cdots & \sum_{i=1}^{n} x_{ik}^2 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} y_1 \\ \sum_{i=1}^{n} x_{i1}y_1 \\ \vdots \\ \sum_{i=1}^{n} x_{ik}y_1 \end{bmatrix} \tag{3}$$

Esse modelo é utilizado com o objetivo de entender como *Y* se comporta após uma mudança em uma ou mais variáveis independentes. Dessa forma, é possível fazer inferências sobre a variável resposta, tais como realizar predições de seu comportamento e obter estimativas por intervalo (CHARNET *et al*., 1999; MONTGOMERY; RUNGER, 2008).

Dentre outros trabalhos que empregam técnicas de regressão linear aplicadas ao mercado imobiliário, pode-se citar Steiner *et al*. (2007), que realizaram um estudo imobiliário na cidade de Campo Mourão, no Paraná, utilizando a análise de agrupamento para criar grupos de imóveis com características mais homogêneas, para posteriormente utilizar a regressão linear múltipla, com o intuito de estimar preços de imóveis que seriam colocados à venda. Gazola (2002) coletou dados referentes a apartamentos da cidade de Criciúma, no estado de Santa Catarina, visando à estimação de preços de outros apartamentos a partir de suas características por meio de uma regressão linear múltipla. As variáveis utilizadas foram a área total do imóvel, consumo de energia, distância à escola, acessibilidade, idade do imóvel, dormitórios, meio ambiente, região homogênea, zona fiscal, padrão de entrada, classificação, conservação, garagem, suíte, dependência de empregada, elevador e pólos de valorização. Além disso, utilizou-se a técnica de *Ridge Regression*, que evita a multicolinearidade, que ocorre quando as variáveis independentes de uma regressão possuem relações lineares exatas ou aproximadamente exatas.

No mesmo âmbito, Alves (2005) avaliou preços de imóveis na cidade de Campo Mourão, no estado do Paraná, através de modelos de regressão linear e com o auxílio de um programa computacional denominado AMI (Análise Multivariada de Imóveis), desenvolvido através da linguagem computacional MATLAB. A *interface* do programa oferece como opções três diferentes tipos de regressões e fornece o resultado de maneira imediata ao usuário. Além disso, Couto (2007) realizou um estudo imobiliário na cidade de Porto, em Portugal, através de ferramentas estatísticas, dentre elas, modelos de regressão linear múltipla, tendo sido seu trabalho voltado principalmente para imóveis destinados à habitação e com maior concentração na tributação imobiliária.

Braulio (2005) desenvolveu um modelo através de métodos estatísticos multivariados para avaliar imóveis em função de suas principais características na cidade de Campo Mourão, assim como Alves (2005). Porém, para apurar os dados coletados e garantir a confiabilidade do modelo final, foram utilizadas técnicas de análise multivariada, como análise de agrupamento e diversos testes de seleção de variáveis, como a análise de todas as regressões possíveis, o teste *Stepwise* (passo a passo), Seleção *forward* e Eliminação *backward*. O resultado obtido foi um modelo de regressão linear múltipla de alto nível de precisão no que diz respeito à predição de preços de casas, apartamentos e terrenos da cidade de Campo Mourão. Já no estudo de Rozenbaun e Macedo-Soares (2007), foi construído um índice de preço de imóveis através de um modelo de regressão linear múltipla. Estes autores, citando Sirmans *et al*. (2005), elencam as seguintes variáveis entre as mais presentes nos estudos de avaliação do preço de imóveis, sendo elas: área privativa, número de quartos, localização, amenidades e idade do imóvel. Porém, a diferença do trabalho de Rozenbaun e Macedo-Soares (2007) com o desenvolvido por Braulio (2005) é que no primeiro não foram utilizados métodos de seleção de variáveis formais, mas estas foram selecionadas de maneira subjetiva através de um modelo hedônico, que permite analisar a importância relativa a cada variável.

3. MATERIAL E MÉTODO DA PESQUISA

3.1. Material

Foi coletada uma amostra de 150 observações a partir dos *websites* de diversas imobiliárias da cidade de Sorocaba-SP. Foram consideradas as informações disponíveis nos *sites*, sendo as variáveis disponíveis: o preço de venda da casa (Y), quantidade de dormitórios (x_1), área construída em m² (x_2), área do terreno em m² (x_3), número de vagas na garagem (x_4) e a localização do imóvel indicada pelo bairro (d), sendo que a amostra obtida cobriu todas as regiões da cidade.

3.2. Métodos

Os dados mencionados na seção anterior foram tratados utilizando-se técnicas de regressão linear múltipla, na qual se procurou construir um modelo para o preço de casas em função das demais características disponíveis. E a fim de elencar as variáveis realmente importantes na formação do preço, foram empregados os métodos de seleção de variáveis. No que se segue são apresentados os métodos utilizados.

3.2.1. Seleção de Variáveis para a Modelagem do Preço

Com o intuito de selecionar as variáveis relevantes para a construção de um modelo relacionando o preço de venda das casas residenciais e suas características disponíveis, foram utilizados alguns métodos de seleção de variáveis.

Muitas vezes, nem todas as variáveis ou regressores são relevantes para o modelo, nesse caso, é necessário obter um subconjunto de variáveis que contenha apenas as que influenciam no sentido de melhorar o 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. Porém, quanto maior for o número de regressores, maior é o gasto de recursos para se trabalhar com o modelo, como o custo de manutenção e a dificuldade de utilização do modelo. Assim, a seleção de variáveis na verdade é um problema de otimização, em que o objetivo é encontrar o melhor subconjunto de regressores que gere um modelo fiel (MONTGOMERY; RUNGER, 2008). Os métodos de seleção de variáveis utilizados são citados a seguir.

3.2.1.1. Todas as regressões possíveis

Nessa abordagem, para se encontrar o melhor modelo, foram testadas todas as equações de regressão possíveis, considerando as quatro variáveis quantitativas disponíveis. Ou seja, foram ajustadas todas as equações existentes com apenas uma das variáveis candidatas, todas existentes com duas e assim por diante, obtendo um total de 2^4 equações diferentes. Assim, todas as equações foram avaliadas de acordo com alguns critérios apresentados a seguir, visando encontrar o melhor modelo.

Um dos critérios utilizados para analisar e comparar as equações é o coeficiente de determinação múltipla (R_p^2), representado pela Equação 4.

$$R_p^2 = \frac{SQ_R\,(p)}{SQ_T} = 1 - \frac{SQ_E\,(p)}{SQ_T} \tag{4}$$

Em que $SQ_R\,(p)$ é a soma quadrática da regressão, $SQ_E\,(p)$ é a soma quadrática dos erros e SQ_t é a soma quadrática total, para um modelo com p variáveis. Conforme p aumenta, ocorre também um aumento em R_p. Assim, adicionam-se variáveis ao modelo até o ponto que é visível que o aumento em R_p é praticamente desprezível. Essa técnica é importante, pois mostra que existem modelos de regressão bons com números diferentes de regressores. Existe um momento em que se aumenta o número de regressores e a qualidade do modelo aumenta pouquíssimo, o que gera maior gasto de recursos pelo maior número de regressores.

Outro critério utilizado é o quadrado médio do erro, dado pela Equação 5.

$$MQ_E\,(p) = \frac{SQ_E\,(p)}{(\mathrm{n}-p)} \tag{5}$$

Normalmente ocorre uma diminuição no $MQ_E\,(p)$ quando p aumenta. Escolhe-se o mínimo $MQ_E\,(p)$, pois a média quadrática devido ao erro seria menor, não prejudicando a qualidade do modelo.

Um terceiro critério utilizado foi a média quadrática total do erro, C_p, para o modelo de regressão. Essa medida é definida através da Equação 6.

$$C_p = \frac{SQ_E\,(p)}{\hat{\sigma}^2} = n + 2p \tag{6}$$

As equações que possuem tendenciosidade negligenciável têm valores de C_p próximos de p, enquanto aquelas com tendenciosidades significantes terão C_p relativamente maiores do que p. Obviamente, o modelo escolhido é o que possui a média quadrática do erro mais próxima do valor de p, pois é o que possui menor tendenciosidade.

Outro critério empregado, apresentado na Equação 7, é o chamado R_p^2 ajustado, que é basicamente uma modificação em R_p^2 que considera o número de variáveis no modelo.

$$\overline{R}_p^2 = 1 - \frac{(n-1)}{(n-p)} = (1 - R_p^2) \tag{7}$$

Percebe-se que $\overline{R}_p^2$ decresce à medida que p aumenta, se a diminuição de $(n-1)(1-R_p^2)$ não for compensada pela perda de um grau de liberdade $n - p$. Além disso, o modelo selecionado é o de valor máximo do $\overline{R}_p^2$, que na verdade é o mesmo que selecionar o valor mínimo de $MQ_E\,(p)$ (MONTGOMERY; RUNGER, 2008).

3.2.1.2. Regressão por etapas

Segundo Charnet *et al*. (1999), o método de regressão por etapas é o mais utilizado na seleção de variáveis de modelos de regressão, em que é construída uma sequência de modelos adicionando ou removendo variáveis, em cada etapa. Os três métodos de regressão por etapas foram utilizados, sendo o "Passo atrás", "Passo a frente" e "Passo a Passo". No primeiro caso, inicialmente foram utilizadas todas as variáveis do modelo e feitos testes de significância (Teste F) por etapas, sendo que a cada etapa, uma variável poderia ser eliminada. A partir do momento em que não foi eliminada nenhuma variável, as variáveis que restaram no processo foram as selecionadas. No método "Passo a frente", iniciaram-se os testes com apenas uma variável, a de maior coeficiente de correlação amostral com a variável resposta *y*. Assim, foram realizados os testes, em que a cada etapa, poderia ser adicionada uma variável. Da mesma forma que no método "Passo atrás", no momento em que nenhuma variável foi adicionada, o teste foi interrompido e foram utilizadas no modelo final as variáveis que restaram no conjunto.

O método "Passo a Passo" é semelhante ao "Passo a frente", possuindo como diferença, o fato de que em cada etapa, alguma variável poderia ser descartada também. Ou seja, neste método, variáveis foram adicionadas e descartadas a cada etapa. Obtivemos o subconjunto de variáveis selecionado para utilização no modelo final quando nenhuma variável foi incluída ou excluída do modelo (CHARNET *et al*., 1999).

3.2.2. Variáveis *Dummy*

No presente estudo, foram utilizadas variáveis quantitativas, como quantidade de vagas na garagem, número de quartos, área útil e área do terreno. Porém, como no caso da utilização da variável bairro, muitas vezes existe a necessidade de utilizar variáveis não numéricas, chamadas de variáveis qualitativas e conhecidas na econometria como variáveis *Dummy*, que são binárias. Assim, em casos como informações sobre gênero (masculino ou feminino), pessoas que possuem ensino superior ou não, empresas que disponibilizam determinado serviço ou não, as variáveis *Dummy* são utilizadas (WOOLDRIDGE, 2010).

Geralmente essas variáveis possuem valor 1 para uma das opções e zero para a outra, como ilustra o exemplo na Tabela 1.

Tabela 1 – Exemplo de utilização de variáveis *Dummy* para variáveis com 2 níveis.

Bairro de Sorocaba	Variável *Dummy*
Campolim	1
Não Campolim	0

Fonte: Dados da pesquisa.

Segundo Montgomery (2008), quando a variável qualitativa possui mais de dois valores, é necessária utilização de mais de uma variável *Dummy*, sendo que uma variável com *t* níveis pode ser modelada com *t* – 1 variáveis indicativas, como ilustra o exemplo na Tabela 2.

Tabela 2 – Exemplo da utilização de variáveis *Dummy* para variáveis com mais de 2 níveis.

Bairros de Sorocaba	Variável *Dummy* 1	Variável *Dummy* 2	Variável *Dummy* 3
Campolim	1	0	0
Jardim Vera Cruz	0	1	0
Jardim São Paulo	0	0	1
Centro	0	0	0

Fonte: Dados da pesquisa.

A princípio, no presente trabalho, os preços dos imóveis foram avaliados em função das variáveis quantitativas, tendo sido a variável bairro trabalhada posteriormente, por ser uma variável *Dummy* e demandar tratamento estatístico diferenciado.

Para todos os tratamentos estatísticos realizados no presente artigo, foi utilizado o *software* R Development Core Team (2010).

4. ANÁLISE DE DADOS E RESULTADOS

Antes de se iniciar o processo de ajuste de modelos e seleção de variáveis foi realizada uma análise da correlação entre as variáveis quantitativas, candidatas a regressores, x_1 (número de dormitórios), x_2 (área do terreno), x_3 (área construída), x_4 (número de vagas na garagem) e a variável resposta Y (preço do imóvel). O resultado obtido está expresso na Tabela 3.

Tabela 3 – Correlação entre as variáveis independentes quantitativas e a variável dependente.

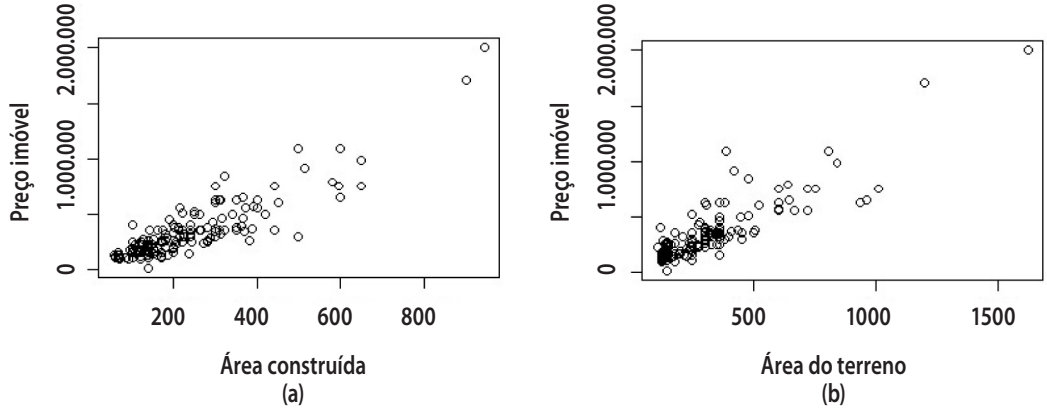
	x_1	x_2	x_3	x_4	Y
x_1	-	0,48	0,35	0,17	0,39
x_2	0,48	-	0,84	0,45	0,88
x_3	0,35	0,84	-	0,51	0,86
x_4	0,17	0,45	0,51	-	0,48
Y	0,39	0,88	0,86	0,48	-

Fonte: Dados da pesquisa.

Percebe-se que as variáveis x_2, x_3 apresentam fortes correlações com a variável y. Isso significa que a relação entre a área do terreno e área construída das casas de Sorocaba é mais próxima de uma relação linear com o preço de venda, e havendo um crescimento de uma dessas variáveis, o valor do preço da casa acompanhará esse crescimento.

A Figura 1 corrobora com os resultados obtidos na Tabela 3, em que se nota forte associação entre as variáveis área do terreno e área construída das casas com o preço de venda. Porém, observa-se ainda que, a partir de uma certa medida do terreno (em torno de 400m^2), os pontos ficam um pouco mais dispersos (Figura 1b), enquanto que para a área construída embora a dispersão também aumente com o aumento da área, os pontos ainda se concentram mais próximos de uma reta. Esses resultados indicam que a partir de um determinado ponto, a área do terreno já não é tão determinante para o preço quanto a área construída.

Figura 1 – Diagrama de dispersão entre preço e área construída da casa (a) e Diagrama de dispersão entre preço e área do terreno da casa (b).



Fonte: Dados da pesquisa.

4.1. Seleção das Variáveis

Inicialmente foi utilizado o método "Todas as regressões possíveis", sendo ajustados todos os possíveis modelos de regressão, considerando as variáveis independentes (x_1, x_2, x_3 e x_4). Os modelos candidatos foram comparados através de alguns indicadores, como coeficiente de determinação múltipla (R^2), coeficiente de determinação múltipla ajustado ($\overline{R}_p^2$), quadrado médio do erro (MQ_E) e média quadrática total do erro (C_p). A Tabela 4 apresenta os resultados dos indicadores para cada modelo de regressão linear. As linhas em negrito representam os melhores indicadores a cada grupo de equações com as mesmas quantidades de variáveis.

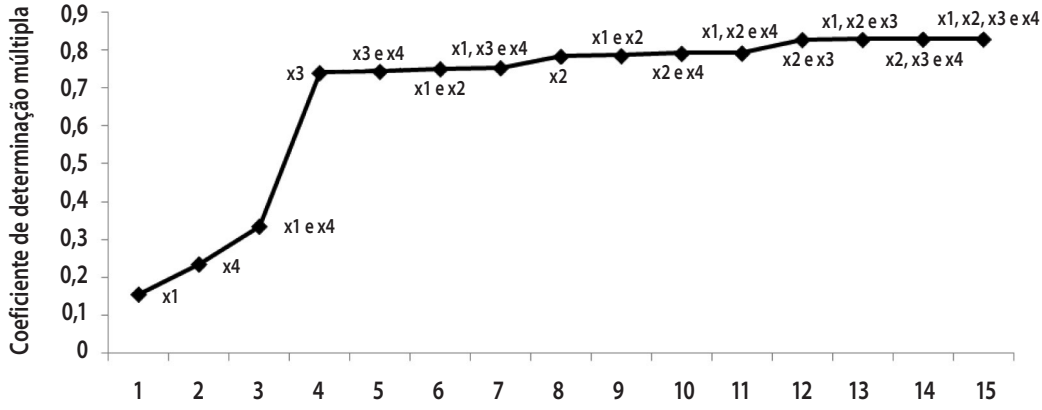
Tabela 4 – Cálculo dos indicadores para seleção de variáveis de cada modelo candidato.

Variáveis usadas na equação candidata	R^2	R^2 ajustado	MQ_E	C_p	$C_p - p$
x_1	0,1550	0.1268	7,0756E+10	469,8126	467,8126
x_2	**0,7823**	**0.7751**	**1,8223E+10**	**31,2468**	**29,2468**
x_3	0,7393	0.7306	2,1823E+10	60,5690	58,5690
x_4	0,2343	0.2088	6,4114E+10	414,3482	412,3482
x_1 e x_2	0,7843	0.7771	1,8058E+10	31,9109	28,9109
x_1 e x_3	0,7497	0.7413	2,0958E+10	56,0978	53,0978
x_1 e x_4	0,3341	0.3119	5,5754E+10	346,5429	343,5429
x_2 e x_3	**0,8268**	**0.8211**	**1,4495E+10**	**2,0181**	**-1.181**
x_2 e x_4	0,7902	0.7832	1,7562E+10	27,6176	24,6176
x_3 e x_4	0,7429	0.7343	2,1527E+10	60,7042	57,7042
x_1, x_2 e x_3	0,8271	0.8214	1,447E+10	3,6163	-1.163
x_1, x_2 e x_4	0,7911	0.7841	1,7491E+10	28,6226	24,6226
x_1, x_3 e x_4	0,7523	0.7441	2,0734E+10	56,1425	54,1425
x_2, x_3 e x_4	**0,8281**	**0.8224**	**1,439E+10**	**3,1344**	**-1.344**
x_1, x_2, x_3 e x_4	**0,8283**	**0.8226**	**1,4371E+10**	**4,9605**	**-1.605**

Fonte: Dados da pesquisa.

A última linha da Tabela 4, representada pela equação que possui todas as variáveis, apresentou os melhores indicadores. Porém, como mostra o gráfico do comportamento do coeficiente de correlação múltipla, ilustrado na Figura 2, existe um ponto em que a adição de variáveis ao modelo não gera grandes diferenças nos indicadores. A equação que representa este ponto é a que possui apenas as variáveis x_2 e x_3, sugerindo que talvez não seja vantajoso incluir as variáveis x_1 e x_4 no modelo, fato que não implicaria em melhora significativa na qualidade do mesmo.

Figura 2 – Comportamento do coeficiente de determinação múltipla para cada equação.



Fonte: Dados da pesquisa.

Também foram aplicados os procedimentos de seleção de variáveis "Passo a Passo", "Passo a Frente" e "Passo Atrás". Esses procedimentos apresentaram os mesmos resultados que o método "Todas as Regressões". Dessa forma, na Tabela 5 são apresentados os resultados apenas do método "Passo a Passo" divididos por etapas. As linhas em negrito na Tabela 5 indicam as variáveis que devem compor o modelo a cada etapa.

Assim como no método "Todas as Regressões", observa-se na Tabela 5 que o método "Passo a Passo" indica não haver necessidade de se utilizar um modelo com mais que duas variáveis regressoras dentre as variáveis testadas.

Tabela 5 – Cálculo dos indicadores para seleção de variáveis de cada modelo candidato usando o método "Passo a Passo".

Etapa	Variável presente na Equação	Teste F	p-valor (Resultado)
1	x_1	22,564	5.56e-06 ***
	x_2	**441,87**	**2.2e-16** ***
	x_3	350,65	2.2e-16 ***
	x_4	37,645	1.069e-08 ***
2	x_2 e x_1	1,079	Não Significativo
	x_2 e x_3	**31,488**	**1.278e-07** ***
	x_2 e x_4	4,6869	0.03234 *
3	x_2, x_3 e x_1	0,1984	Não Significativo
	x_2, x_3 e x_4	0,89	Não Significativo

Fonte: Dados da pesquisa.

Os resultados dos métodos de seleção de variáveis aplicados indicam a necessidade de se trabalhar apenas com as variáveis x_2 e x_3, o que resultou em um modelo expresso através da Equação 8. Neste modelo, observa-se que a cada unidade aumentada na área útil das casas (x_2), o valor do preço terá um acréscimo de R$ 989,40, considerando a área do terreno fixa, ou seja, a área construída é um dos atributos principais a serem considerados pelos investidores imobiliários, sendo este mais relevante do que o tamanho do terreno.

$$y = -34792{,}9 + 989{,}4x_2 + 43{,}4x_3 \quad (8)$$

O fato das variáveis x_1 e x_4 (número de dormitórios e número de vagas na garagem) terem sido eliminadas do modelo de regressão linear pode parecer não fazer sentido, pelo fato da estimação do preço de venda de uma casa ser baseada apenas na área útil e na área do terreno. No entanto, é perceptível o motivo que gerou este fato, sendo que na maioria das vezes, quanto maior a área útil de uma casa, maior o número de dormitórios presentes nela e quanto maior a área de um terreno, maior a possível quantidade de vagas na garagem; ou seja, para um terreno com área grande, mesmo que a quantidade de vagas anunciadas não seja grande, espera-se que seja possível disponibilizar espaço para esse fim. Assim, as variáveis x_1 e x_4, de certo modo, são explicadas pelas variáveis x_2 e x_3, o que torna irrelevante a presença delas no modelo.

4.2. Variável Bairro

Na amostra coletada, existem casas de 23 bairros diferentes, havendo a necessidade da utilização de 22 variáveis *Dummy*, fato que prejudica a eficiência do modelo, visto que o objetivo do estudo é desenvolver um modelo de regressão linear com o melhor desempenho possível e utilizando apenas variáveis relevantes. Dessa forma, foi realizada uma análise da diferença existente entre cada bairro, auxiliando no agrupamento dos bairros de características semelhantes. Para isso, consideraram-se as variáveis selecionadas anteriormente, x_2 e x_3, e ajustou-se um modelo incluindo as 22 variáveis *Dummy*, representando os 23 bairros. Esse modelo foi ajustado diversas vezes, tomando como base um bairro diferente em cada ajuste, ou seja, a cada ajuste um bairro diferente era representado por todas as variáveis *Dummy* assumindo valores iguais a zero. Com isso foi possível verificar, através do teste de significância dos coeficientes de regressão, quais bairros eram diferentes do bairro base. Os bairros que não apresentaram diferenças significativas foram colocados no mesmo grupo.

Através do teste de significância dos coeficientes de regressão, bairros que possuem casas de características semelhantes foram agrupados. Dessa forma, foram criados 5 grupos que incluem os 23 bairros coletados na amostra, como ilustra a Tabela 6. Cada grupo formado representa, de certa forma, a similaridade social e de infra-estrutura dos bairros que compõem o grupo. O grupo 1, por exemplo, contém os bairros nobres, onde reside a elite da cidade e onde se concentram *shopping centers* e *hipermercados*. No grupo 2, pode-se dizer que se concentram bairros populares, no grupo 3 bairros de classe média, no grupo 4 bairros antigos e bem localizados, enquanto que no grupo 5 bairros localizados nos extremos da cidade e distantes do centro.

Tabela 6 – Grupos criados no agrupamento dos bairros.

Grupo 1	Grupo 2	Grupo 3	Grupo 4	Grupo 5
Campolim	Nova Manchester	Pq Vitória Régia	Centro	Cajuru
Jd dos Estados	Jd Santo André	Éden	Vila Santana	São Bento
Trujilo	Pq das Laranjeiras	Jd Casa Branca	Jd Vera Cruz	Vila Fiori
Santa Rosália		Simus		Jd São Paulo
		Jd Europa		Jd Tatiana
		São Conrado		Jd Bertanha
		Jd Guaíba		

Fonte: Dados da pesquisa.

Assim, o número de variáveis *Dummy* no modelo foi reduzido de 22 variáveis para apenas 4. Dessa forma, para realizar a estimação do valor de uma casa através do modelo final, deve-se utilizar a Tabela 6 para saber em qual grupo a casa em questão está inserida.

4.3. Modelo Final

Após a realização dos diversos testes para seleção de variáveis que explicam a variável y (preço do imóvel), chegou-se a um conjunto de regressores composto pelas variáveis x_2 e x_3, que representam área construída e área do terreno respectivamente, além das variáveis *Dummy* d_1, d_2, d_3 e d_4, que representam os 5 grupos de bairros. O modelo de regressão final é apresentado através da Equação 9.

$$y = 29020{,}7 + 887{,}7x_2 + 451{,}9x_3 + 88051{,}8d_1 - 66601{,}1d_2 - 360261d_3 - 32951{,}7d_4 - \varepsilon \qquad (9)$$

Para utilização do modelo, devem-se substituir as variáveis x_2, x_3 na equação pela área construída e área do terreno da casa que se deseja estimar o preço de venda. Para as variáveis *Dummy*, é preciso enquadrar o bairro do imóvel em um dos 5 grupos expostos na Tabela 6. Assim, deve-se utilizar a Tabela 7 para atribuir os valores às 4 variáveis, dependendo do grupo que está sendo considerado. Por exemplo, no caso da necessidade de se estimar o preço de venda de uma casa no bairro Jardim Europa, percebe-se através da Tabela 6 que ele pertence ao grupo 3. Portanto, de acordo com a Tabela 7, a variável d_3 deve ser igual a 1 e todas as outras variáveis *Dummy* devem ser nulas.

Tabela 7 – Valores das variáveis *Dummy* para cada grupo de bairros.

Grupo	d1	d2	d3	d4
1	1	0	0	0
2	0	1	0	0
3	0	0	1	0
4	0	0	0	1
5	0	0	0	0

Fonte: Dados da pesquisa.

4.4. Validação Preditiva

A fim de realizar um teste de validação acerca da qualidade do modelo de regressão linear obtido, foram selecionadas ao acaso 10 casas da amostra utilizada, com o intuito de empregar o modelo para calcular seus preços de venda a partir de suas áreas construídas, áreas do terreno e bairros. O resultado obtido encontra-se na Tabela 8.

Tabela 8 – Utilização do modelo de regressão na previsão de preços de casas contidas na amostra utilizada.

Bairro	x_2	x_3	Preço observado	Preço estimado	Erro (%)
Campolim	202	325	R$ 380.000,00	R$ 443.255,40	16,65%
Campolim	341	360	R$ 500.000,00	R$ 582.462,20	16,50%
Trujilo	515	420	R$ 920.000,00	R$ 764.036,00	-16,95%
Santa Rosália	364	644	R$ 650.000,00	R$ 721.218,90	10,95%
Pq das Larenjeiras	160	250	R$ 190.000,00	R$ 217.426,60	14,43%
Jd Europa	200	152	R$ 260.000,00	R$ 239.223,40	-8,00%
Pq São Bento	70	144	R$ 130.000,00	R$ 156.233,30	20,17%
Trujilo	240	300	R$ 300.000,00	R$ 311.037,60	3,67%
Trujilo	213	664	R$ 550.000,00	R$ 606.214,20	10,22%
Pq Vitória Régia	69	125	R$ 100.000,00	R$ 110.733,40	19,73%
				Erro absoluto médio	13,73%

Fonte: Dados da pesquisa.

Como se pode perceber, existe uma diferença entre os preços de venda utilizados no mercado e os estimados através do modelo de regressão. No bairro Campolim, por exemplo, observou-se erros de previsão de aproximadamente 16,5% para as duas casas sorteadas. Porém, isso não implica em um padrão de erros aproximadamente iguais dentro dos bairros. Os valores dos erros de previsão nesse bairro poderiam ser diferentes caso tivessem sido sorteadas outras casas para a previsão. A exemplo disso, observa-se, por exemplo, o bairro Trujilo, em que o erro de previsão apresentou uma variação de -16,95% a 10,22%. Essa variação possivelmente poderia ser explicada por variáveis não medidas, como a localização do imóvel dentro do bairro (proximidade com o *shopping center* ou com a universidade localizados no bairro).

De uma forma geral, o erro absoluto médio de previsão foi estimado em cerca de 13,70%. Esta diferença pode ser causada pela falta de algumas variáveis que poderiam agregar qualidade ao modelo, como idade, estado de conservação do imóvel e proximidade com centros comerciais e escolas. Dessa forma, é possível que a falta de dados gere algumas imperfeições no modelo de regressão final.

Estes resultados indicam que o modelo proposto pode contribuir na avaliação do preço de casas da cidade de Sorocaba, visto que não ocorreu nenhuma diferença maior do que 20,17% entre o preço estimado e o preço observado. Os resultados, no entanto, mostram também a necessidade de se disponibilizarem mais informações a respeito do imóvel, tais como idade, estado de conservação, entre outras, como já citado anteriormente, a fim de se avaliar de maneira mais eficiente o imóvel e consequentemente estimar o preço de forma mais precisa.

5. CONSIDERAÇÕES FINAIS

A partir do modelo desenvolvido para estimação de preços de venda de casas na cidade de Sorocaba, percebe-se a grande importância da prática de testes de seleção de variáveis antes da construção efetiva de um modelo. Isso se deve ao fato de que em alguns casos, a utilização de uma variável a mais no modelo traz um aumento de qualidade tão ínfimo no resultado final, que não justifica a utilização da variável em questão, pelo simples fato do objetivo principal, ao se trabalhar com regressão linear múltipla, ser a busca por um modelo eficaz e eficiente, ou seja, um modelo que cumpra com os objetivos propostos e ao mesmo tempo seja desenvolvido com utilização mínima de recursos.

Outra vantagem dos testes de seleção de variáveis, que pode ser vista no presente estudo, ocorre quando são identificadas variáveis que não precisam fazer parte do modelo, pois são explicadas por alguma outra variável já presente. É o caso das variáveis "número de quartos" e "número de vagas na garagem", que foram retiradas do modelo, pois eram desnecessárias, dado que as variáveis "área útil" e "área do terreno" estavam no modelo. A aplicação dos diversos testes de seleção de variáveis indicou a irrelevância das variáveis que foram descartadas para o estudo proposto, porém, é importante haver uma interpretação do motivo pelo qual as variáveis estatisticamente não devem compor o modelo, o que não ocorre comumente. No presente caso, entende-se que em geral, uma casa com grande área construída possui maior número de dormitórios do que uma de menor área construída. Do mesmo modo, quanto maior a área do terreno de uma casa, maior a possibilidade desta possuir mais vagas na garagem. Assim, resultados estatísticos são interpretados de maneira mais eficiente e torna-se mais fácil a exposição dos mesmos para público em geral.

Com relação ao modelo final, é nítida a contribuição que este pode oferecer no mercado imobiliário, pelo fato de o preço da casa ser estimado com base em suas características e ser semelhante ao preço de venda das casas dos bairros de características semelhantes. Desta forma, situações de casas superestimadas ou subestimadas seriam evitadas e, ao mesmo tempo, ocorreria grande facilidade da geração do preço do imóvel demandando pouquíssimo tempo e recursos, bastando apenas a utilização de um computador.

Como sugestão para futuros estudos, seria interessante a utilização de uma gama maior de variáveis relativas às casas, de modo a buscar um modelo de regressão ainda mais realista. Seria, também, interessante cogitar a possibilidade do desenvolvimento de um modelo de regressão linear que contemplasse não só casas, mas apartamentos e imóveis comercias, o que ampliaria a gama de utilização do modelo e traria muitos benefícios para o mercado imobiliário, podendo inclusive ser utilizado em imobiliárias na estimação dos preços de imóveis à venda. Além disso, pelo fato do grande crescimento da população universitária da cidade de Sorocaba, seria de suma importância a criação de um modelo que envolvesse preços de aluguéis de imóveis na cidade, os quais vêm tendo demanda crescente nos últimos anos, e que com o crescimento do ensino universitário público na cidade, tende a crescer ainda mais.

REFERÊNCIAS BIBLIOGRÁFICAS

ALVES, V. **Avaliação de imóveis urbanos baseada em métodos estatísticos multivariados**. Dissertação de Mestrado – Programa de Pós Graduação em Métodos Numéricos em Engenharia, Universidade Federal do Paraná, UFPR, Campo Mourão, PR, 2005.

BRAULIO, S. N. **Proposta de uma metodologia para avaliação de imóveis urbanos baseados em métodos estatísticos multivariados**. Dissertação de Mestrado - Programa de Pós Graduação em Métodos Numéricos em Engenharia, Universidade Federal do Paraná, UFPR, Curitiba, PR, 2005.

CHARNET, R.; FREIRE, C. A. L.; CHARNET, E. M. R.; BONVINO, H. **Análise de Modelos de Regressão Linear com Aplicações**. Campinas: Editora da Unicamp, 1999. 356p.

COUTO, P. M. **Avaliação Patrimonial de Imóveis para Habitação**. 2007. 566 f. Tese de Doutorado, Laboratório Nacional de Engenharia Civil, Universidade do Porto, Porto. 2007.

GAZOLA, S. **Construção de um modelo de regressão para avaliação de imóveis**. Dissertação de Mestrado – Programa de Pós-Graduação em Engenharia de Produção, Universidade Federal de Santa Catarina, UFSC, Florianópolis, SC, 2002.

MONTGOMERY, D. C.; RUNGER, G. C. **Estatística Aplicada e Probabilidade para Engenheiros**. 2 ed. Rio de Janeiro: LTC Editora, 2008. 463 p.

NADAL, C. A.; JULIANO, K.A.; RATTON, E. Testes Estatísticos Utilizados para a Validação de Regressões Múltiplas Aplicadas na Avaliação de Imóveis Urbanos. **Bol. Ciênc. Geod.**, Curitiba, v. 9, nº 2, p. 243-262, 2003.

STEINER, M.T.A.; NETO, A.C.; BRAULIO, S.N.; ALVES, V.. Métodos Estatísticos Multivariados Aplicados à Engenharia de Avaliações. **Gest. Prod.**, São Carlos, v. 15, n. 1, p. 23-32, jan.-abr. 2008.

Portal da Cidade de Sorocaba. Texto. Disponível em: <http: www.sorocaba.com.br >. Acesso em: 02 setembro 2010.

R Development Core Team (2010). 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.

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.

SIRMANS, S.G.; MACPHERSON, D.A.; ZIETZ, E.N. The composition of hedonic pricing models. **Journal of Real Estate Literature**, v. 13, n. 1, p. 3-43, 2005.

WOOLDRIDGE, J. M. **Introdução à econometria:** uma abordagem moderna. 4 ed. São Paulo: Cengage Learning Editora, 2010. 701 p.