INFORMAÇÕES GERAIS DO TRABALHO

Título do Trabalho: Estatística Multivariada Aplicada à Engenharia de Avaliações

Autor (es): Aguinaldo Manoel da Silva Junior, Denilson Junio Marques Soares, Thiago Pastre Pereira, Germano de Oliveira Mattosinho.

Palavras-chave: Estatística Multivariada, Engenharia de Avaliaçãoes, Regressão Linear Múltipla.

Campus: Piumhi

Área do Conhecimento (CNPq): 1.02.02.00-5 – Estatística

Tipo de Bolsa: Programa Institucional de Bolsas de Iniciação Cientifica (PIBIC) com fomento interno.

RESUMO

A Engenharia de avaliações é um ramo da engenharia que procura determinar tecnicamente o valor de um bem, de seus direitos, frutos e custos de reprodução. Para isto, a principal metodologia utilizada está na análise de regressão, que consiste na realização de análises estatísticas que buscam identificar uma relação funcional entre uma variável dependente com uma ou mais variáveis independentes.

Para estimação dos parâmetros de um modelo de regressão linear múltipla, um dos métodos mais utilizados é o dos mínimos quadrados, que consiste em minimizar a soma dos quadrados dos erros de um modelo. Cálculos envolvendo álgebra matricial são necessários para a determinação desses parâmetros, cuja significância pode ser avaliada utilizando a estatística F-Snedicor e o teste t para comparação de médias. Para analisar a qualidade do ajuste é comum utilizarmos o coeficente de determinação ajustado.

Este trabalho visa apresentar uma metodologia, composta por técnicas estatísticas multivariadas, para a construção de um modelo de regressão linear múltipla para avaliação de imóveis de um determinado município, considerando a localização (distância do imóvel à prefeitura do município, em metros), área do terreno (em m²) e área edificada (em m²), buscando um modelo com um ajuste adequado aos dados e cuja capacidade preditiva seja satisfatória.

Para isto, considerou-se uma amostra hipotética de tamanho $n = 12$ e verificou-se que, para esta amostra, a área do terreno não foi significativa para o modelo. Assim, estimou-se um novo modelo, considerando a localização e a área edificada. Essas variáveis foram significativas e o modelo apresentou um

coeficiente de determinação ajustado satisfatório, indicando que o modelo se ajustava bem aos dados.

É importante ressaltar que outras variáveis independentes podem ser inseridas no modelo, podendo acarretar em um modelo mais completo. Dessa forma, podemos perceber que as técnicas da análise multivariada podem ser utilizadas no âmbito da Engenharia de avaliações, servindo como um instrumento facilitador na determinação de valores de bens.

INTRODUÇÃO:

Segundo Walpole et. al (2009), a análise de regressão múltipla é uma técnica utilizada a fim de estudar a relação linear entre várias variáveis explicativas de um determinado processo. Diversos problemas relacionados a área da engenharia faz uso das técnicas e conceitos relacionados a esta importante área da estatística.

Suponhamos que exista uma relação linear entre uma variável dependente Y e K variáveis independentes X. Neste caso, o modelo estatístico que ilustra a situação é dado por:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \dots + \beta_k X_{ki} + \varepsilon_i$$

Fazendo i variar de 1 até n, temos:

$$Y_1 = \beta_0 + \beta_1 X_{11} + \beta_2 X_{21} + \cdots + \beta_k X_{k1} + \epsilon_i$$

$$Y_2 = \beta_0 + \beta_1 X_{21} + \beta_2 X_{22} + \cdots + \beta_k X_{k2} + \epsilon_2$$

$$\vdots \qquad\qquad \ddots \qquad\qquad \vdots$$

$$Y_n = \beta_0 + \beta_1 X_{1n} + \beta_2 X_{2n} + \cdots + \beta_k X_{kn} + \epsilon_n$$

que em notação matricial pode ser escrito como

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_{11} & X_{21} & \dots & X_{k1} \\ 1 & X_{12} & X_{22} & \dots & X_{k2} \\ \vdots & & \ddots & & \vdots \\ 1 & X_{21} & X_{2n} & \dots & X_{kn} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix} \Rightarrow \boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon} \qquad (1)$$

Para determinar o valor dos elementos da matriz $\boldsymbol{\beta}$, pode-se utilizar o Método dos Mínimos Quadrados (CASELLA & BERGER, 2002), que consiste em minimizar a soma dos quadrados dos erros, a fim de encontrar o melhor ajuste para um conjunto de dados.

Este trabalho tem o objetivo de apresentar como é realizado este ajuste, além de apresentar técnicas para verificação da sua qualidade. Também será apresentada uma aplicação em Engenharia de Avaliações que, de acordo com Dantas (2012), trata-se de um ramo da engenharia que procura determinar tecnicamente o valor de um bem, como um imóvel, considerando variáveis que possam apresentar significância para a construção do modelo, como a localização, área do terreno, área edificada, pavimentação, topografia, entre outras.

METODOLOGIA:

Considerando que exista uma relação linear entre as variáveis independentes e a variável dependente, que os resíduos são independentes e possuem distribuição normal e que as variâncias são homogêneas, podemos tomar a soma dos quadrados dos desvios, obtidos através da equação (1), obtendo:

$$\boldsymbol{\epsilon} = \boldsymbol{Y} - \boldsymbol{X\beta} \Rightarrow L = \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} \boldsymbol{\epsilon}_{\boldsymbol{i}}^{\boldsymbol{2}} = \boldsymbol{\epsilon}'\boldsymbol{\epsilon} = (\boldsymbol{Y} - \boldsymbol{X\beta})'\boldsymbol{Y} - \boldsymbol{X\beta})$$

em que o produto desta relação resulta em um escalar, a notação $\boldsymbol{X}'$ representa o trasposto da matriz $\boldsymbol{X}$ enquanto que $\boldsymbol{Y}'$ e $\boldsymbol{\beta}'$ representam os transpostos dos vetores $\boldsymbol{Y}$ e $\boldsymbol{\beta}$ respectivamente. Ao utilizar a técnica de derivação matricial obtemos

$$\frac{\partial L}{\partial \boldsymbol{\beta}} = -\boldsymbol{2X'Y} + \boldsymbol{2X'X\beta}$$

Igualando a zero e substituindo o vetor $\boldsymbol{\beta}$ por $\widehat{\boldsymbol{\beta}}$, temos

$$(\boldsymbol{X'X})\widehat{\boldsymbol{\beta}} = \boldsymbol{X'Y}$$

De forma geral, a matriz $(\boldsymbol{X'X})$ é não singular e portanto é invertível, assim os estimadores para parâmetros são dados pelo vetor

$$\boldsymbol{\beta} = (\boldsymbol{X'X})^{-\boldsymbol{1}}\boldsymbol{X'Y}$$

que, através de manipulações em álgebra matricial, resulta:

$$\boldsymbol{\beta} = \frac{1}{n\sum X_i^2 - \left(\sum X_i\right)^2} \begin{bmatrix} \sum X_i^2 & -\sum X_i \\ -\sum X_i & n \end{bmatrix} \begin{bmatrix} \sum Y_i \\ \sum X_i Y_i \end{bmatrix}$$

Podemos verificar se existe relação linear entre a variável dependente Y e alguma variável independente $X_1, X_2, \ldots, X_k$. Para isto, realizamos um teste para significância da regressão, cuja hipótese de nulidade é que os estimadores para os parâmetros são todos nulos. Para que tenhamos pelo menos uma das variáveis independentes significativa para o modelo, devemos rejeitar esta hipótese. A Tabela 1 serve de auxílio na obtenção da estatística F que possui distribuição F-Snedicor (central) com $p, n-p-1$ graus de liberdade (BUSSAB & MORETTIN, 2017).

Tabela 1: Análise de variância da regressão linear múltipla

Fonte	Graus de liberdade	Soma dos quadrados	Quadrado médio	Quociente F
Regressão	k	SQR	QMR = $\frac{SQR}{k}$	$F = \frac{QMR}{QME}$
Resíduo	n – (k+1)	SQE	QME = $\frac{SQE}{n-(k+1)}$	
Total	n - 1	SQT		

em que

$$SQR = \boldsymbol{\beta X'Y} - \frac{\left(\sum_{i=1}^{n} Y_i\right)^2}{n} \; ;$$

$$SQE = SQT - SQR \; ;$$

$$SQT = \boldsymbol{Y'Y} - \frac{\left(\sum_{i=1}^{n} Y_i\right)^2}{n}$$

Para um análise individual dos coeficientes da regressão, podemos utilizar o Teste t, associado a $n-k-1$ graus de liberdade, cuja estatística é obtida por:

$$t = \frac{\widehat{\beta_j} - \beta_j}{\hat{\sigma}\,(\beta_j)},$$

Neste caso, rejeitamos a hipótese H_0: $\beta_j = 0$, caso $|t_{cal}| > t_{(\beta, n-k+1)}$ (BUSSAB & MORETTIN, 2017).

O coeficiente de determinação para regressão linear múltipla é definido pela razão entre a soma de quadrados da regressão e a soma de quadrados total. De acordo com Gurajati & Porter (2011), quanto maior no número de variáveis independentes no modelo de regressão, maior tende a ser o coeficiente de regressão. Para contornar esta situação, podemos utilizar o coeficiente de determinação ajustado $\overline{R^2}$ para fins de analisar a qualidade do ajustamento, dado por:

$$\overline{R^2} = \frac{R^2\,(n-1) - k}{n-k-1}$$

Segundo Cecon et. al. (2012), um R^2 em torno de 0,40 é suficiente para que o modelo seja considerado significativo a 5% de probabilidade.

Neste trabalho, consideramos uma amostra hipotética de tamanho $n = 12$, que relaciona o valor venal do metro quadrado de imóveis com a distância, em metros, em que se encontra da prefeitura do município, a área do terreno em que se encontra e a área de edificação, ambas em metros quadrados. Os resultados foram obtidos com o auxílio do software estatístico R (R, 2018).

RESULTADOS E DISCUSSÕES:

O primeiro modelo obtido, considerando as três variáveis: localização (X_1), área do terreno (X_2), e área edificada (X_3), é o que segue:

$$Y = 8957{,}8940 \ - 0{,}9217X_1 - 0{,}9587X_2 + 4{,}7539\ X_3$$

As Tabelas 2 e 3 trazem os resultados da análise da estatística F e do Teste t, para este modelo:

Tabela 2: Análise de variância da amostra

Fonte	Graus de liberdade	Soma dos quadrados	Quadrado médio	Quociente F
Localização	1	26978973	26978973	1531,7601*
Área do terreno	1	189040	189040	10,7330*
Área edificada	1	140249	140249	7,9628*
Resíduos	8	140904	17613	
Total	11	27449166		

Tabela 3: Teste t da amostra

	Estimativa	Valor t	p-valor
Intercepto	8957,89402	47,744	4.10e-11*
Localização	-0,92167	-24,242	8.95e-09*
Área do terreno	-0,95871	-0,735	0,4835
Área edificada	4,75392	2,822	0,0224*

A estatística F foi significativa. Entretanto, a realização do Teste t indicou que a variável área do terreno (X_2), não foi significativa, considerando 5% de significância. Dessa forma, um novo modelo foi estimado, considerando as variáveis localização (X_1) e área edificada (X_2), obtendo o seguinte modelo:

$$Y = 8932{,}1591 - 0{,}9196\ X_1 + 3{,}6926X_2$$

As Tabelas 4 e 5 trazem os resultados da análise da estatística F e do Teste t, para este modelo:

Tabela 4: Análise de variância da amostra

Fonte	Grau de liberdade	Soma dos quadrados	Quadrado médio	Quociente F
Localização	1	26978973	26978973	1.824e-11*
Área edificada	1	319783	319783	0,0017*
Resíduos	9	150411	16712	
Total	11	27449167		

Tabela 5: Teste t da amostra

	Estimativa	Valor t	p-valor
Intercepto	8932,15910	49,747	2.69e-12*
Localização	-0,91962	-24,898	1.31e-09*
Área edificada	3,69256	4,374	0,00179*

Os parâmetros desse modelo foram significativos, à 5% de significância, dessa forma, podemos estimar que o valor venal do metro quadrado de um imóvel que encontra-se a 1 km da prefeitura e cuja área edificada é de $120\ m^2$ é:

$$Y = 8932{,}1591 - 0{,}9196 \times 1000 + 3{,}6926 \times 120 = 8455{,}6711$$

O coeficiente de determinação ajustado foi superior a $0{,}9$, indicando uma boa qualidade dos dados tendo uma variabilidade muito pequena explicada pela regressão.

CONCLUSÕES:

Os modelos de regressão linear podem ser utilizados em diversas áreas, como instrumento para analisar o grau de associação entre variáveis. Neste trabalho, mostramos que o uso de técnicas multivariadas pode ser utilizado também em Engenharia de Avaliações, em que podemos estimar o valor de um bem, com base em característica que julgarmos serem significativas na composição de seu preço.

A construção de um modelo estatístico para estimar os valores venais de imóveis urbanos pode auxiliar diretamente na arrecadação tributária de municípios. Especialmente, no caso de municípios de pequeno porte, a implementação de técnicas significativas para auxiliar na avaliação de imóveis é uma tarefa desafiadora, devido, entre outros fatores, à falta de mão de obra qualificada e de recursos financeiros a serem destinados para este fim. Em suma, espera-se que este trabalho, entre outras funções, possa servir como um instrumento de difusão dessas técnicas na avaliação de bens.

AGRADECIMENTOS:

Agradecemos ao Instituto Federal de Minas Gerais (IFMG) pelo suporte e auxílio financeiro para o desenvolvimento deste trabalho de pesquisa.

SEMINÁRIO DE INICIAÇÃO CIENTÍFICA

INSTITUTO FEDERAL Minas Gerais
Pró-Reitoria de Pesquisa, Inovação e Pós-Graduação - PRPPG

REFERÊNCIAS BIBLIOGRÁFICAS:

BUSSAB, W. O.; MORETTIN, P. A. **Estatística básica**. São Paulo: Editora Saraiva, 2017. 540 p.

CASELLA, G.; BERGER, R. L. **Statistical inference**. Wadsworth and Brooks/Cole: Pacific Grove, 1990. CA: Duxbury, 2002. 650 p.

CECON, P. R. et al. **Métodos estatísticos**. Viçosa: Editora UFV, 2012. 229 p.

DANTAS, R. A. **Engenharia de Avaliações: uma Introdução à Metodologia Científica**. São Paulo: Pini, 2012. 262 p.

GUJARATI, D. N.; PORTER, D. C. **Econometria Básica,** vol. 5. Amgh Editora, 2011. 924p.

R DEVELOPMENT CORE TEAM. **R**: A Language and Environment for Statistical. Computing. Vienna: R Foundation on Statistical Computing, 2017. Disponível em: https://www.r-project.org. Acesso: 31/03/2018.

WALPOLE, R. E. et al. **Probabilidade e Estatística para engenharia e ciências**. São Paulo: Editora Pearson, 2009. 491 p.

Participação em Congressos, publicações e/ou pedidos de proteção intelectual:

Esta é a primeira vez em que este trabalho será apresentado.