

Inferência Estatística - Seleção de Variáveis Preditoras na Regressão Múltipla

Ênfase em Multicolinearidade entre variáveis independentes

Arquiteta Urbanista Ana Maria de Biazzi Dias de Oliveira

anabiazzi@uol.com.br



REALIZAÇÃO



PATROCÍNIO



Motivação

- Na modelagem da Regressão Múltipla (RM) para avaliação de bens pelo Método Comparativo Direto de Dados de Mercado (ABNT- NBR 14.653-2), a **seleção das variáveis independentes**, é uma das etapas mais importantes para compor os dados da **amostra**.
- A existência de **preditores redundantes** (ou **correlacionados entre si**) é uma particularidade da formação do Valor do Imóvel.
- Com efeito, é razoável supor, p.ex., que apartamentos, com maior área, tenha mais dormitórios, mais vagas de garagem, mais banheiros, mais suítes... *(Naturalmente, tal situação não é uma regra)*.
- Neste contexto, surge o problema da **Multicolinearidade** - um fenômeno estatístico em que uma variável preditora pode ter combinação linear com outras variáveis predictoras incluídas no modelo de regressão.

? **Como estamos tratando esta questão nos trabalhos de avaliação?**

Seleção de Variáveis na Regressão Múltipla

No modelo de regressão múltiplo (MRM), a utilização de muitas variáveis independentes é uma escolha do profissional em relação à justificativa para a inclusão de determinada variável no vetor de variáveis X, de forma que consiga fundamentar a inserção dessas variáveis no *modelo de regressão*.

O conceito de ceteris paribus (mantidas as demais condições constantes) deve ser utilizado, uma vez que a interpretação de cada variável é feita de forma isolada.

Para que o uso desta equação seja eficaz fundamental examinar previamente os pressupostos da RM, bem como **identificar as consequências da sua violação**.

Dentre as suposições inerentes à análise de regressão - normalidade dos erros, linearidade do fenômeno medido, variância constante dos termos de erro (homoscedasticidade), independência dos termos de erro - **que são problemas estatísticos**, tem-se **a Multicolinearidade, que é um problema dos dados**

Dessa forma, é possível existir Multicolinearidade, mesmo quando os demais pressupostos do modelo linear de mínimos quadrados são respeitados.

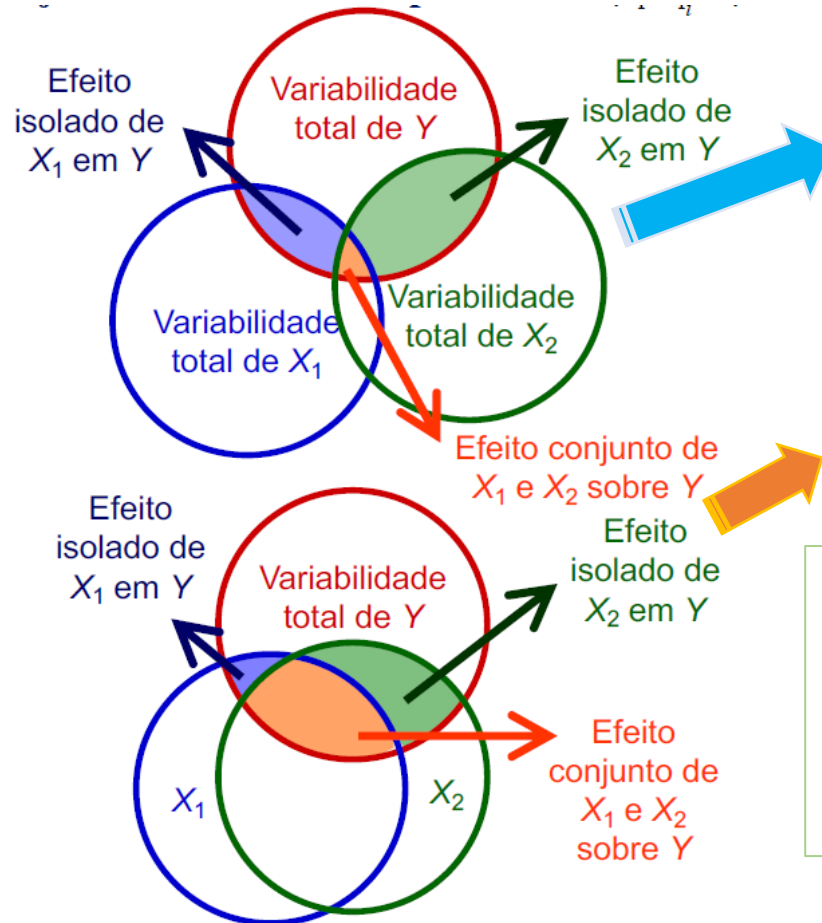
Multicolinearidade - Conceito

Seja o modelo definido por:

$$Y = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + e_i$$

Na relação entre as variáveis independentes X_1 e X_2 , o efeito sobre Y pode ser dividido em 3 partes:

i) efeito isolado de X_1 ; ii) isolado de X_2 ; e iii) efeito conjunto de X_1 e X_2 .



Se X_1 e X_2 são independentes, a variabilidade de Y explicada pelo modelo de RLM divide-se em duas partes: *o efeito isolado de X_1 e o efeito isolado de X_2 .*

Quando há uma **forte relação linear entre X_1 e X_2** , difícil identificar os efeitos isolados de X_1 e X_2 sobre Y . Ou seja, a maior parcela da variabilidade de Y é explicada pelo efeito conjunto de X_1 e X_2 .

O problema é que a multicolinearidade torna muitas vezes as estimativas dos coeficientes dos parâmetros (B 's) insignificantes, já que, cada um, pressupõe, por definição, a variação em Y dada uma variação unitária em X , mantendo-se constantes as demais informações. Ou seja, se duas variáveis independentes são fortemente correlacionadas, será muito difícil haver variação em uma, sem que haja em outra.

Detecção de Multicolinearidade

- ✓ Pela observação de particularidades no modelo inferido

Algumas indicações da presença de Multicolinearidade:

- valores **altos do coeficiente de correlação na matriz correlação** das variáveis explicativas;
- **grandes alterações nas estimativas dos coeficientes de regressão**, quando uma variável independente for adicionada ou retirada do modelo, ou quando uma observação for alterada ou eliminada;
- **conflito entre estatísticas R^2 e F do modelo e testes t para os parâmetros B** : as estatísticas R^2 e F podem indicar um modelo significativo, enquanto testes t de parâmetros B 's seriam insignificantes.
- obtenção de estimativas para os **coeficientes de regressão com sinais algébricos contrários** àqueles que seriam esperados a partir de conhecimentos teóricos disponíveis ou de experiências anteriores sobre o fenômeno estudado e
- obtenção de intervalos de confiança com elevadas amplitudes para os coeficientes de regressão, associados a variáveis independentes importantes.

Detecção de Multicolinearidade

- ✓ Pela matriz correlação das variáveis independentes

É a forma mais simples de detectar a multicolinearidade- quanto **maior a magnitude dos coeficientes, maiores os eventuais problemas.**

A norma da ABNT NBR 14.653-2 em seu item A.2.1.5.2, recomenda que esta verificação seja feita, e indica como premissa "....*atenção especial para resultados superiores a 0,80*"

Limitação: A matriz de correlação permite avaliar a **existência de dependência linear entre par de variáveis**. Porém, quando mais de dois regressores estão envolvidos na dependência linear, a matriz de correlação não é eficiente.

Detecção de Multicolinearidade

✓ Diagnóstico pelo uso de técnicas analíticas e estatísticas

Fator de Inflação de Variância - FIV, que é uma medida que indica o quanto a variância de cada coeficiente estimado aumenta devido a Multicolinearidade. Dado pela expressão:

$$VIF_j = 1 / (1 - R^2_{i.j}), \text{ onde}$$

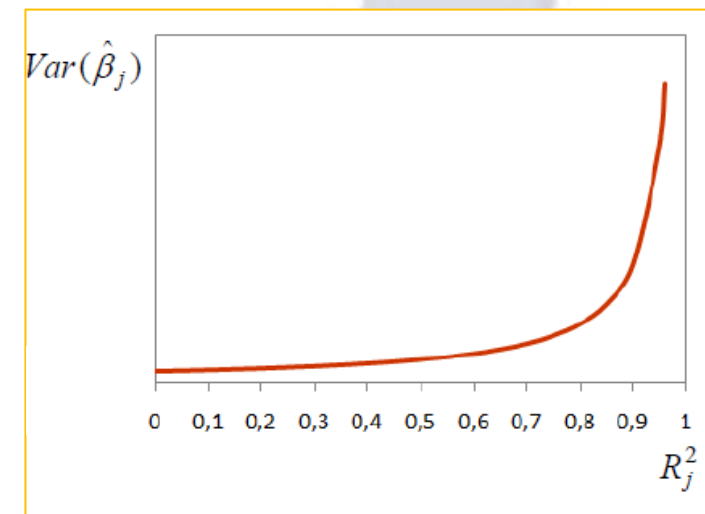
$R^2_{i.j}$ é o Coeficiente de Determinação entre a variável independente i e as j variáveis independentes do modelo.

Se as colunas de **X** forem ortogonais (variáveis não correlacionadas), então $VIF_j = 1$ para todo j e **menor a preocupação com a Multicolinearidade**.

À medida que aproxima de uma relação exata ($R^2_j = 1$), o FIV_j tenderá a infinito.

Valores para FIV superiores a 1 indicam a presença de multicolinearidade.

Quanto maior o valor do fator, mais severa é a multicolinearidade.



Alerta não consenso na literatura referência para Fatores de Inflação de Variância num eventual descarte do modelo, ou mesmo para a tomada de medidas corretivas. Alguns autores, indicam que valores superiores a 10 indicam multicolinearidade severa, comprometendo o modelo.

FIV = 4 ou FIV = 10 não devem ser adotadas de forma direta, sem uma análise de outros aspectos do modelo.

REALIZAÇÃO



PATROCÍNIO



Tratamento da Multicolinearidade

Detectada a ocorrência de **multicolinearidade severa** no modelo inferido, analisar os procedimentos a serem tomados para o tratamento da mesma.

❖ **Adicionar dados** *Nem sempre é possível*

Por restrição econômica, inexistência de dados; ou

Ser inviável **quando a multicolinearidade é devido a restrição no modelo ou na população.**

(Exemplo: considere os regressores área da casa e área do terreno - adicionar dados será irrelevante, se esta relação é uma característica estrutural da população, já que os dados adicionais exibirão esse comportamento.

❖ **Eliminar variáveis do modelo** - *procedimento que exige cautela*

A retirada de variáveis explicativas do modelo, **não deve ser feita por causa de seu baixo valor de t, sob risco de retirar-se importante variável**

Pode ser eficaz se os regressores eliminados não tem um alto poder de influência na variabilidade da variável resposta.

Recomendação é eliminação após o reconhecimento do problema da Multicolinearidade no modelo – por métodos de *stepwise* e usando as estatísticas *R2 Ajustado* e *VIFs*.

REALIZAÇÃO



PATROCÍNIO



Tratamento da Multicolinearidade

❖ adoção de técnicas estatísticas mais avançadas, a exemplo do uso de regressão de componentes principais.

A análise de componentes principais substitui um conjunto de variáveis correlacionadas (X) por um conjunto de novas variáveis não-correlacionadas (Y), sendo essas combinações lineares das variáveis iniciais e colocadas em ordem decrescente por suas variâncias

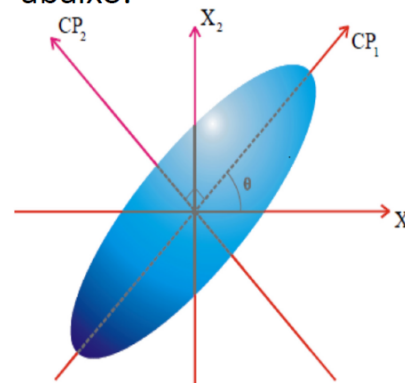
$\text{Var } Y_1 > \text{Var } Y_2 > \dots > \text{Var } Y_p$

Algebricamente, as componentes principais são combinações lineares das “p” variáveis aleatórias $X_1, X_2, \dots, X_p$;

Geometricamente, as combinações lineares representam um novo sistema de coordenadas obtido pela **translação** e **rotação** do sistema original com $X_1, X_2, \dots, X_p$ como eixos;

Os novos eixos representam as direções com as maiores variabilidades e fornecem uma descrição mais simples e mais parcimoniosa da estrutura de covariância, pois as componentes principais são não correlacionadas

Para duas variáveis, x_1 e x_2 , espera-se que os dados estejam correlacionados, os dados num diagrama de dispersão são representados pela elipse na figura abaixo.



CP_1 tem maior variância, corresponde ao maior eixo da elipse.

CP_2 tem menor variância e é perpendicular ao eixo maior.

As variâncias são proporcionais aos autovalores.

13

REALIZAÇÃO



PATROCÍNIO



Exemplo (conceitual)

Avaliação de Terreno –alta correlação (0,84) entre duas variáveis X1 (Dicotômica - *com construções ou não*) e X2 (- *Distancia ao Polo Valorizante* em metros)

matriz correlação (sem transformações)

Matriz Correlações de 1a ordem	Com Construções (Sim 1 ou não 0)	Distancia Polo (m)	Transação/ Oferta (0-1)	Área Terreno (m2)	Coef Aproveitamento	Valor Unitário
Com Construções (Sim ou não)	1					
Distancia Polo	-0,844690	1				
Transação/Oferta	0,247717	-0,242899	1			
Área	-0,369491	0,412283	0,057312	1		
Coef Aproveitamento	0,353553	-0,547919	0,155700	0,278212	1	
Valor Unitário	0,757604	-0,689601	0,358960	-0,714114	0,034850	1

matriz correlação (variáveis transformadas)

Matriz Correlações de 1a ordem	Com Construções (Sim 1 ou não 0)	Distancia Polo (m)	Transação/ Oferta (0-1)	Área Terreno (m2)	Coef Aproveitamento	Valor Unitário
Com Construções (Sim ou não)	1					
Distancia Polo (1/x)	0,704557	1				
Transação/Oferta	0,247717	0,268054	1			
Área	-0,369491	-0,516880	0,057312	1		
Coef Aproveitamento	0,353553	0,289554	0,155700	0,278212	1	
Valor Unitário (LN)	0,709265	0,627087	0,346889	-0,749213	0,2209	1

Problemas– Inviável as duas variáveis no modelo – quando inserida a dicotômica (com ou sem construções), **a distancia a polo valorizante inverte o sinal (a medida que distancia do polo o valor unitário aumenta e o teste do regressor – aceita a hipótese nula**

Valor Unitário =

b0 1.565,11

b1 -0.2216 * Distancia Polo

b2 +12,51 * Transação/Oferta

b3 -0.38175 * Área

b4 + 19,6521 * Coef Aproveitamento

Valor Unitário =

b0 1.113

b1 +228.3477857 * Com Construções

b2 +0.008483718201 * Distancia Polo

b3 +117.8717494 * Evento

b4 -0.4287949774 * Área

b5 -38.35613931 * Coef Aproveitamento

Valor Unitário = e^ (

710,89

b1 +0.25665 * Com Construções

b2 -10.8217 / Distancia Polo

b3 +0.1534 * Evento

b4 -0.0006063987569 * Área

b5 +0.1297 * Coef Aproveitamento)

		t test - Sig	R2	F calc	Valor Unitário
Modelo 1	Dist Polo	1,14%	79%	15,09	R\$ 1.010,00

		t test - Sig	R2	F calc	Valor Unitário
Modelo 2	Dist Polo	67,00%	85%	19,39	R\$ 740,00
	Com Const ou Não	1,5%			

		t test - Sig	R2	F calc	Valor Unitário
Modelo 3	Dist Polo	43,41%	86%	18,77	R\$ 695,00
	Com Const ou Não	0,35%			

REALIZAÇÃO



PATROCÍNIO



Análise prática – Cont.

Análise FIV- Fator de Inflação da Variância

$$FIV = \frac{1}{(1 - R_f^2)} = \frac{1}{(1 - 0,889)} = 8,98$$

Há Multicolinearidade severa, o que estaria dificultando a identificação dos impactos isolados dos regressores. O FIV é 9 vezes superior ao que seria na ausência de relação entre os regressores.

Análise - Componentes Principais

A análise de componentes principais – desenvolvida no **Matlab** – resultou o modelo com 2 componentes principais(resumindo as 5 variáveis analisadas).

No modelo foi possível obter o Valor unitário de R\$ 850/m2 - o qual dentre outras variáveis ,, leva em conta a posição do imóvel (próximo ao polo (300m) e a o fato de não conter edificações (o) e obter

		t test - Sig	R2	F calc	Valor Unitário
Modelo 1	Dist Polo	1,14%	79%	15,09	R\$ 1.010,00
Modelo 2	Dist Polo	67,00%	85%	19,39	R\$ 740,00
	Com Const ou Não	1,5%			
Modelo 3	Dist Polo	43,41%	86%	18,77	R\$ 695,00
	Com Const ou Não	0,35%			
Modelo de regressão - por PCA - componetes principais					R\$ 1.010,00

REALIZAÇÃO



PATROCÍNIO



O que diz as Normas da ABNT 14.653-2

Colinearidade ou Multicolinearidade (item A.2.1.2)

A.2.1.5.1 Uma forte dependência linear entre duas ou mais variáveis independentes provoca degenerações no modelo e limita a sua utilização. As variâncias das estimativas dos parâmetros podem ser muito grandes e **acarretar a aceitação da hipótese nula e a eliminação de variáveis fundamentais**.

A.2.1.5.2 Para verificação da multicolinearidade deve-se, em primeiro lugar, analisar a matriz das correlações, que espelha as dependências lineares de primeira ordem entre as variáveis independentes, **com atenção especial para resultados superiores a 0,80**. Como também é possível ocorrer multicolinearidade, mesmo quando a matriz de correlação apresenta coeficientes de valor baixo, recomenda-se, também, verificar o correlacionamento de cada variável com subconjuntos de outras variáveis independentes, por meio de regressões auxiliares, como pela **análise de variância por partes**.

A.2.1.5.3 Para tratar dados na presença de multicolinearidade, é recomendável que sejam tomadas medidas corretivas, como a ampliação da amostra ou adoção de técnicas estatísticas mais avançadas, a exemplo do uso de regressão de componentes principais.

A.2.1.5.4 Nos casos em que o imóvel avaliando segue os padrões estruturais do modelo, a existência de multicolinearidade pode ser negligenciada.

A.2 Pressupostos básicos A.2.1

f) em caso de correlação linear elevada entre quaisquer subconjuntos de variáveis independentes, isto é, multicolinearidade, **deve-se examinar a coerência das características do imóvel avaliando com a estrutura de multicolinearidade inferida, vedada a utilização do modelo em caso de incoerência;**

Conclusões

- ✓ A análise da presença de Multicolinearidade nos modelos pode ser diagnosticada por diversas formas – desde a observação de particularidades no modelo inferido como por uso de técnicas analíticas e estatísticas e um deles é o Fator de Inflação de Variância – FIV
- ✓ A matriz de correlações não é suficiente e resultados superiores a 0,80 (recomendado pelas normas) entre par de variáveis, pode não ser suficiente para resolver o problema:
A exclusão de variáveis correlacionadas nem sempre é a solução (pode ocorrer a omissão de variáveis importantes) e a manutenção pode levar a modelos incoerentes
- ✓ A norma da ABNT 14.653-2 e a literatura indicam outros caminhos (como, por exemplo, Análise de Componentes Principais) e nós temos que trilhar por eles.

Obrigada

“ Uma resposta aproximada para o problema certo vale muito mais do que uma resposta exata para um problema aproximado. ”

(John Wilder Tukey)

Arquiteta Urbanista Ana Maria de Biazzzi Dias de Oliveira

anabiazzi@uol.com.br

055 (11) 98109-7733



REALIZAÇÃO



PATROCÍNIO

