Multicolinearidade

Econometria

Alexandre Gori Maia

Ementa:

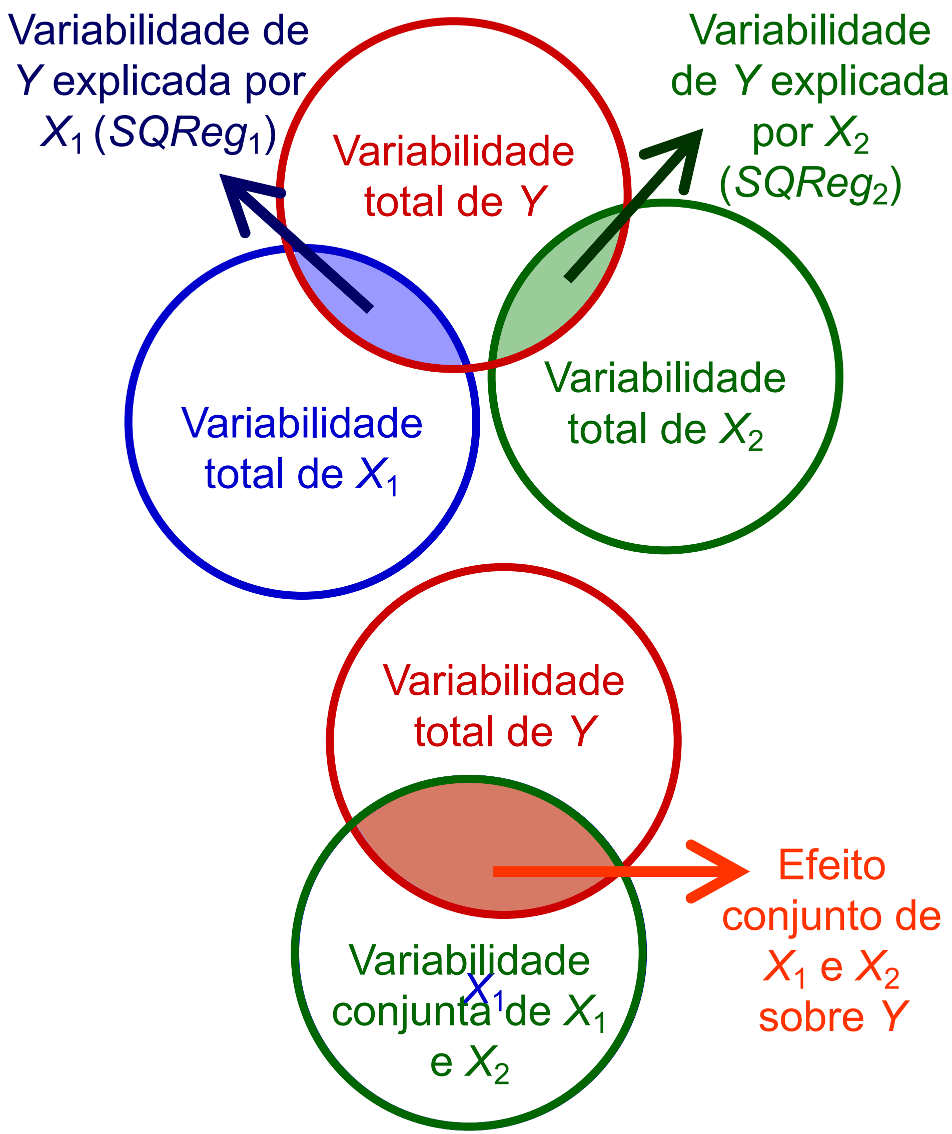
- Definição;
- Fator Inflacionário da Variância;
- Identificação: Estatística Conflitantes e Ajsute entre Regressores;
- Medidas Paliativas;

Bibliografia Básica:

- Maia, Alexandre Gori (2017). *Econometria: conceitos e aplicações*. Cap. 10.

Multicolinearidade - Conceito

Seja o modelo definido por: $Y = \alpha + \beta_1 X_{1_i} + \beta_2 X_{2_i} + e_i$



Se X_1 e X_2 são independentes, a variabilidade de Y explicada pelo modelo de RLM divide-se em duas partes disjuntas: o efeito isolado de X_1 ($SQReg_1$) e o efeito isolado de X_2 ($SQReg_2$).

Teríamos ainda os mesmos coeficientes angulares das regressões simples:

$$Y = \alpha_1 + \beta_1 X_{1_i} + e_{1_i} \Rightarrow SQ\mathrm{Re}\,g_1$$

$$Y = \alpha_2 + \beta_2 X_{2_i} + e_{2_i} \Rightarrow SQ\mathrm{Re}\,g_2$$

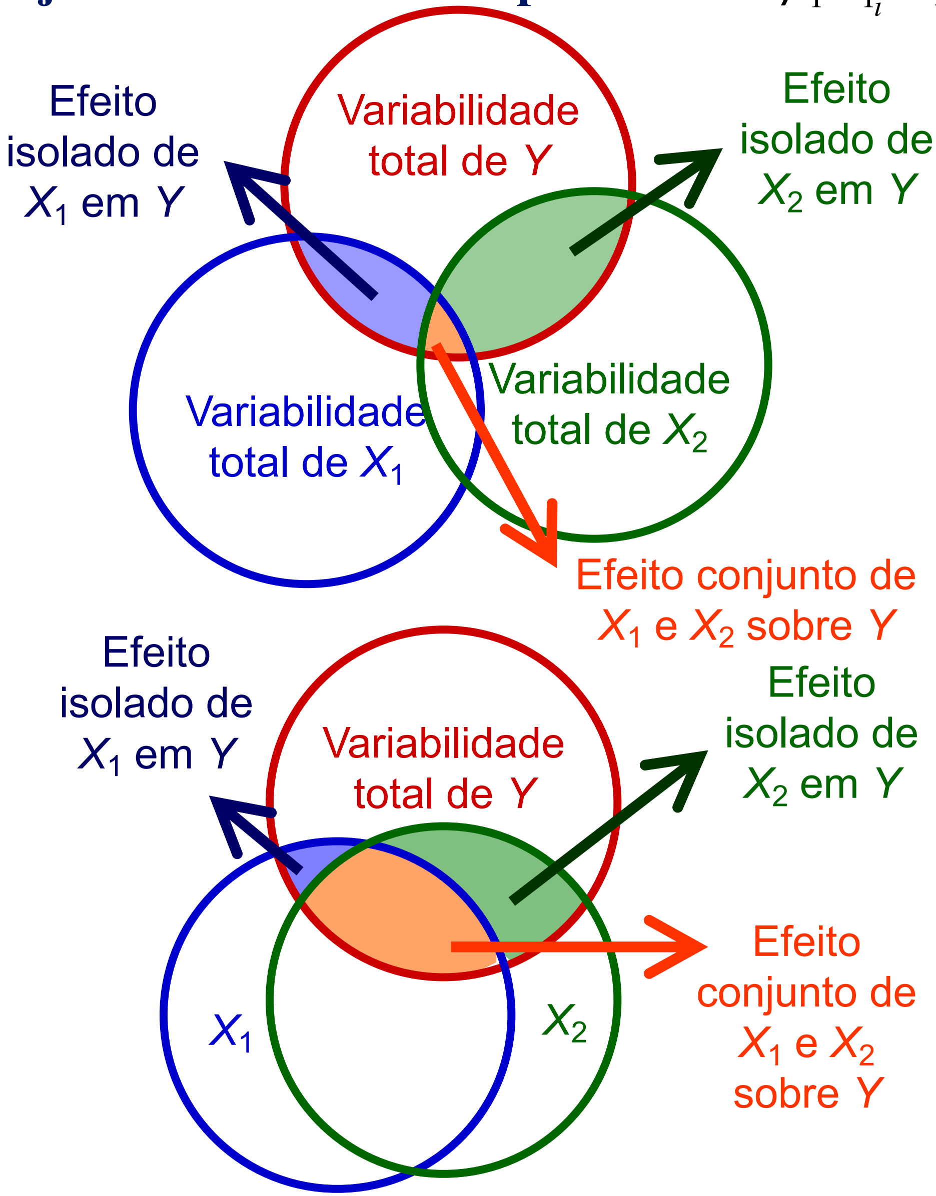
No outro extremo, poderíamos ter uma relação linear exata entre X_1 e X_2 (perfeita colinearidade), ou seja:

$$X_{1_i} = \lambda_2 X_{2_i}$$

Nesta situação, seria impossível estimar os efeitos isolados de X_1 e X_2 sobre Y.

Multicolinearidade - Conceito

Seja o modelo definido por: $Y = \alpha + \beta_1 X_{1_i} + \beta_2 X_{2_i} + e_i$



Frequentemente observamos relação entre as variáveis independentes X_1 e X_2 e, nesses casos, oefeito sobre Y poderá 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.

O problema é que quando há uma forte relação linear entre X_1 e X_2 (*multicolinearidade*) pode ficar muito 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.

Algebricamente, essa relação de multicolinearidade seria dada por:

$$X_{1_i} = \lambda_2 X_{2_i} + v_i$$

Multicolinearidade - Definição

Colinearidade Perfeita

Duas variáveis são ditas perfeitamente colineares quando há uma relação linear exata entre essas. De maneira genérica, podemos representar a linearidade perfeita por: $X_{j_i} = \lambda_1 X_{1_i} + \lambda_2 X_{2_i} + \ldots + \lambda_k X_{k_i}$

Multicolinearidade

Há multicolinearidade em um modelo de regressão múltipla quando duas ou mais variáveis independentes são fortemente relacionadas linearmente entre si. Nesse caso, teríamos: $X_{j_i} = \lambda_1 X_{1_i} + \lambda_2 X_{2_i} + \ldots + \lambda_k X_{k_i} + v_i$

Conseqüências da Multicolinearidade

A existência de uma colinearidade exata entre duas ou mais variáveis independentes torna impossível a obtenção dos coeficientes dos parâmetros por MQO. Por sua vez, na presença de multicolinearidade os estimadores de MQO continuam sendo os MELNV. O problema é que a multicolinearidade torna muitas vezes as estimativas dos coeficientes dos parâmetros (β'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.

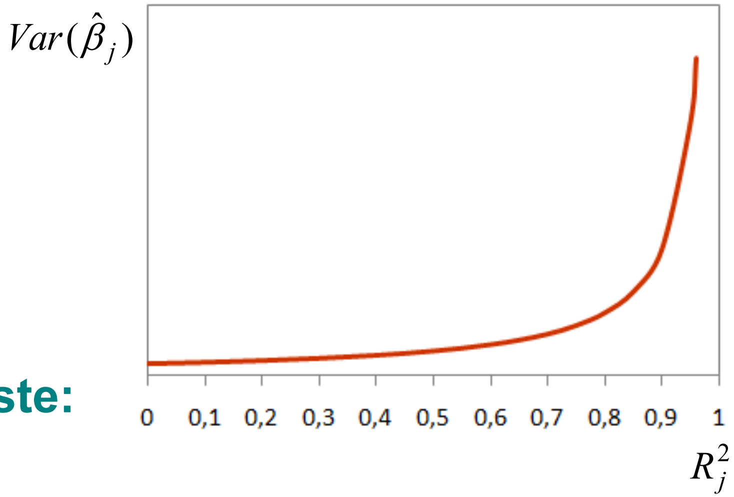
Fator Inflacionário da Variância

O vetor de variância e covariância dos estimadores será dada por:

$$Var(\hat{\boldsymbol{\beta}}) = (\mathbf{X}^T\mathbf{X})^{-1}\sigma^2$$

E a variância de cada estimador β_j será dada por:

$$Var(\hat{\beta}_j) = \frac{\sigma^2}{\sum_{i=1}^{n} x_{j_i}^2 (1-R_j^2)} = \frac{\sigma^2}{\sum_{i=1}^{n} x_{j_i}^2}\frac{1}{(1-R_j^2)} = \frac{\sigma^2}{\sum_{i=1}^{n} x_{j_i}^2} FIV_j$$



Onde R^2_j é o coeficiente de determinação do ajuste:

$$X_{j_i} = \lambda_0 + \lambda_1 X_{1_i} + \ldots + \lambda_{k-1} X_{k_i} + \nu_i$$

Fator Inflacionário da Variância - FIV

Representa o quanto a variância de $\hat{\beta}_j$ está sendo inflacionada pela relação de multcolinearidade entre X_j e as demais variáveis independentes. Quando não houver relação entre as variáveis independentes (R^2_j =0) o FIV_j será igual a 1 e, à medida que aproximamo-nos de uma relação exata (R^2_j =1), o FIV_j tenderá a infinito.

Multicolinearidade - Identificação

Alguns sinais de multicolinearidade:

Conflito entre estatísticas R^2 e F do modelo e testes t para os parâmetros β: as estatísticas R^2 e F podem indicar um modelo significativo, enquanto os testes t dos parâmetros β's seriam insignificantes.

$$Y = \alpha + \beta_1 X_{1_i} + \beta_2 X_{2_i} + e_i \quad \Rightarrow \quad H_0 : \beta_1 = \beta_2 = 0 \quad \mathbf{X} \quad H_0 : \beta_1 = 0 \ \text{e} \ H_0 : \beta_2 = 0$$

Ajuste linear significativo entre as variáveis independentes: uma das variáveis independentes apresentaria forte relação de linearidade com as demais.

$$X_{j_i} = \lambda_0 + \lambda_1 X_{1_i} + \ldots + \lambda_{k-1} X_{k_i} + v_i \quad \Rightarrow \quad R_j^2$$

Fator Inflacionário da Variância: medida de inflacionamento da variância devido à multicolinearidade.

$$R_j^2 \quad \Rightarrow \quad FIV_j$$

Multicolinearidade - Exemplo

Sejam as informações sobre emissões de CO^2, PIB e população de 8 países:

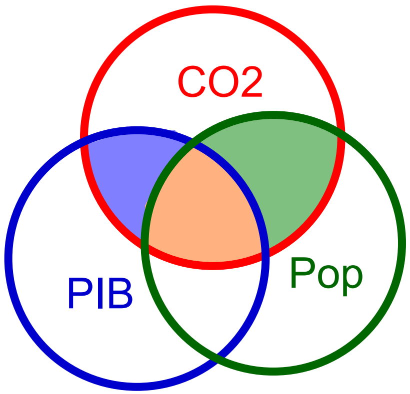
CO2	1.5	8.7	2.8	9.4	4.4	8.4	3.2	0.9
PIB	13.2	197.0	128.6	286.4	72.6	167.8	114.4	58.0
Pop	3.2	35.5	19.1	40.4	3.1	22.3	8.4	9.0

As estimativas de MQO para o modelo linear são:

$$CO2 = \beta_0 + \beta_1 \, PIB + \beta_2 \, Pop + e \quad \Rightarrow \quad CO2 = 0{,}472 + 0{,}030 PIB + 0{,}028 \, Pop + \hat{e}$$

Os resultados da tabela ANOVA:

Fonte	*gl*	*SQ*	*QM*	*F*	*p*
Regressão	2	63.9	31.9	8.80	0.023
Resíduos	5	18.2	3.6		
Total	7	82.0			



E os testes *t* para os coeficientes:

Variável	$\hat{\beta}$	$S_{\hat{\beta}}$	*t*	*p*
Intercepto	0.472	1.328	0.356	0.737
PIB	0.030	0.025	1.226	0.275
Pop	0.028	0.150	0.183	0.862

Embora o ajuste seja significativo no conjunto (teste *F*), as contribuições marginais de cada variável são insignificantes. Resultados que sugerem a presença de multicolinearidade.

Multicolinearidade - Exemplo

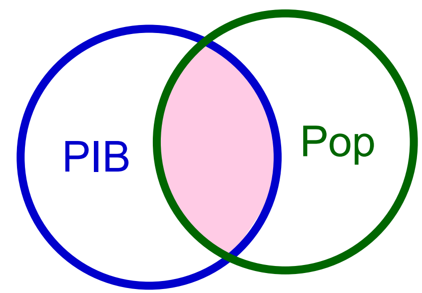
Para nos certificarmos da presença de multicolinearidade, relacionamos os regressores *PIB* e *Pop*:

O ajuste de MQO forneceu os seguintes resultados para a relação entre os regressores:

$$PIB = \lambda_0 + \lambda_1\, Pop + v \quad \Rightarrow \quad PIB = 29{,}2 + 5{,}7\, Pop + \hat{v}$$

Os resultados da tabela ANOVA:

Fonte	*gl*	*SQ*	*QM*	*F*	*p*
Regression	1	46.871,2	46.871,2	47,88	0,0005
Residual	6	5.873,5	978,9		
Total	7	52.744,6			



$$R^2_{PIB} = 0{,}889$$

Podemos ainda calcular o FIV por:

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

Há multicolinearidade nos regressores, 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.

Multicolinearidade - Correção

Alguma medidas paliativas na presença de multicolinearidade:

Aumentar o tamanho da amostra: aumentando o tamanho da amostra estaremos aumentando a variabilidade de X_j e, consequentemente, reduzindo a variância do estimador de β_j.

$$Y = \alpha + \beta_1 X_{1_i} + \beta_2 X_{2_i} + e_i \quad \Rightarrow \quad Var(\hat{\beta}_j) = \frac{\sigma^2}{\boxed{\sum_{i=1}^{n} x_{j_i}^2}} FIV_j$$

Tranformar as variáveis: a multicolinearidade pode ser eliminada a partir de funções das variáveis independentes.

$$Y = \alpha + \beta_1 X_{1_i} + \beta_2 X_{2_i} + e_i \quad \Rightarrow \quad Y = \delta_0 + \delta_0 Z_i + u_i \quad \text{onde} \quad Z_i = f(X_{1_i}, X_{2_i})$$

Omitir regressor que apresentar alta colinearidade com os demais: uma solução simples, mas perigosa, é excluir uma ou mais variáveis que apresentam multicolinearidade. A exclusão de variáveis essenciais para compreensão do problema pode, entretanto, gerar **viés de omissão**.

$$Y = \alpha + \beta_1 X_{1_i} + \beta_2 X_{2_i} + e_i \quad \Rightarrow \quad Y = \alpha + \beta_1 X_{1_i} + e_i$$

Correção Multicolinearidade - Exemplo

Sejam as informações sobre emissões de CO^2, PIB e população de 8 países:

O ajuste de MQO nos forneceu os seguintes resultados :

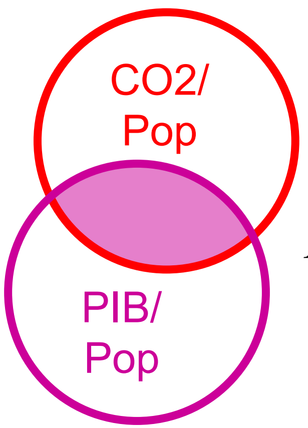
$$CO2 = \beta_0 + \beta_1\, PIB + \beta_2\, Pop + e \;\Rightarrow\; CO2 = 0{,}472 + 0{,}030 PIB + 0{,}028\, Pop + \hat{e}$$

Como havia multicolinearidade e os coeficientes eram insignificantes estatisticamente, podemos propor um novo ajuste:

$$\frac{CO2}{Pop} = \delta_0 + \delta_1 \frac{PIB}{Pop} + v \;\Rightarrow\; \frac{CO2}{Pop} = -0{,}123 + 0{,}058 \frac{PIB}{Pop} + \hat{v}$$

Os resultados da ANOVA seriam:

Fonte	*gl*	*SQ*	*QM*	*F*	*p*
Regressão	1	0,96	0,96	20,1	0,004
Resíduos	6	0,29	0,05		
Total	7	1,25			



$$R^2 = 0{,}77$$

O ajuste é significativo. Como se trata de um modelo de RLS, a significância do teste *t* será igual à da estatística *F*.

Correção Multicolinearidade - Exemplo

Sejam as informações sobre emissões de CO^2, PIB e população de 8 países:

O ajuste de MQO nos forneceu os seguintes resultados :

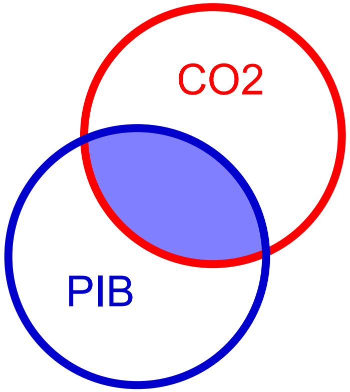
$$CO2 = \beta_0 + \beta_1 PIB + \beta_2 Pop + e \quad \Rightarrow \quad CO2 = 0{,}472 + 0{,}030 PIB + 0{,}028 Pop + \hat{e}$$

Uma solução para eliminar a multicolinearidade seria excluir um dos regressores:

$$CO2 = \beta_0 + \beta_1 PIB + e \quad \Rightarrow \quad CO2 = 0{,}401 + 0{,}035 PIB + \hat{e}$$

Os resultados da ANOVA seriam:

Fonte	*gl*	*SQ*	*QM*	*F*	*p*
Regressão	1	63,77	63,77	20,93	0,004
Resíduos	6	18,28	3,05		
Total	7	82,05			



As estimativas dos dois modelos para a variável PIB diferem apenas apenas ligeiramente. E a relação entre CO2 e PIB passou a ser significativa. Neste caso, a omissão parece não ter ocasionado um viés de omissão muito grave.

Mas a exclusão de variáveis deve ser analisada com muita cautela!