

Multicolinearidade no Modelo de MQO

1 Introdução

Considere o modelo de regressão linear com dois regressores:

$$y_i = \beta_1 + \beta_2 x_{2i} + \beta_3 x_{3i} + u_i, \quad i = 1, \dots, n. \quad (1)$$

A multicolinearidade ocorre quando duas ou mais variáveis explicativas apresentam forte associação linear entre si. Neste exemplo, o problema aparece quando x_2 e x_3 são fortemente correlacionadas.

O objetivo desta nota é mostrar, de forma algébrica, como a correlação entre x_2 e x_3 afeta a matriz de variância-covariância dos estimadores de MQO. A ideia central é que a multicolinearidade não gera viés, desde que as hipóteses clássicas sejam satisfeitas. O seu principal efeito é aumentar a variância dos estimadores, tornando os coeficientes menos precisos.

2 Centralização das variáveis

Defina as variáveis centradas em relação à média amostral:

$$y_i^* = y_i - \bar{y}, \quad x_{2i}^* = x_{2i} - \bar{x}_2, \quad x_{3i}^* = x_{3i} - \bar{x}_3. \quad (2)$$

A média amostral da equação original é:

$$\bar{y} = \beta_1 + \beta_2 \bar{x}_2 + \beta_3 \bar{x}_3 + \bar{u}. \quad (3)$$

Subtraindo essa equação da equação original, temos:

$$y_i - \bar{y} = \beta_2(x_{2i} - \bar{x}_2) + \beta_3(x_{3i} - \bar{x}_3) + (u_i - \bar{u}). \quad (4)$$

Portanto:

$$y_i^* = \beta_2 x_{2i}^* + \beta_3 x_{3i}^* + u_i^*, \quad (5)$$

em que:

$$u_i^* = u_i - \bar{u}. \quad (6)$$

Assim, ao trabalhar com as variáveis centradas, o intercepto desaparece. A regressão passa a ser escrita apenas em termos dos desvios em relação às médias.

3 Forma matricial da regressão centrada

Escreva a regressão centrada em forma matricial:

$$\mathbf{y}^* = \mathbf{X}^* \boldsymbol{\beta} + \mathbf{u}^*, \quad (7)$$

em que:

$$\mathbf{y}^* = \begin{bmatrix} y_1^* \\ y_2^* \\ \vdots \\ y_n^* \end{bmatrix}, \quad \mathbf{X}^* = \begin{bmatrix} x_{21}^* & x_{31}^* \\ x_{22}^* & x_{32}^* \\ \vdots & \vdots \\ x_{2n}^* & x_{3n}^* \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_2 \\ \beta_3 \end{bmatrix}, \quad \mathbf{u}^* = \begin{bmatrix} u_1^* \\ u_2^* \\ \vdots \\ u_n^* \end{bmatrix}. \quad (8)$$

O estimador de MQO para os coeficientes de inclinação é:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^{*'} \mathbf{X}^*)^{-1} \mathbf{X}^{*'} \mathbf{y}^*. \quad (9)$$

Portanto, a existência e a precisão do estimador dependem diretamente da matriz $\mathbf{X}^{*'} \mathbf{X}^*$.

4 A matriz $\mathbf{X}^{*'} \mathbf{X}^*$

A matriz $\mathbf{X}^{*'} \mathbf{X}^*$ é:

$$\mathbf{X}^{*'} \mathbf{X}^* = \begin{bmatrix} \sum_{i=1}^n (x_{2i}^*)^2 & \sum_{i=1}^n x_{2i}^* x_{3i}^* \\ \sum_{i=1}^n x_{2i}^* x_{3i}^* & \sum_{i=1}^n (x_{3i}^*)^2 \end{bmatrix}. \quad (10)$$

Defina:

$$S_{22} = \sum_{i=1}^n (x_{2i}^*)^2, \quad S_{33} = \sum_{i=1}^n (x_{3i}^*)^2, \quad S_{23} = \sum_{i=1}^n x_{2i}^* x_{3i}^*. \quad (11)$$

Então:

$$\mathbf{X}^{*'} \mathbf{X}^* = \begin{bmatrix} S_{22} & S_{23} \\ S_{23} & S_{33} \end{bmatrix}. \quad (12)$$

O termo S_{23} mede a covariação amostral, não normalizada, entre x_2 e x_3 . Quando x_2 e x_3 se movem juntas, S_{23} é grande em valor absoluto.

A correlação amostral entre x_2 e x_3 é dada por:

$$r_{23} = \frac{S_{23}}{\sqrt{S_{22}S_{33}}}. \quad (13)$$

Logo:

$$S_{23} = r_{23} \sqrt{S_{22}S_{33}}. \quad (14)$$

Isso mostra que o elemento fora da diagonal da matriz $\mathbf{X}^{*'} \mathbf{X}^*$ está diretamente ligado à correlação entre os regressores.

5 A matriz de variância-covariância dos estimadores

Sob as hipóteses clássicas do modelo de regressão linear, temos:

$$E(\mathbf{u} \mid \mathbf{X}) = \mathbf{0} \quad (15)$$

e

$$Var(\mathbf{u} \mid \mathbf{X}) = \sigma^2 \mathbf{I}. \quad (16)$$

Nesse caso, a matriz de variância-covariância do estimador de MQO é:

$$Var(\hat{\boldsymbol{\beta}} \mid \mathbf{X}) = \sigma^2 (\mathbf{X}^{*'} \mathbf{X}^*)^{-1}. \quad (17)$$

Como:

$$\mathbf{X}^{*'} \mathbf{X}^* = \begin{bmatrix} S_{22} & S_{23} \\ S_{23} & S_{33} \end{bmatrix}, \quad (18)$$

sua inversa é:

$$(\mathbf{X}^{*'} \mathbf{X}^*)^{-1} = \frac{1}{S_{22}S_{33} - S_{23}^2} \begin{bmatrix} S_{33} & -S_{23} \\ -S_{23} & S_{22} \end{bmatrix}. \quad (19)$$

Portanto:

$$Var(\hat{\boldsymbol{\beta}} \mid \mathbf{X}) = \frac{\sigma^2}{S_{22}S_{33} - S_{23}^2} \begin{bmatrix} S_{33} & -S_{23} \\ -S_{23} & S_{22} \end{bmatrix}. \quad (20)$$

Agora, usando:

$$r_{23} = \frac{S_{23}}{\sqrt{S_{22}S_{33}}}, \quad (21)$$

temos:

$$S_{23}^2 = r_{23}^2 S_{22} S_{33}. \quad (22)$$

Logo:

$$S_{22}S_{33} - S_{23}^2 = S_{22}S_{33} - r_{23}^2 S_{22}S_{33} \quad (23)$$

$$= S_{22}S_{33}(1 - r_{23}^2). \quad (24)$$

Substituindo esse resultado na matriz de variância-covariância, obtemos:

$$Var(\hat{\beta} \mid \mathbf{X}) = \frac{\sigma^2}{S_{22}S_{33}(1 - r_{23}^2)} \begin{bmatrix} S_{33} & -S_{23} \\ -S_{23} & S_{22} \end{bmatrix}. \quad (25)$$

Essa expressão mostra exatamente onde a correlação entre x_2 e x_3 aparece: ela entra no denominador por meio do termo $1 - r_{23}^2$.

6 Variâncias e covariância dos coeficientes

A variância de $\hat{\beta}_2$ é o primeiro elemento da diagonal principal:

$$Var(\hat{\beta}_2 \mid \mathbf{X}) = \frac{\sigma^2 S_{33}}{S_{22}S_{33}(1 - r_{23}^2)}. \quad (26)$$

Simplificando:

$$Var(\hat{\beta}_2 \mid \mathbf{X}) = \frac{\sigma^2}{S_{22}(1 - r_{23}^2)}. \quad (27)$$

De forma análoga, a variância de $\hat{\beta}_3$ é:

$$Var(\hat{\beta}_3 \mid \mathbf{X}) = \frac{\sigma^2 S_{22}}{S_{22}S_{33}(1 - r_{23}^2)}. \quad (28)$$

Logo:

$$Var(\hat{\beta}_3 \mid \mathbf{X}) = \frac{\sigma^2}{S_{33}(1 - r_{23}^2)}. \quad (29)$$

A covariância entre $\hat{\beta}_2$ e $\hat{\beta}_3$ é:

$$Cov(\hat{\beta}_2, \hat{\beta}_3 \mid \mathbf{X}) = -\frac{\sigma^2 S_{23}}{S_{22}S_{33}(1 - r_{23}^2)}. \quad (30)$$

Como:

$$S_{23} = r_{23}\sqrt{S_{22}S_{33}}, \quad (31)$$

então:

$$Cov(\hat{\beta}_2, \hat{\beta}_3 | \mathbf{X}) = -\frac{\sigma^2 r_{23}}{\sqrt{S_{22}S_{33}}(1 - r_{23}^2)}. \quad (32)$$

Portanto, a matriz de variância-covariância pode ser representada como:

$$Var(\hat{\beta} | \mathbf{X}) = \begin{bmatrix} \frac{\sigma^2}{S_{22}(1 - r_{23}^2)} & -\frac{\sigma^2 r_{23}}{\sqrt{S_{22}S_{33}}(1 - r_{23}^2)} \\ -\frac{\sigma^2 r_{23}}{\sqrt{S_{22}S_{33}}(1 - r_{23}^2)} & \frac{\sigma^2}{S_{33}(1 - r_{23}^2)} \end{bmatrix}. \quad (33)$$

7 O fator de inflação da variância

O termo central para entender o efeito da multicolinearidade é:

$$\frac{1}{1 - r_{23}^2}. \quad (34)$$

Esse termo é chamado de fator de inflação da variância, ou *variance inflation factor*:

$$VIF = \frac{1}{1 - r_{23}^2}. \quad (35)$$

No caso com apenas dois regressores, o R^2 da regressão auxiliar de x_2 sobre x_3 é exatamente:

$$R_2^2 = r_{23}^2. \quad (36)$$

Portanto:

$$VIF_2 = \frac{1}{1 - R_2^2} = \frac{1}{1 - r_{23}^2}. \quad (37)$$

De modo equivalente, o R^2 da regressão auxiliar de x_3 sobre x_2 também é:

$$R_3^2 = r_{23}^2, \quad (38)$$

e:

$$VIF_3 = \frac{1}{1 - R_3^2} = \frac{1}{1 - r_{23}^2}. \quad (39)$$

Assim, quanto maior for a correlação entre x_2 e x_3 , maior será r_{23}^2 , menor será $1 - r_{23}^2$, e maior será a variância dos estimadores.

8 Interpretação econômica e estatística

A multicolinearidade não implica, por si só, que o estimador de MQO seja viesado. Sob as hipóteses clássicas, continua valendo:

$$E(\hat{\beta}_2 | \mathbf{X}) = \beta_2 \quad \text{e} \quad E(\hat{\beta}_3 | \mathbf{X}) = \beta_3. \quad (40)$$

O problema é a perda de precisão.

O coeficiente β_2 mede o efeito parcial de x_2 sobre y , mantendo x_3 constante. Porém, se x_2 e x_3 se movem quase juntas, existe pouca variação em x_2 que não esteja também associada a x_3 . Nesse caso, o MQO tem pouca informação para identificar separadamente o efeito de x_2 .

Da mesma forma, β_3 mede o efeito parcial de x_3 sobre y , mantendo x_2 constante. Se x_2 e x_3 são altamente correlacionadas, há pouca variação independente em x_3 . Como consequência, a estimativa de β_3 também se torna imprecisa.

A intuição é simples: quando duas variáveis explicativas carregam quase a mesma informação, o modelo pode até saber que o conjunto dessas variáveis ajuda a explicar y , mas tem dificuldade em separar quanto do efeito deve ser atribuído a x_2 e quanto deve ser atribuído a x_3 .

9 Consequências práticas

A multicolinearidade tende a gerar:

- i) erros-padrão maiores;
- ii) estatísticas t menores;
- iii) intervalos de confiança mais largos;
- iv) coeficientes mais sensíveis a pequenas mudanças na amostra;
- v) possível troca de sinal dos coeficientes;
- vi) dificuldade para rejeitar hipóteses nulas individuais;
- vii) coexistência de alto poder explicativo conjunto com baixa significância individual dos coeficientes.

Portanto, em uma regressão com forte multicolinearidade, pode ocorrer a seguinte situação: o R^2 do modelo é elevado, o teste F conjunto é estatisticamente significativo,

mas os testes t individuais para β_2 e β_3 não são significativos. Isso acontece porque x_2 e x_3 , em conjunto, explicam y , mas o modelo tem dificuldade em separar seus efeitos individuais.

10 Exemplo numérico

Suponha inicialmente que a correlação entre x_2 e x_3 seja:

$$r_{23} = 0,8. \quad (41)$$

Nesse caso:

$$VIF = \frac{1}{1 - r_{23}^2} \quad (42)$$

$$= \frac{1}{1 - 0,8^2} \quad (43)$$

$$= \frac{1}{1 - 0,64} \quad (44)$$

$$= \frac{1}{0,36} \quad (45)$$

$$= 2,78. \quad (46)$$

Isso significa que a variância do estimador é aproximadamente 2,78 vezes maior do que seria se x_2 e x_3 não fossem correlacionadas.

Agora suponha:

$$r_{23} = 0,95. \quad (47)$$

Então:

$$VIF = \frac{1}{1 - r_{23}^2} \quad (48)$$

$$= \frac{1}{1 - 0,95^2} \quad (49)$$

$$= \frac{1}{1 - 0,9025} \quad (50)$$

$$= \frac{1}{0,0975} \quad (51)$$

$$= 10,26. \quad (52)$$

Neste caso, a variância do estimador fica mais de dez vezes maior. Como o erro-

padrão é a raiz quadrada da variância, o erro-padrão aumenta por um fator de:

$$\sqrt{10,26} \approx 3,20. \quad (53)$$

Assim, mesmo que o modelo esteja corretamente especificado, a forte correlação entre os regressores pode tornar os coeficientes individualmente pouco precisos.

11 O caso extremo: multicolinearidade perfeita

Se x_2 e x_3 forem perfeitamente correlacionadas, então:

$$r_{23} = 1 \quad \text{ou} \quad r_{23} = -1. \quad (54)$$

Nesse caso:

$$1 - r_{23}^2 = 0. \quad (55)$$

Logo:

$$S_{22}S_{33} - S_{23}^2 = 0. \quad (56)$$

Isso significa que a matriz $\mathbf{X}^{*'}\mathbf{X}^*$ não é inversível. Portanto, não é possível calcular:

$$(\mathbf{X}^{*'}\mathbf{X}^*)^{-1}. \quad (57)$$

Como o estimador de MQO é:

$$\hat{\beta} = (\mathbf{X}^{*'}\mathbf{X}^*)^{-1}\mathbf{X}^{*'}\mathbf{y}^*, \quad (58)$$

a inexistência da inversa impede a estimação separada de β_2 e β_3 .

Portanto, no caso de multicolinearidade perfeita, o modelo não consegue distinguir o efeito de x_2 do efeito de x_3 . Já no caso de multicolinearidade alta, mas não perfeita, os coeficientes ainda podem ser estimados, porém com variâncias elevadas e baixa precisão estatística.

12 Conclusão

No modelo:

$$y_i = \beta_1 + \beta_2 x_{2i} + \beta_3 x_{3i} + u_i, \quad (59)$$

a centralização das variáveis permite escrever:

$$y_i^* = \beta_2 x_{2i}^* + \beta_3 x_{3i}^* + u_i^*. \quad (60)$$

A matriz relevante para a estimação por MQO é:

$$\mathbf{X}^{*'} \mathbf{X}^* = \begin{bmatrix} S_{22} & S_{23} \\ S_{23} & S_{33} \end{bmatrix}, \quad (61)$$

em que:

$$r_{23} = \frac{S_{23}}{\sqrt{S_{22}S_{33}}}. \quad (62)$$

A matriz de variância-covariância dos estimadores depende do termo:

$$\frac{1}{1 - r_{23}^2}. \quad (63)$$

À medida que r_{23}^2 se aproxima de 1, esse termo cresce rapidamente. Como consequência, as variâncias de $\hat{\beta}_2$ e $\hat{\beta}_3$ aumentam, os erros-padrão ficam maiores e os testes individuais perdem poder.

Assim, a multicolinearidade não implica, por si só, viés nos estimadores de MQO. O problema é que ela reduz a capacidade do pesquisador de identificar separadamente os efeitos individuais de regressores altamente correlacionados.