Econometria Aplicada

Multicolinearidade

João Ricardo Costa Filho

“The most important questions of life are, for the most part, really only problems in probability."

Laplace (1812)

“In God we trust. All others must bring data."

William Edwards Deming

A regressão linear

Motivação (tudo começa com uma pergunta)

Qual é a relação entre as emissões de CO_2 e a população dos países?

Referências

- Como responder a questão que motivou a nossa análise?

Referências

- Como responder a questão que motivou a nossa análise?
  - Econometria: capítulo 3 de Wooldridge (2006).

Variância do Estimador $\hat{\beta}_j$

Regressão múltipla

Consideremos o modelo geral:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + \epsilon_i,$$

onde Y_i é a variável dependente, X_{ji} é o j-ésimo regressor cujo coeficiente populacional é β_j; ϵ_i é o termo de erro, com $E[\epsilon_i|\mathbf{X}] = 0$ e $Var[\epsilon_i|\mathbf{X}] = \sigma^2$.

Teorema de Frisch-Waugh-Lovell

O estimador de MQO para $\hat{\beta}_j$ em um modelo de regressão múltipla é numericamente idêntico ao estimador obtido de uma regressão de Y sobre os **resíduos da regressão de X_j sobre todas as outras variáveis explicativas**.

Teorema de Frisch-Waugh-Lovell

O estimador de MQO para $\hat{\beta}_j$ em um modelo de regressão múltipla é numericamente idêntico ao estimador obtido de uma regressão de Y sobre os **resíduos da regressão de X_j sobre todas as outras variáveis explicativas**.

Sejam u_{ji} os resíduos da regressão auxiliar de X_j sobre as outras variáveis $X_1, \ldots, X_{j-1}, X_{j+1}, \ldots, X_k$:

Teorema de Frisch-Waugh-Lovell

O estimador de MQO para $\hat{\beta}_j$ em um modelo de regressão múltipla é numericamente idêntico ao estimador obtido de uma regressão de Y sobre os **resíduos da regressão de X_j sobre todas as outras variáveis explicativas**.

Sejam u_{ji} os resíduos da regressão auxiliar de X_j sobre as outras variáveis $X_1, \ldots, X_{j-1}, X_{j+1}, \ldots, X_k$:

$$X_j = \alpha_0 + \alpha_1 X_1 + \cdots + \alpha_{j-1} X_{j-1} + \alpha_{j+1} X_{j+1} + \cdots + \alpha_k X_k + u_j,$$

onde u_{ji} é a parte de X_{ji} que não explicada pelas outras variáveis.

Teorema de Frisch-Waugh-Lovell

O estimador $\hat{\beta}_j$ pode ser expresso como:

$$\hat{\beta}_j = \frac{\sum_{i=1}^n u_{ji} Y_i}{\sum_{i=1}^n u_{ji}^2}.$$

Substituindo $Y_i = \beta_0 + \sum_{p=1}^k \beta_p X_{pi} + \epsilon_i$:

$$\hat{\beta}_j = \frac{\sum_{i=1}^n u_{ji} \left(\beta_0 + \sum_{p=1}^k \beta_p X_{pi} + \epsilon_i\right)}{\sum_{i=1}^n u_{ji}^2}.$$

Teorema de Frisch-Waugh-Lovell

Devido à propriedade de ortogonalidade dos resíduos (u_{ji} é não correlacionado com as X_p para $p \neq j$, e $\sum u_{ji} = 0$), muitos termos se cancelam, levando a:

$$\hat{\beta}_j = \beta_j + \frac{\sum_{i=1}^{n} u_{ji}\epsilon_i}{\sum_{i=1}^{n} u_{ji}^2}.$$

Variância amostral de $\hat{\beta}_j$

Agora, calculamos a variância de $\hat{\beta}_j$:

$$Var\left[\hat{\beta}_j\right] = Var\left[\beta_j + \frac{\sum_{i=1}^{n} u_{ji}\epsilon_i}{\sum_{i=1}^{n} u_{ji}^2}\right].$$

Considerando β_j e $\sum u_{ji}^2$ como constantes:

$$Var\left[\hat{\beta}_j\right] = \frac{1}{(\sum_{i=1}^{n} u_{ji}^2)^2} Var\left[\sum_{i=1}^{n} u_{ji}\epsilon_i\right].$$

Variância amostral de $\hat{\beta}_j$

Assumindo $Var(\epsilon_i) = \sigma^2$ e $Cov(\epsilon_i, \epsilon_k) = 0$ para $i \neq k$:

$$Var\left[\sum_{i=1}^{n} u_{ji}\epsilon_i\right] = \sum_{i=1}^{n} u_{ji}^2 Var[\epsilon_i] = \sum_{i=1}^{n} u_{ji}^2\sigma^2.$$

Substituindo de volta, temos:

$$Var\left[\hat{\beta}_j\right] = \frac{1}{(\sum_{i=1}^{n} u_{ji}^2)^2}\left(\sum_{i=1}^{n} u_{ji}^2\sigma^2\right) = \frac{\sigma^2}{\sum_{i=1}^{n} u_{ji}^2}.$$

Variância amostral de $\hat{\beta}_j$

- Lembremos que u_{ji} são os resíduos da regressão de X_j sobre as outras variáveis.

Variância amostral de $\hat{\beta}_j$

- Lembremos que u_{ji} são os resíduos da regressão de X_j sobre as outras variáveis. A soma dos quadrados desses resíduos é a Soma dos Quadrados dos Resíduos (SQR) dessa regressão auxiliar: $SQR_j = \sum_{i=1}^{n} u_{ji}^2$.
- O coeficiente de determinação (R_j^2) para essa regressão auxiliar é definido como: $R_j^2 = 1 - \frac{SQR_j}{SQT_j}$, onde $SQT_j = \sum_{i=1}^{n}(X_{ji} - \bar{X}_j)^2$.
- Rearranjando a fórmula de R_j^2:

$$\frac{SQR_j}{SQT_j} = 1 - R_j^2 \iff SQR_j = SQT_j(1 - R_j^2).$$

Portanto, temos $\sum_{i=1}^{n} u_{ji}^2 = SQT_j(1 - R_j^2)$.

Variância amostral de $\hat{\beta}_j$

Substituindo $\sum_{i=1}^{n} u_{ji}^2$ na expressão da variância de $\hat{\beta}_j$:

$$Var\left[\hat{\beta}_j\right] = \frac{\sigma^2}{SQT_j(1 - R_j^2)}.$$

Variância amostral de $\hat{\beta}_j$

Substituindo $\sum_{i=1}^{n} u_{ji}^2$ na expressão da variância de $\hat{\beta}_j$:

$$Var\left[\hat{\beta}_j\right] = \frac{\sigma^2}{SQT_j(1 - R_j^2)}.$$

Onde:

Variância amostral de $\hat{\beta}_j$

Substituindo $\sum_{i=1}^{n} u_{ji}^2$ na expressão da variância de $\hat{\beta}_j$:

$$Var\left[\hat{\beta}_j\right] = \frac{\sigma^2}{SQT_j(1 - R_j^2)}.$$

Onde:

- σ^2: variância dos **erros**;

Variância amostral de $\hat{\beta}_j$

Substituindo $\sum_{i=1}^{n} u_{ji}^2$ na expressão da variância de $\hat{\beta}_j$:

$$Var\left[\hat{\beta}_j\right] = \frac{\sigma^2}{SQT_j(1 - R_j^2)}.$$

Onde:

- σ^2: variância dos **erros**;
- $SQT_j = \sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2$: variância amostral de x_j;

Variância amostral de $\hat{\beta}_j$

Substituindo $\sum_{i=1}^{n} u_{ji}^2$ na expressão da variância de $\hat{\beta}_j$:

$$Var\left[\hat{\beta}_j\right] = \frac{\sigma^2}{SQT_j(1 - R_j^2)}.$$

Onde:

- σ^2: variância dos **erros**;
- $SQT_j = \sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2$: variância amostral de x_j;
- R_j^2: R^2 da regressão de x_j sobre os demais regressores.

Variância amostral de $\hat{\beta}_j$ e σ^2

"[M]ais 'ruído' da equação (um maior σ^2) torna mais difícil estimar o efeito parcial de qualquer uma das variáveis independentes sobre y, e isso é refletido nas variâncias maiores dos estimadores de incliniação de MQO." [Wooldridge (2006); p. 92]

Variância amostral de $\hat{\beta}_j$ e SQT_j

"[Q]uanto maior a variação total em x_j, menor é $Var(\hat{\beta}_j)$."
[Wooldridge (2006); p. 93]

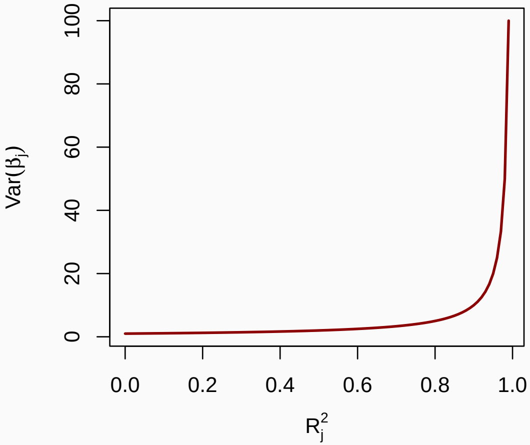
- Note que $\lim_{n\to\infty} SQT_j = \infty \implies \lim_{n\to\infty} Var\left[\hat{\beta}_j\right] = 0$.

Variância amostral de $\hat{\beta}_j$ e o grau de multicolinearidade R_j^2

Se x_j for altamente explicado pelas outras variáveis, temos que:

$$\lim_{R_j^2 \to 1} \text{Var}\left[\hat{\beta}_j\right] = \infty.$$

Variância amostral de $\hat{\beta}_j$ e o grau de multicolinearidade R_j^2



Variância amostral de $\hat{\beta}_j$ e o grau de multicolinearidade R_j^2

Ou seja, quanto maior R_j^2, mais difícil isolar o efeito de x_j em y. Quando não houver relação entre as variáveis,

$$Var\left[\hat{\beta}_j\right] = \frac{\sigma^2}{SQT_j}.$$

O que é multicolinearidade?

A multicolinearidade ocorre quando há correlações lineares fortes entre duas ou mais variáveis explicativas em um modelo de regressão linear.

- **Perfeita**: uma variável é combinação linear exata de outras;
- **Imperfeita**: correlação alta, mas não perfeita.

Principais consequências da multicolinearidade

Estimadores continuam não-viesados, mas:

- Alta variância dos coeficientes;
- Interpretação dos coeficientes comprometida;
- Estatísticas t podem ser insignificantes mesmo com R^2 alto;
  - Colinearidade perfeita: não é possível calcular $\hat{\beta} = (X'X)^{-1}X'Y$.
- Dificuldade para selecionar variáveis relevantes.

Sinais de multicolinearidade

Sinais de multicolinearidade

- R^2 alto, rejeitar a hipótese nula no teste de significância conjunta (teste F), mas não rejeitar a hipótese nula no teste de significância individual (teste t).

Sinais de multicolinearidade

- R^2 alto, rejeitar a hipótese nula no teste de significância conjunta (teste F), mas não rejeitar a hipótese nula no teste de significância individual (teste t).

Wooldridge, Jeffrey M. 2006. *Introdução à Econometria: Uma Abordagem Moderna*. Cengage Learning.