

Análise de Regressão Múltipla

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k + u$$

1. Estimação

Paralelo com Regressão Simples

- β_0 continua sendo o intercepto
- β_1 a β_k são todos chamados de parâmetros de inclinação
- u continua sendo o termo de erro
- Continua sendo necessária a hipótese média condicional zero, mas assumimos que
- $E(u/x_1, x_2, \dots, x_k) = 0$
- Continuamos minimizando a soma do quadrado dos resíduos, mas temos $k+1$ condições de primeira ordem

Exemplos

Equação de salários

$$w = \hat{\beta}_0 + \hat{\beta}_1 educ + \hat{\beta}_2 exper + u$$

Efeito do gasto público sobre a nota média

$$notamed = \hat{\beta}_0 + \hat{\beta}_1 gasto + \hat{\beta}_2 rendfam + u$$

Exemplos

Consumo das famílias como função quadrática da renda

$$C = \hat{\beta}_0 + \hat{\beta}_1 \text{renda} + \hat{\beta}_2 \text{renda}^2 + u$$

Efeito marginal da renda

$$\frac{\partial \text{cons}}{\partial \text{rend}} \approx \hat{\beta}_1 + 2\hat{\beta}_2 \text{renda}$$

Interpretando Regressão Múltipla

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \dots + \hat{\beta}_k x_k, \text{ então}$$

$$\Delta \hat{y} = \Delta \hat{\beta}_1 x_1 + \Delta \hat{\beta}_2 x_2 + \dots + \Delta \hat{\beta}_k x_k,$$

então mantendo x_2, \dots, x_k fixos implica que

$\Delta \hat{y} = \Delta \hat{\beta}_1 x_1$, isto significa que cada β tem
uma interpretação *ceteris paribus*

Outra interpretação

Considere o caso onde $k = 2$, i.e.

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2, \text{ então}$$

$$\hat{\beta}_1 = \left(\sum \hat{r}_{i1} y_i \right) / \sum \hat{r}_{i1}^2, \text{ tal que } \hat{r}_{i1} \text{ são}$$

os resíduos da regressão estimada

$$\hat{x}_1 = \hat{\gamma}_0 + \hat{\gamma}_2 \hat{x}_2$$

Outra Interpretação

- A equação anterior implica que regredindo y sobre x_1 e x_2 fornece o mesmo efeito de x_1 como regredindo y sobre os resíduos de uma regressão de x_1 sobre x_2
- Isto significa apenas a parte de x_{i1} que não é correlacionada com x_{i2} estão sendo relacionadas com y_i então estamos estimando o efeito de x_1 sobre y após x_2 ter sido eliminada

Estimativa Simples vs Múltipla

Compare a regressão simples $\tilde{y} = \tilde{\beta}_0 + \tilde{\beta}_1 x_1$
com a regressão múltipla $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$

Genericamente, $\tilde{\beta}_1 \neq \hat{\beta}_1$ ao menos que:

$\hat{\beta}_2 = 0$ (i.e. não há efeito parcial de x_2) ou
 x_1 e x_2 são não correlacionados na amostra

Grau de ajuste

Nós podemos pensar que cada observação é feita de uma parte explicada, e outra não-explicada,

$y_i = \hat{y}_i + \hat{u}_i$ Podemos definir o seguinte:

$\sum (y_i - \bar{y})^2$ é a soma total dos quadrados (SST)

$\sum (\hat{y}_i - \bar{y})^2$ é a soma explicada dos quadrados (SSE)

$\sum \hat{u}_i^2$ é o resíduo da soma dos quadrados (SSR)

Então $SST = SSE + SSR$

Grau de ajuste

- ◆ Como podemos pensar em como a reta de regressão se ajusta ao dados da nossa amostra?
- ◆ Podemos computar a fração da soma total dos quadros (SST) que é explicada pelo modelo, a que chamamos de R-quadrado da regressão
- ◆ $R^2 = SSE/SST = 1 - SSR/SST$

Grau de Ajuste

Também podemos pensar no R^2 como sendo igual ao quadrado do coeficiente de correlação entre o y_i real e os valores preditos \hat{y}_i

$$R^2 = \frac{\left(\sum (y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}}) \right)^2}{\left(\sum (y_i - \bar{y})^2 \right) \left(\sum (\hat{y}_i - \bar{\hat{y}})^2 \right)}$$

R -quadrado

- R^2 nunca pode diminuir quando outra variável independente é adicionada a uma regressão, na verdade costuma aumentar
- Porque o R^2 irá aumentar com o número de variáveis independentes, ele não é uma boa forma de se comparar modelos

Hipóteses para Inexistência de Viés

- ◆ O modelo populacional é linear nos parâmetros: $y = \beta_0 + \beta_1x_1 + \beta_2x_2 + \dots + \beta_kx_k + u$
- ◆ Podemos usar uma amostra aleatória de tamanho n , $\{(x_{i1}, x_{i2}, \dots, x_{ik}, y_i): i=1, 2, \dots, n\}$, do modelo populacional, tal que o modelo seja $y_i = \beta_0 + \beta_1x_{i1} + \beta_2x_{i2} + \dots + \beta_kx_{ik} + u_i$
- ◆ $E(u/x_1, x_2, \dots, x_k) = 0$, implicando que todas as variáveis explicativas são exógenas
- ◆ Nenhum dos x 's são constantes, e não há dependência linear entre eles

Muitas ou poucas variáveis

- O que acontece se nós incluimos variáveis em nossa especificação que não pertence ao modelo?
- Não há impacto em nossa estimativa dos parâmetros e, o estimador OLS permanece não-viesado
- O que ocorre se excluirmos uma variável que pertence ao modelo verdadeiro?
- OLS será, em geral, viesado

Viés de Variável Omitida

Suponha que o modelo verdadeiro
seja $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$, mas
estimamos $\tilde{y} = \tilde{\beta}_0 + \tilde{\beta}_1 x_1 + u$, então

$$\tilde{\beta}_1 = \frac{\sum (x_{i1} - \bar{x}_1) y_i}{\sum (x_{i1} - \bar{x}_1)^2}$$

Viés de Variável Omitida

Relembre que o modelo verdadeiro é

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + u_i, \text{ então}$$

o numerador se torna

$$\sum (x_{i1} - \bar{x}_1)(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + u_i) =$$

$$\beta_1 \sum (x_{i1} - \bar{x}_1)^2 + \beta_2 \sum (x_{i1} - \bar{x}_1)x_{i2} + \sum (x_{i1} - \bar{x}_1)u_i$$

Viés de Variável Omitida

$$\tilde{\beta} = \beta_1 + \beta_2 \frac{\sum (x_{i1} - \bar{x}_1) x_{i2}}{\sum ((x_{i1} - \bar{x}_1)^2)} + \frac{\sum (x_{i1} - \bar{x}_1) u_i}{\sum ((x_{i1} - \bar{x}_1)^2)}$$

dado que $E(u_i) = 0$, tomando expectativas temos:

$$E(\tilde{\beta}_1) = \beta_1 + \beta_2 \frac{\sum (x_{i1} - \bar{x}_1) x_{i2}}{\sum ((x_{i1} - \bar{x}_1)^2)}$$

Viés de Variável Omitida

Considere a regressão de x_2 sobre x_1

$$\tilde{x}_2 = \tilde{\delta}_0 + \tilde{\delta}_1 x_1 \quad \text{então} \quad \tilde{\delta}_1 = \frac{\sum (x_{i1} - \bar{x}_1) x_{i2}}{\sum ((x_{i1} - \bar{x}_1)^2)}$$

$$\text{então } E(\tilde{\beta}_1) = \beta_1 + \beta_2 \tilde{\delta}_1$$

Sumário do Viés

	$\text{Corr}(x_1, x_2) > 0$	$\text{Corr}(x_1, x_2) < 0$
$\beta_2 > 0$	Viés Positivo	Negativo
$\beta_2 < 0$	Negativo	Positivo

Sumário Viés de Variável Omitida

- Dois casos onde o viés é igual a zero
 - $\beta_2 = 0$, isto é x_2 não pertence realmente ao modelo
 - x_1 e x_2 não são correlacionados na amostra
- Se a correlação entre x_2 , x_1 e x_2 , y tem a mesma direção, o viés será positivo
- Se a correlação entre x_2 , x_1 e x_2 , y tem direção oposta, o viés será negativo

O Caso Geral

- Tecnicamente, podemos apenas sinalizar o vies para o caso geral se todos os x 's incluídos são não-correlacionados
- Tipicamente, interpretamos o viés assumindo que os x 's não são correlacionados, isto serve como um bom guia do viés mesmo que esta hipótese não seja verdadeira.

Viés de omissão

- Vamos supor o exemplo com o seguinte modelo

$$\textit{salar}ioh = \beta_0 + \beta_1 \textit{educ} + \beta_2 \textit{exper} + \beta_3 \textit{aptid} + u$$

Variância dos Estimadores OLS

- ◆ Agora nós sabemos que a distribuição amostral de nossa estimativa é centrada em torno do parâmetro verdadeiro
- ◆ Desejamos pensar sobre o quanto esta distribuição está “espalhada”
- ◆ É muito fácil pensar sobre esta variância sob a hipótese adicional:
- ◆ $\text{Var}(u/x_1, x_2, \dots, x_k) = \sigma^2$
(Homocedasticidade)

Variância do OLS

- Faça \mathbf{x} representar (x_1, x_2, \dots, x_k)
- Assumindo que $\text{Var}(u|\mathbf{x}) = \sigma^2$ também implica que $\text{Var}(y|\mathbf{x}) = \sigma^2$
- As 4 hipóteses para inexistência de viés, mais esta hipótese de homocedasticidade são conhecidas como as hipóteses de ***Gauss-Markov***

Variância do OLS

Dadas as hipóteses Gauss-Markov

$$\text{Var}(\hat{\beta}_j) = \frac{\sigma^2}{SST_j (1 - R_j^2)}, \text{ onde}$$

$$SST_j = \sum (x_{ij} - \bar{x}_j)^2 \text{ e } R_j^2 \text{ é o } R^2$$

de regredir x_j sobre todos os outros x 's

Componentes das Variâncias OLS

- A variância do erro: um σ^2 maior implica variância maior para os estimadores OLS
- A variação amostral total: um SST_j maior implica em uma variância menor dos estimadores
- Relações lineares entre as variáveis independentes: um R_j^2 maior implica em uma variância maior dos estimadores

Modelos Mal Especificados

Considere novamente o modelo com especificação errada

$$\tilde{y} = \tilde{\beta}_0 + \tilde{\beta}_1 x_1, \text{ tal que } Var(\tilde{\beta}_1) = \frac{\sigma^2}{SST_1}$$

Portanto, $Var(\tilde{\beta}_1) < Var(\hat{\beta}_1)$ ao menos que

x_1 e x_2 sejam não correlacionados, então eles serão os mesmos

Modelos Mal Especificados

- Enquanto a variância do estimador é menor para o modelo mal-especificado, ao menos que $\beta_2 = 0$ o modelo mal-especificado é viesado
- A medida que a amostra cresce, a variância de cada estimador se aproxima de zero, fazendo a diferença da variância menos importante

Estimando a Variância do Erro

- ◆ Nós não sabemos qual é a variância do erro, σ^2 , pois não observamos o erro, u_i
- ◆ O que observamos são os resíduos, \hat{u}_i
- ◆ Nós podemos usar os resíduos para formar uma estimativa da variância do erro

Estimando a Variância do Erro

$$\hat{\sigma}^2 = \left(\sum \hat{u}_i^2 \right) / (n - k - 1) \equiv SSR / df$$

portanto, $se(\hat{\beta}_j) = \hat{\sigma} / \left[SST_j (1 - R_j^2) \right]^{1/2}$

- $df = n - (k + 1)$, ou $df = n - k - 1$
- df (i.e. graus de liberdade) é o (número de observações) – (número de parâmetros estimados)

O Teorema Gauss-Markov

- Dadas nossas 5 hipóteses Gauss-Markov, podemos mostrar que o estimador OLS é “**BLUE**”
- **B**est (melhor)
- **L**inear (linear)
- **U**nbiased (não-viesado)
- **E**stimator (estimador)
- Portanto, se as hipóteses valem, então use OLS