Insper | Instituto de Ensino e Pesquisa

Análise de Regressão Linear Múltipla III

Aula 06

Heij et al., 2004 – Capítulo 3

Suposições e Propriedades

Suposições e Propriedades

MLR.1 – O modelo de regressão é linear nos parâmetros

O modelo na população pode ser escrito como

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \varepsilon$$

em que

β_0, β_1, ..., β_k – são parâmetros desconhecidos (constantes);

ε – termo de erro aleatório não observável.

Suposições e Propriedades

MLR.2 – Amostragem Aleatória

Temos uma amostra aleatória de *n* observações

$$(x_{1i}, x_{2i}, ..., x_{ki}, y_i), \quad i = 1, 2, ..., n,$$

do modelo populacional descrito em MLR.1.

MLR.3 – Ausência de Colinearidade Perfeita

Na amostra (e, portanto, na população) nenhum regressor é constante e não há relação linear <u>PERFEITA</u> entre os regressores (a matriz X apresenta posto completo).

Suposições e Propriedades

MLR.4 – Média Condicional Zero

O valor esperado do vetor de erro aleatório, ε, condicionado na matriz de explicação X, é igual a zero.

Ou seja,

$$E(\varepsilon \mid X) = 0.$$

Teorema 1. Sob as suposições MLR.1 a MLR.4, condicionado nos valores do regressores, os estimadores de MQO para os parâmetros do modelo de regressão múltipla são não-viesados, ou seja, $E(\hat{\beta}_j) = \beta_j$, j = 0, 1, 2, ..., k.

Demonstração do Teorema 1

$$(i)\quad E\left(\hat{\underset{\sim}{\boldsymbol{\beta}}} \mid \underset{\sim}{\mathbf{X}}\right) = E\left[\left(\underset{\sim}{\mathbf{X}}'\underset{\sim}{\mathbf{X}}\right)^{-1}\underset{\sim}{\mathbf{X}}'\underset{\sim}{\mathbf{y}} \mid \underset{\sim}{\mathbf{X}}\right] = E\left[\left(\underset{\sim}{\mathbf{X}}'X\right)^{-1}\underset{\sim}{\mathbf{X}}'\left(\underset{\sim}{\mathbf{X}}\underset{\sim}{\boldsymbol{\beta}} + \underset{\sim}{\boldsymbol{\varepsilon}}\right) \mid \underset{\sim}{\mathbf{X}}\right] =$$

$$= E\left[\left(\underset{\sim}{\mathbf{X}}'\underset{\sim}{\mathbf{X}}\right)^{-1}\underset{\sim}{\mathbf{X}}'\underset{\sim}{\mathbf{X}}\underset{\sim}{\boldsymbol{\beta}} + \left(\underset{\sim}{\mathbf{X}}'\underset{\sim}{\mathbf{X}}\right)^{-1}\underset{\sim}{\mathbf{X}}'\underset{\sim}{\boldsymbol{\varepsilon}} \mid \underset{\sim}{\mathbf{X}}\right] =$$

$$= E\left[\underset{\sim}{\boldsymbol{\beta}} + \left(\underset{\sim}{\mathbf{X}}'\underset{\sim}{\mathbf{X}}\right)^{-1}\underset{\sim}{\mathbf{X}}'\underset{\sim}{\boldsymbol{\varepsilon}} \mid \underset{\sim}{\mathbf{X}}\right] = E\left[\underset{\sim}{\boldsymbol{\beta}} \mid \underset{\sim}{\mathbf{X}}\right] + E\left[\left(\underset{\sim}{\mathbf{X}}'\underset{\sim}{\mathbf{X}}\right)^{-1}\underset{\sim}{\mathbf{X}}'\underset{\sim}{\boldsymbol{\varepsilon}} \mid \underset{\sim}{\mathbf{X}}\right] =$$

$$= \underset{\sim}{\boldsymbol{\beta}} + \left(\underset{\sim}{\mathbf{X}}'\underset{\sim}{\mathbf{X}}\right)^{-1}\underset{\sim}{\mathbf{X}}'E\left[\underset{\sim}{\boldsymbol{\varepsilon}} \mid \underset{\sim}{\mathbf{X}}\right] = \underset{\sim}{\boldsymbol{\beta}}$$

$$(ii)\quad E\left(\hat{\underset{\sim}{\boldsymbol{\beta}}}\right) \overset{\text{Lei das Expectativas Iteradas}}{=} E\left[E\left(\hat{\underset{\sim}{\boldsymbol{\beta}}} \mid \underset{\sim}{\mathbf{X}}\right)\right] \overset{de\,(i)}{=} E\left[\underset{\sim}{\boldsymbol{\beta}}\right] = \underset{\sim}{\boldsymbol{\beta}}$$

Suposições e Propriedades

SUPOSIÇÃO FUNDAMENTAL:

$$E(\varepsilon \mid X) = 0.$$

Ou seja, todos os fatores contidos em ε devem ser não correlacionados com as variáveis explicativas, e deve ter sido usada a forma funcional correta.

Suposições e Propriedades

SUPOSIÇÃO FUNDAMENTAL: (cont)

Como pode falhar?

- ***Omissão de variável explicativa importante, correlacionada com*** x_1, x_2, ***... ou*** x_k;
- ***Forma funcional especificada incorretamente;***
- ***Erro de medida em*** x_1, x_2, ***... ou*** x_k;
- ***Simultaneidade entre*** y ***e*** x_1, x_2, ***...ou*** x_k;

Inclusão e Exclusão de Regressores

ANÁLISE DE DOIS CASOS ESPECIAIS:

A) Inclusão de variável irrelevante

$\Rightarrow$ **não prejudica a propriedade de ausência de viés**

B) Omissão de variável relevante

$\Rightarrow$ **modelo correto tem *k* = 2, mas usamos *k* = 1**

Resultado:

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

Inclusão e Exclusão de Regressores

Direção do Viés

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

Inclusão e Exclusão de Regressores

Observações

⇒ **viés depende tanto dos sinais quanto das magnitudes;**

⇒ **em geral, se $k > 1$, omissão de qualquer variável relevante faz com que todos os estimadores de mínimos quadrados sejam viesados;**

⇒ **a menos que a variável omitida seja irrelevante ou não-correlacionada com as demais variáveis explicativas presentes no modelo, os estimadores de mínimos quadrados serão viesados.**

Suposições e Propriedades

MLR.5 – Homocedasticidade

A variância do vetor de erro aleatório, condicional na matriz de explicação, é diagonal (com todos os elementos da diagonal iguais a σ^2).

Ou seja,

$$Var\left(\underset{\sim}{\boldsymbol{\varepsilon}} \mid \underset{\sim}{\mathbf{X}}\right) = E\left(\underset{\sim}{\boldsymbol{\varepsilon}}\underset{\sim}{\boldsymbol{\varepsilon}}' \mid \underset{\sim}{\mathbf{X}}\right) = \begin{pmatrix} \sigma^2 & 0 & 0 & \dots & 0 \\ 0 & \sigma^2 & 0 & \dots & 0 \\ 0 & 0 & \sigma^2 & & \vdots \\ 0 & 0 & 0 & \ddots & 0 \\ 0 & 0 & \dots & 0 & \sigma^2 \end{pmatrix} = \sigma^2 \underset{\sim}{\mathbf{I}_n}$$

(matriz de variâncias e covariâncias associada ao vetor de erros)

Observação 1

As suposições MLR.1 a MLR.5 conjuntamente são conhecidas como suposições de Gauss-Markov.

Sob as suposições MLR.1 a MLR.5:

$$(iii)\,Var\left(\underset{\sim}{\hat{\boldsymbol{\beta}}}\,|\,\underset{\sim}{\mathbf{X}}\right)=Var\left[\left(\underset{\sim}{\mathbf{X}}'\underset{\sim}{\mathbf{X}}\right)^{-1}\underset{\sim}{\mathbf{X}}'\underset{\sim}{\mathbf{y}}\,|\,\underset{\sim}{\mathbf{X}}\right]=Var\left[\left(\underset{\sim}{\mathbf{X}}'\underset{\sim}{\mathbf{X}}\right)^{-1}\underset{\sim}{\mathbf{X}}'\left(\underset{\sim}{\mathbf{X}}\underset{\sim}{\boldsymbol{\beta}}+\underset{\sim}{\boldsymbol{\varepsilon}}\right)|\,\underset{\sim}{\mathbf{X}}\right]=$$

$$=Var\left[\left(\underset{\sim}{\mathbf{X}}'\underset{\sim}{\mathbf{X}}\right)^{-1}\underset{\sim}{\mathbf{X}}'\underset{\sim}{\mathbf{X}}\underset{\sim}{\boldsymbol{\beta}}+\left(\underset{\sim}{\mathbf{X}}'\underset{\sim}{\mathbf{X}}\right)^{-1}\underset{\sim}{\mathbf{X}}'\underset{\sim}{\boldsymbol{\varepsilon}}\,|\,\underset{\sim}{\mathbf{X}}\right]=Var\left[\underset{\sim}{\boldsymbol{\beta}}+\left(\underset{\sim}{\mathbf{X}}'\underset{\sim}{\mathbf{X}}\right)^{-1}\underset{\sim}{\mathbf{X}}'\underset{\sim}{\boldsymbol{\varepsilon}}\,|\,\underset{\sim}{\mathbf{X}}\right]=$$

$$=Var\left[\left(\underset{\sim}{\mathbf{X}}'\underset{\sim}{\mathbf{X}}\right)^{-1}\underset{\sim}{\mathbf{X}}'\underset{\sim}{\boldsymbol{\varepsilon}}\,|\,\underset{\sim}{\mathbf{X}}\right]=\left(\underset{\sim}{\mathbf{X}}'\underset{\sim}{\mathbf{X}}\right)^{-1}\underset{\sim}{\mathbf{X}}'Var\left[\underset{\sim}{\boldsymbol{\varepsilon}}\,|\,\underset{\sim}{\mathbf{X}}\right]\left(\left(\underset{\sim}{\mathbf{X}}'\underset{\sim}{\mathbf{X}}\right)^{-1}\underset{\sim}{\mathbf{X}}'\right)'=$$

$$=\left(\underset{\sim}{\mathbf{X}}'\underset{\sim}{\mathbf{X}}\right)^{-1}\underset{\sim}{\mathbf{X}}'\sigma^2\underset{\sim}{\mathbf{I}_\mathbf{n}}\underset{\sim}{\mathbf{X}}\left(\underset{\sim}{\mathbf{X}}'\underset{\sim}{\mathbf{X}}\right)^{-1}=\sigma^2\left(\underset{\sim}{\mathbf{X}}'\underset{\sim}{\mathbf{X}}\right)^{-1}\underset{\sim}{\mathbf{X}}'\underset{\sim}{\mathbf{I}_\mathbf{n}}\underset{\sim}{\mathbf{X}}\left(\underset{\sim}{\mathbf{X}}'\underset{\sim}{\mathbf{X}}\right)^{-1}=$$

$$=\sigma^2\left(\underset{\sim}{\mathbf{X}}'\underset{\sim}{\mathbf{X}}\right)^{-1}$$

Observação 2

De *(iii)*, se

$$\left(\underset{\sim}{\mathbf{X}}'\underset{\sim}{\mathbf{X}}\right)^{-1} = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1k} \\ a_{21} & a_{22} & & a_{2k} \\ \vdots & & \ddots & \\ a_{k1} & a_{k2} & & a_{kk} \end{pmatrix}$$

então

$$Var(\hat{\beta}_j) = a_{jj}\sigma^2 \quad \textbf{e} \quad Cov(\hat{\beta}_i, \hat{\beta}_j) = a_{ij}\sigma^2$$

Variância dos Estimadores de MQO

Teorema 2. Sob as suposições MLR.1 a MLR.5, condicionadas aos valores amostrais das variáveis explicativas

$$Var(\hat{\beta}_j) = \frac{\sigma^2}{SQT_{x_j}(1 - R_{x_j}^2)}, \quad j = 0, 1, ..., k$$

em que

σ^2 = variância do erro;

SQT_{x_j} = SQT do *j-ésimo* regressor na amostra;

$R_{x_j}^2$ = R^2 da regressão de x_j contra todas as outras variáveis explicativas (incluindo um intercepto).

Componentes da Variância dos Estimadores de Mínimos Quadrados

$$Var(\hat{\beta}_j) = \frac{\sigma^2}{SQT_{x_j}(1 - R^2_{x_j})}$$

$\Rightarrow$ **<u>Variância da v.a. *u*</u>: σ^2 alto implica num estimador de mínimos quadrados com alta variância;**

$\Rightarrow$ ***SQT_{x_j}* : se a *j-ésima* variável explicativa apresentar uma variação total alta, então, a variância do *i-ésimo* estimador, associado à esta variável explicativa, será pequena;**

Componentes da Variância dos Estimadores de Mínimos Quadrados

$$Var(\hat{\beta}_j) = \frac{\sigma^2}{SQT_{x_j}(1 - R^2_{x_j})}$$

⇒ **Relações lineares entre as variáveis explicativas: altos valores de $R_{x_j}^2$ implicam numa alta variância para os estimadores.**

⇒ **$1/(1-R_{x_j}^2)$ – conhecido como fator de inflação de variância ou, *VIF*, em inglês.**

⇒ **Inclusão de variável irrelevante geralmente aumenta as variâncias dos demais estimadores de MQO**

Estimação de σ^2

Como σ^2 em geral é desconhecida, utilizaremos o estimador:

$$\boxed{\hat{\sigma}^2 = MSR = \frac{SSR}{n-(k+1)}}$$

MSR (Quadrado Médio devido aos Resíduos)

SSR perde k+1 graus de liberdade, devido às k+1 restrições impostas pelas condições de primeira ordem de MQO.

Estimação de σ^2

Teorema 3. **Sob as suposições de Gauss-Markov (MLR.1 a MLR.5),**

$$E(\hat{\sigma}^2) = E(MSR) = \sigma^2.$$

Observação

$\hat{\sigma} = \sqrt{MSR}$ **: erro padrão da regressão.**

Erro Padrão dos Estimadores de MQO

Dessa forma, o erro-padrão dos estimadores de mínimos quadrados podem ser obtidos através da expressão

$$\hat{\sigma}_{\hat{\beta}_j} = \sqrt{\frac{\hat{\sigma}^2}{SQT_{x_j}(1 - R_{x_j}^2)}}$$

Eficiência dos Estimadores de MQO

Teorema 4. (TEOREMA DE GAUSS-MARKOV)

Sob as suposições MLR.1 a MLR.5,

$$\hat{\beta}_0,\ \hat{\beta}_1,\ \ldots,\ \hat{\beta}_k$$

são os melhores estimadores, na classe dos lineares não-viesados (BLUE) para β_0, β_1, ..., β_k, respectivamente.

Eficiência dos Estimadores de MQO

⇒ **<u>Restringindo</u> a classe de estimadores não viesados a todos os estimadores lineares em y, o teorema de Gauss-Markov prova que o estimador de mínimos quadrados é o "melhor" (no sentido em que apresenta variância mínima)**

⇒ **Diz-se que, sob as suposições MLR.1 a MLR.5, os estimadores de mínimos quadrados são BLUEs (*best linear unbiased estimators*)**

Suposições e Propriedades

MLR.6 – O vetor de erro estocástico ε é independente dos regressores e segue uma distribuição normal multivariada, com vetor de médias igual a zero e matriz de variâncias e covariâncias igual a $\sigma^2 \underset{\sim}{\mathbf{I}_n}$.

Suposições e Propriedades

Observações

1) Para aplicações de regressão com dados do tipo *cross-sectional*, as suposições MLR.1 a MLR.6 são conhecidas como suposições do modelo linear clássico (suposições CLM).

2) Uma maneira sucinta de resumir as suposições CLM na população é

$$y \mid (x_1, x_2, \ldots, x_k) \sim N(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k;\ \sigma^2).$$

3) Sob as suposições CLM os estimadores de mínimos quadrados são estimadores não-viesados de variância mínima.

Propriedades dos Estimadores

(iv) Sob as suposições clássicas do modelo de regressão linear e, também, sabendo que $\hat{\underset{\sim}{\boldsymbol\beta}}$ *é linear em* $\underset{\sim}{\mathbf{y}}$ ***temos que***

$$\hat{\underset{\sim}{\boldsymbol\beta}} \sim N_k\left(\underset{\sim}{\boldsymbol\beta};\ \sigma^2\left(\underset{\sim}{\mathbf{X}}'\underset{\sim}{\mathbf{X}}\right)^{-1}\right)$$

Observação: O vetor de estimadores é normalmente distribuído devido ao fato de ser formado por uma combinação linear dos elementos do vetor resposta, que são normais e independentes, uma vez que os erros assim o são.

Propriedades dos Estimadores

Desta maneira, cada um dos componentes de $\hat{\underset{\sim}{\boldsymbol{\beta}}}$, tem a seguinte distribuição

$$\hat{\beta}_j \sim N\left(\beta_j;\ \sigma^2 a_{jj}\right),$$

em que

a_{jj} é o *j-ésimo* elemento da diagonal da matriz $\left(\underset{\sim}{\mathbf{X}}'\underset{\sim}{\mathbf{X}}\right)^{-1}$.

Distribuição amostral de $\hat{\beta}_j$

Teorema 4.1 – Sob as suposições CLM (MLR.1 a MLR.6), condicionado nos valores amostrais das variáveis explicativas,

$$\hat{\beta}_j \sim N\left(\beta_j ;\ \frac{\sigma^2}{SQT_{x_j}(1-R_{x_j}^2)} \right)$$

Distribuição amostral de $\hat{\beta}_j$

Do teorema anterior segue que,

$$\frac{\hat{\beta}_j - \beta_j}{\sqrt{\dfrac{\sigma^2}{SQT_{x_j}(1-R_{x_j}^2)}}} \sim N(0;1)$$

Como σ^2 é um parâmetro desconhecido, então será proposto um estimador para tal parâmetro. Dessa maneira, será necessário estudar a distribuição de probabilidades da nova v.a. que será gerada.