Econometria Aplicada

Regressão linear simples

João Ricardo Costa Filho

Econometria Aplicada

O que é Econometria

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

Laplace (1812)

O que é Econometria

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

William Edwards Deming

O que é Econometria

"A econometria é baseada no desenvolvimento de
métodos estatísticos para estimar relações
econômicas,

O que é Econometria

"A econometria é baseada no desenvolvimento de métodos estatísticos para estimar relações econômicas, testar teorias,

O que é Econometria

"A econometria é baseada no desenvolvimento de métodos estatísticos para estimar relações econômicas, testar teorias, avaliar

O que é Econometria

"A econometria é baseada no desenvolvimento de métodos estatísticos para estimar relações econômicas, testar teorias, avaliar e implementar políticas de governo e de negócios." (Wooldridge, 2006, p. 1.)

Tipos de dados

A econometria foca em problemas com análise de dados **não-experimentais/observacionais** (Wooldridge 2006).

Tipos de dados

A econometria foca em problemas com análise de dados **não-experimentais/observacionais** (Wooldridge 2006).

- Dados de corte transversal (*cross-section*).

Tipos de dados

A econometria foca em problemas com análise de dados **não-experimentais/observacionais** (Wooldridge 2006).

- Dados de corte transversal (*cross-section*).
    - É possível ter observações em diferentes períodos de tempo (coleta em dias ou semanas diferentes, por exemplo): neste caso, ignora-se a dimensão tempo (Wooldridge 2006, 5.).

Tipos de dados

A econometria foca em problemas com análise de dados **não-experimentais/observacionais** (Wooldridge 2006).

- Dados de corte transversal (*cross-section*).
    - É possível ter observações em diferentes períodos de tempo (coleta em dias ou semanas diferentes, por exemplo): neste caso, ignora-se a dimensão tempo (Wooldridge 2006, 5.).
- Série de tempo.

Tipos de dados

A econometria foca em problemas com análise de dados **não-experimentais/observacionais** (Wooldridge 2006).

- Dados de corte transversal (*cross-section*).
  - É possível ter observações em diferentes períodos de tempo (coleta em dias ou semanas diferentes, por exemplo): neste caso, ignora-se a dimensão tempo (Wooldridge 2006, 5.).
- Série de tempo.
- Cortes transversais agrupados: diferentes indivíduos em diferentes instantes do tempo.

Tipos de dados

A econometria foca em problemas com análise de dados **não-experimentais/observacionais** (Wooldridge 2006).

- Dados de corte transversal (*cross-section*).
  - É possível ter observações em diferentes períodos de tempo (coleta em dias ou semanas diferentes, por exemplo): neste caso, ignora-se a dimensão tempo (Wooldridge 2006, 5.).
- Série de tempo.
- Cortes transversais agrupados: diferentes indivíduos em diferentes instantes do tempo.
  - Agrupam-se cortes transversais para aumentar o tamanho da amostra.

Tipos de dados

A econometria foca em problemas com análise de dados **não-experimentais/observacionais** (Wooldridge 2006).

- Dados de corte transversal (*cross-section*).
  - É possível ter observações em diferentes períodos de tempo (coleta em dias ou semanas diferentes, por exemplo): neste caso, ignora-se a dimensão tempo (Wooldridge 2006, 5.).
- Série de tempo.
- Cortes transversais agrupados: diferentes indivíduos em diferentes instantes do tempo.
  - Agrupam-se cortes transversais para aumentar o tamanho da amostra.
- Dados em painel (ou longitudinais).

Ferramental

Ferramental

- Linguagem: R

Ferramental

- Linguagem: R
- Como?

Ferramental

- Linguagem: R
- Como?
  - RStudio

Ferramental

- Linguagem: R
- Como?
  - RStudio
  - Google Colab: https://colab.research.google.com/#create=true&language=r

Ferramental

- Linguagem: R
- Como?
  - RStudio
  - Google Colab: https://colab.research.google.com/#create=true&language=r
  - R Studio online

Vamos ao Google Colab.

A regressão linear

Motivação (tudo começa com uma pergunta)

Como a distância entre países afeta as trocas comerciais deles?

Vamos coletar os dados e "olhar" para eles.

Por que visualizar os dados é tão importante assim?

O quarteto de Anscombe

O quarteto de Anscombe (Anscombe 1973)

Imagine quatro conjuntos de dados.

O quarteto de Anscombe (Anscombe 1973)

Imagine quatro conjuntos de dados. Em cada um deles, temos duas variáveis (X e Y). Em todos, temos...

O quarteto de Anscombe (Anscombe 1973)

Imagine quatro conjuntos de dados. Em cada um deles, temos duas variáveis (X e Y). Em todos, temos...

- ... a mesma média e o mesmo desvio-padrão de X.

O quarteto de Anscombe (Anscombe 1973)

Imagine quatro conjuntos de dados. Em cada um deles, temos duas variáveis (X e Y). Em todos, temos...

- ... a mesma média e o mesmo desvio-padrão de X.
- ... a mesma média e o mesmo desvio-padrão de Y.

O quarteto de Anscombe (Anscombe 1973)

Imagine quatro conjuntos de dados. Em cada um deles, temos duas variáveis (X e Y). Em todos, temos...

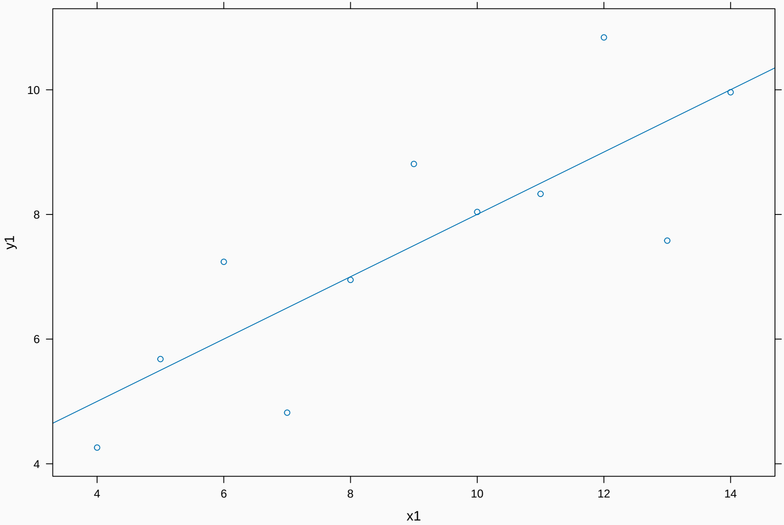
- ...a mesma média e o mesmo desvio-padrão de X.
- ...a mesma média e o mesmo desvio-padrão de Y.
- ...a mesma correlação entre X e Y.

O quarteto de Anscombe (Anscombe 1973)

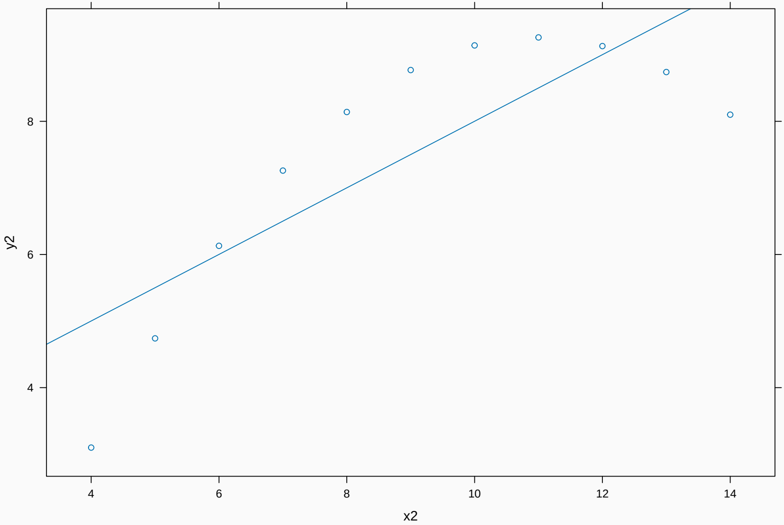
Imagine quatro conjuntos de dados. Em cada um deles, temos duas variáveis (X e Y). Em todos, temos...

- ...a mesma média e o mesmo desvio-padrão de X.
- ...a mesma média e o mesmo desvio-padrão de Y.
- ...a mesma correlação entre X e Y.
- ...os mesmos coeficientes estimados para uma regressão linear de Y em X.

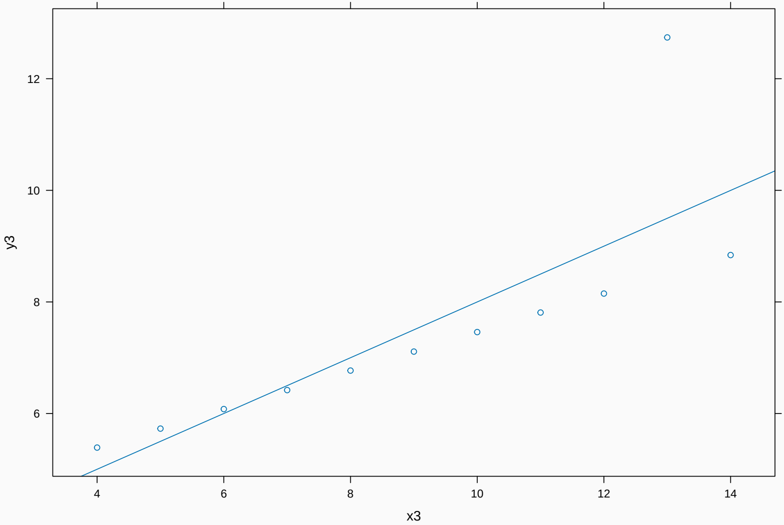
O quarteto de Anscombe - olhem para os dados!



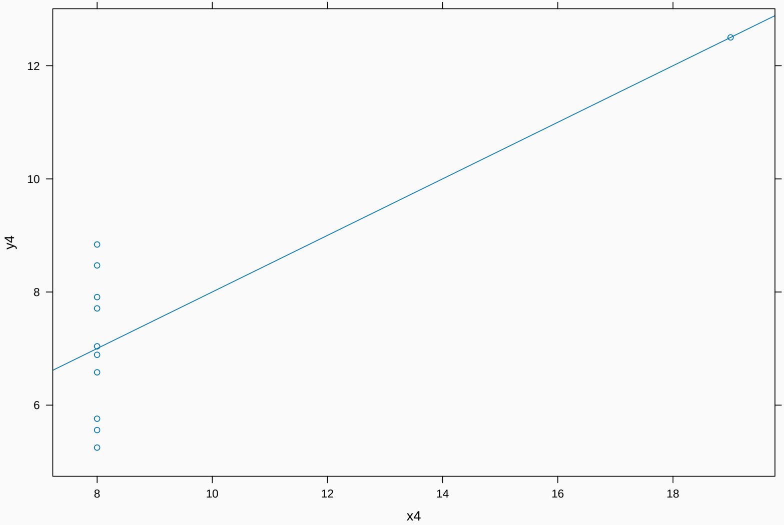
O quarteto de Anscombe - olhem para os dados!



O quarteto de Anscombe - olhem para os dados!



O quarteto de Anscombe - olhem para os dados!



Voltemos à questão dos fluxos comerciais.

Referência

- Como responder a questão que motivou a nossa análise?

Referência

- Como responder a questão que motivou a nossa análise?
  - Utilizemos o capítulo 2 de Wooldridge (2006).

A regressão linear

Assuma que possamos relacionar o fluxo comercial bilateral entre duas economia da seguinte forma:

$$\ln T_i = \beta_0 + \beta_1 \ln Dist_i + \varepsilon_i$$

A regressão linear

Assuma que possamos relacionar o fluxo comercial bilateral entre duas economia da seguinte forma:

$$\ln T_i = \beta_0 + \beta_1 \ln Dist_i + \varepsilon_i$$

O que os parâmetros significam?

A regressão linear

Assuma que possamos relacionar o fluxo comercial bilateral entre duas economia da seguinte forma:

$$\ln T_i = \beta_0 + \beta_1 \ln Dist_i + \varepsilon_i$$

O que os parâmetros significam? Como estimá-los?

Regressão linear com MQO

Assuma que busquemos um estimador que minimize o erro quadrado. Por quê erro quadrado?

Regressão linear com MQO

Assuma que busquemos um estimador que minimize o erro quadrado. Por quê erro quadrado?

- Erro: $\varepsilon_i = \ln T_i - \beta_0 - \beta_1 \ln Dist_i$.

Regressão linear com MQO

Assuma que busquemos um estimador que minimize o erro quadrado. Por quê erro quadrado?

- Erro: $\varepsilon_i = \ln T_i - \beta_0 - \beta_1 \ln Dist_i$.
- $\min_{\widehat{\beta}_0, \widehat{\beta}_1} \sum_{i=1}^{n} \left(\epsilon_i\right)^2 = \sum_{i=1}^{n} \left(\ln T_i - \widehat{\beta_0} - \widehat{\beta_1} \ln Dist_i\right)^2$

Regressão linear com MQO

Assuma que busquemos um estimador que minimize o erro quadrado. Por quê erro quadrado?

- Erro: $\varepsilon_i = \ln T_i - \beta_0 - \beta_1 \ln Dist_i$.
- $\min_{\widehat{\beta}_0, \widehat{\beta}_1} \sum_{i=1}^{n} \left(\epsilon_i\right)^2 = \sum_{i=1}^{n} \left(\ln T_i - \widehat{\beta_0} - \widehat{\beta_1} \ln Dist_i \right)^2$
    - $\widehat{\beta}_1 = \frac{\sum_{i=1}^{n} \left(Dist_i - \overline{Dist}\right)\left(\ln T_i - \overline{\ln T}\right)}{\sum_{i=1}^{n} \left(\ln Dist_i - \overline{\ln Dist}\right)^2}$

Regressão linear com MQO

Assuma que busquemos um estimador que minimize o erro quadrado. Por quê erro quadrado?

- Erro: $\varepsilon_i = \ln T_i - \beta_0 - \beta_1 \ln Dist_i$.
- $\min_{\widehat{\beta}_0,\widehat{\beta}_1} \sum_{i=1}^{n} \left(\epsilon_i\right)^2 = \sum_{i=1}^{n} \left(\ln T_i - \widehat{\beta_0} - \widehat{\beta_1} \ln Dist_i\right)^2$
  - $\widehat{\beta}_1 = \frac{\sum_{i=1}^{n} \left(Dist_i - \overline{Dist}\right)\left(\ln T_i - \overline{\ln T}\right)}{\sum_{i=1}^{n} \left(\ln Dist_i - \overline{\ln Dist}\right)^2}$
  - $\widehat{\beta}_0 = \overline{\ln T} - \widehat{\beta}_1 \overline{Dist}$

Regressão linear com MQO

Genericamente

$$\widehat{\beta_1} = \frac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2} = \frac{cov(X, Y)}{var(X)} = corr(X, Y)\frac{s_X}{s_y}$$

Regressão linear com MQO

Genericamente

$$\widehat{\beta_1} = \frac{\sum_{i=1}^{n} \left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{\sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2} = \frac{cov(X, Y)}{var(X)} = corr(X, Y)\frac{s_X}{s_y}$$

e

$$\widehat{\beta_0} = \bar{Y} - \widehat{\beta_1}\bar{X}$$

Parâmetros vs estimadores

Parâmetros vs estimadores

- Na **população**, temos:

Parâmetros vs estimadores

- Na **população**, temos: $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$

Parâmetros vs estimadores

- Na **população**, temos: $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$
  - Portanto, β_0 e β_1 são **parâmetros populacionais**

Parâmetros vs estimadores

- Na **população**, temos: $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$
  - Portanto, β_0 e β_1 são **parâmetros populacionais** e ε_i representa o **erro**.

Parâmetros vs estimadores

- Na **população**, temos: $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$
  - Portanto, β_0 e β_1 são **parâmetros populacionais** e ε_i representa o **erro**.
- Na mostra: $Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + \hat{\varepsilon}_i$

Parâmetros vs estimadores

- Na **população**, temos: $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$
    - Portanto, β_0 e β_1 são **parâmetros populacionais** e ε_i representa o **erro**.
- Na mostra: $Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + \hat{\varepsilon}_i$
    - onde $\hat{\beta}_0$ e $\hat{\beta}_1$ são os estimadores de β_0 e β_1, respectivamente

Parâmetros vs estimadores

- Na **população**, temos: $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$
    - Portanto, β_0 e β_1 são **parâmetros populacionais** e ε_i representa o **erro**.
- Na mostra: $Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + \hat{\varepsilon}_i$
    - onde $\hat{\beta}_0$ e $\hat{\beta}_1$ são os estimadores de β_0 e β_1, respectivamente e $\hat{\varepsilon}_i$ representa o **resíduo**.

Terminologia [Wooldridge (2006); p. 21]

y	x
Variável Explicada	Variável Explicativa
Variável Prevista	Variável Preditora
Regressando	Regressor

Vamos estimar as regressões!

Inferência

Definição importante

Vocês aceitam errar quantas vezes para cada 100 tentativas?

Inferência

Como verificar se a associação entre as variáveis é estatísticamente significativa?

Inferência

Como verificar se a associação entre as variáveis é estatísticamente significativa? Com um teste de hipótese sobre o parâmetro estimado!

Inferência

Como verificar se a associação entre as variáveis é estatísticamente significativa? Com um teste de hipótese sobre o parâmetro estimado!

- Para $\widehat{\beta}_1$ (que ficaria muito mais legal como β_1, β_1 ou β_1):

Como verificar se a associação entre as variáveis é estatísticamente significativa? Com um teste de hipótese sobre o parâmetro estimado!

- Para $\widehat{\beta}_1$ (que ficaria muito mais legal como β_1, β_1 ou β_1):

$$\mathcal{H}_0 : \beta_1 = 0$$
$$\mathcal{H}_a : \beta_1 \neq 0$$

(Não precisa ser apenas com $\neq$ e nem com zero!)

Vamos simular o comportamento de β_0 e β_1 em diferentes amostras?

Vamos simular para entender o que significa um teste de hipótese

Imagine que tenhamos 500 amostras de tamanho 200 com duas variáveis, X e Y.

Vamos simular para entender o que significa um teste de hipótese

Imagine que tenhamos 500 amostras de tamanho 200 com duas variáveis, X e Y. E que saibamos que $Y_i = 2 + 3X_i + \epsilon_i$.

Vamos simular para entender o que significa um teste de hipótese

Imagine que tenhamos 500 amostras de tamanho 200 com duas variáveis, X e Y. E que saibamos que $Y_i = 2 + 3X_i + \epsilon_i$. Ou seja, que $\beta_0 = 2$ e $\beta_1 = 3$.

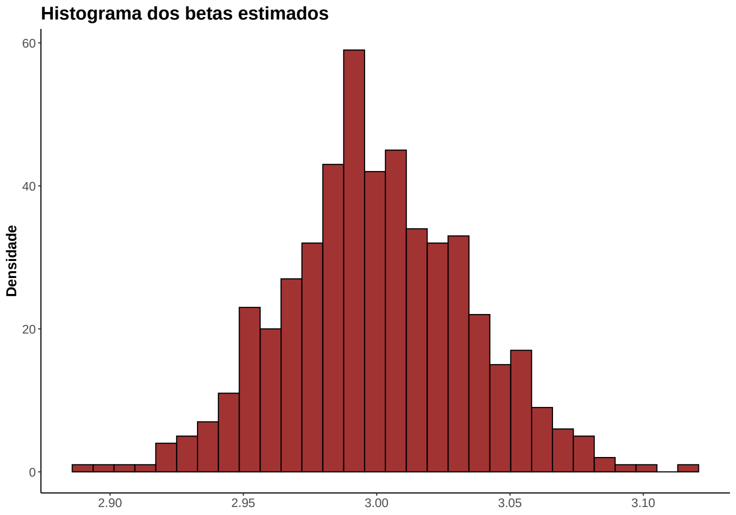
Vamos simular para entender o que significa um teste de hipótese

Imagine que tenhamos 500 amostras de tamanho 200 com duas variáveis, X e Y. E que saibamos que $Y_i = 2 + 3X_i + \epsilon_i$. Ou seja, que $\beta_0 = 2$ e $\beta_1 = 3$. Quais seriam os resultados dos estimadores ($\widehat{\beta}_0$ e $\widehat{\beta}_1$) em cada uma delas?

Vamos simular para entender o que significa um teste de hipótese

Imagine que tenhamos 500 amostras de tamanho 200 com duas variáveis, X e Y. E que saibamos que $Y_i = 2 + 3X_i + \epsilon_i$. Ou seja, que $\beta_0 = 2$ e $\beta_1 = 3$. Quais seriam os resultados dos estimadores ($\widehat{\beta}_0$ e $\widehat{\beta}_1$) em cada uma delas? Podemos identificar algum padrão?

Vamos simular para entender o que significa um teste de hipótese



Distribuição amostral

Ou seja, tanto $\widehat{\beta}_0$ quanto $\widehat{\beta}_1$ são **estatísticas** (i.e. funções dos valores amostrais) e cada estatística possui uma **distribuição**. Em função disso, podemos (i) definir um nível de significância e (ii) fazer um teste de hipótese sobre o parâmetro de interesse.

Teste t

Teste t

- Para $\widehat{\beta}_1$:

Teste t

- Para $\widehat{\beta}_1$:

$$\mathcal{H}_0 : \beta_1 = \mu$$
$$\mathcal{H}_a : \beta_1 \neq \mu$$

Teste t

- Para $\widehat{\beta}_1$:

$$\mathcal{H}_0 : \beta_1 = \mu$$
$$\mathcal{H}_a : \beta_1 \neq \mu$$

A estatística do teste é dada por:

$$t_{\widehat{\beta}_1} = \frac{\widehat{\beta}_1 - \mu}{se(\widehat{\beta}_1)} \sim t_{n-k-1}.$$

Teste t

Como fica o teste para as regressões que estimamos?

Goodness of fit

Goodness of fit

Quanto o modelo explica da variação das trocas comerciais bilaterais?

Goodness of fit

Quanto o modelo explica da variação das trocas comerciais bilaterais?

- Do total da soma (dos quadrados) dos residuos, $\sum_{i=1}^{n} \left(Y_i - \bar{Y} \right)^2 = (n-1)s_Y^2 \ldots$

Goodness of fit

Quanto o modelo explica da variação das trocas comerciais bilaterais?

- Do total da soma (dos quadrados) dos residuos, $\sum_{i=1}^{n} \left(Y_i - \bar{Y} \right)^2 = (n-1)s_Y^2 \ldots$
- …uma parte é explicada pelo modelo, $\sum_{i=1}^{n} \left(\widehat{Y}_i - \bar{Y} \right)^2 = s_{\widehat{Y}}^2 \ldots$

Goodness of fit

Quanto o modelo explica da variação das trocas comerciais bilaterais?

- Do total da soma (dos quadrados) dos residuos, $\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2 = (n-1)s_Y^2 \ldots$
- ... uma parte é explicada pelo modelo, $\sum_{i=1}^{n}\left(\widehat{Y}_i - \bar{Y}\right)^2 = s_{\widehat{Y}}^2 \ldots$
- ... e outra parte é explicada pelo erro, $\sum_{i=1}^{n}\left(\varepsilon_i - 0\right)^2 = s_{\varepsilon}^2 \ldots$

Goodness of fit

Quanto o modelo explica da variação das trocas comerciais bilaterais?

- Do total da soma (dos quadrados) dos residuos, $\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2 = (n-1)s_Y^2 \ldots$
- ... uma parte é explicada pelo modelo, $\sum_{i=1}^{n}\left(\widehat{Y}_i - \bar{Y}\right)^2 = s_{\widehat{Y}}^2 \ldots$
- ... e outra parte é explicada pelo erro, $\sum_{i=1}^{n}\left(\varepsilon_i - 0\right)^2 = s_{\varepsilon}^2 \ldots$
- Assim, podemos definir uma estatística que avalia quão aderente é o modelo aos dados: $R^2 = \frac{s_{\widehat{Y}}^2}{s_Y^2} = 1 - \frac{s_{\epsilon}^2}{s_Y^2}$

Por quê MQO?

BLUE

- Sob algumas hipóteses (Gauss-Markov), o estimador de mínimos quadros é **BLUE** (*best linear unbiased estimator*). Mesmo sem assumirmos a normalidade dos erros!

BLUE

- Sob algumas hipóteses (Gauss-Markov), o estimador de mínimos quadros é **BLUE** (*best linear unbiased estimator*). Mesmo sem assumirmos a normalidade dos erros!
- Sob a hipótese de normalidade dos erros, o estimador de MQO é o mais eficiente entre os estimadores lineares e não-lineares (Cramér–Rao)!

BLUE

- Sob algumas hipóteses (Gauss-Markov), o estimador de mínimos quadros é **BLUE** (*best linear unbiased estimator*). Mesmo sem assumirmos a normalidade dos erros!
- Sob a hipótese de normalidade dos erros, o estimador de MQO é o mais eficiente entre os estimadores lineares e não-lineares (Cramér–Rao)!
- E quais são essas hipóteses?

Hipóteses

Hipóteses

- **Linearidade**: $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$ (linear nos parâmetros; as variáveis podem ser não-lineares).

Hipóteses

- **Linearidade**: $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$ (linear nos parâmetros; as variáveis podem ser não-lineares).
- **Exogeneidade**: $E[\varepsilon_i | X_i] = E[\varepsilon_i] = 0$.

Hipóteses

- **Linearidade**: $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$ (linear nos parâmetros; as variáveis podem ser não-lineares).
- **Exogeneidade**: $E[\varepsilon_i | X_i] = E[\varepsilon_i] = 0$.
- **Multicolinearidade não-perfeita**: se tivermos mais de uma variável X (e.g. $X_1, X_2, \cdots, X_k$), elas não podem ser perfeitamente correlacionadas.

Hipóteses

- **Linearidade**: $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$ (linear nos parâmetros; as variáveis podem ser não-lineares).
- **Exogeneidade**: $E[\varepsilon_i | X_i] = E[\varepsilon_i] = 0$.
- **Multicolinearidade não-perfeita**: se tivermos mais de uma variável X (e.g. $X_1, X_2, \cdots, X_k$), elas não podem ser perfeitamente correlacionadas.
- **Homocedasticidade**: $Var[\varepsilon_i | X_i] = \sigma^2$ e $Cov[\varepsilon_i, \varepsilon_j | X_i] = 0$.

Exogeneidade

Exogeneidade

Esse é um ponto crucial para nós.

Exogeneidade

Esse é um ponto crucial para nós.

- O termo erro (ε_i) inclui, por definição, tudo o que não está no modelo. Ele **não** pode influenciar as variáveis explicativas (X). Se isso acontecer, é porque temos:
  - Variáveis omitidas.

Exogeneidade

Esse é um ponto crucial para nós.

- O termo erro (ε_i) inclui, por definição, tudo o que não está no modelo. Ele **não** pode influenciar as variáveis explicativas (X). Se isso acontecer, é porque temos:
  - Variáveis omitidas.
  - Erro de mensuração das variáveis explicativas.

Exogeneidade

Esse é um ponto crucial para nós.

- O termo erro (ε_i) inclui, por definição, tudo o que não está no modelo. Ele **não** pode influenciar as variáveis explicativas (X). Se isso acontecer, é porque temos:
  - Variáveis omitidas.
  - Erro de mensuração das variáveis explicativas.
  - Simultaneidade

BLUE

BLUE

Sob exogeneidade e Multicolinearidade não-perfeita, o estimador de MQO é **consistente** e **não-viesado**.

BLUE

Sob exogeneidade e Multicolinearidade não-perfeita, o estimador de MQO é **consistente** e **não-viesado**.

- Consistência: $plim_{n\to\infty}|\widehat{\beta}_1 - \beta_1| = 0$.

BLUE

Sob exogeneidade e Multicolinearidade não-perfeita, o estimador de MQO é **consistente** e **não-viesado**.

- Consistência: $plim_{n\to\infty}|\widehat{\beta}_1 - \beta_1| = 0$.
- Não-viesado: $E[\widehat{\beta}_1] = \beta_1$

Referências i

Anscombe, Francis J. 1973. "Graphs in Statistical Analysis." *The American Statistician* 27 (1): 17–21.

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