
# Estatística II

## Licenciatura em Gestão

## 2.º Ano/1.º Semestre

## 2023/2024

# Aulas Teóricas N.ºs 20 e 21 (Semana 11)

**Docente**: Elisabete Fernandes
**E-mail**: efernandes@iseg.ulisboa.pt

https://doity.com.br/estatistica-aplicada-a-nutricao

https://basiccode.com.br/produto/informatica-basica/

# Conteúdos Programáticos

| Aulas Teóricas (Semanas 1 a 5) | Aulas Teóricas (Semanas 5 a 7) | Aulas Teóricas (Semanas 7 a 9) | Aulas Teóricas (Semanas 10 a 13) |
|---|---|---|---|
| • **Capítulo 1**: Estimação | • **Capítulo 2**: Testes de Hipóteses | • **Capítulo 3**: Modelo de Regressão Linear | • **Capítulo 4**: Complementos ao Modelo de Regressão Linear |

**Material didático:** Exercícios do Livro Murteira et al (2015), Formulário e Tabelas Estatísticas

**Bibliografia**: B. Murteira, C. Silva Ribeiro, J. Andrade e Silva, C. Pimenta e F. Pimenta; *Introdução à Estatística*, 2ª ed., Escolar Editora, 2015.

**8ª semana (07/11 e 09/11)**

T14 - Modelo de Regressão Linea (MRL)r

Interpretação dos parâmetros da regressão; exemplos; Resíduos MQ e regressão ajustada; Propriedades dos estimadores MQ dos coeficientes da regressão; Estimador não enviesado da variância da variável residual; Exemplo.

T15 - Modelo de regressão Linear

Coeficiente de determinação e sua interpretação. Hipótese adicional (H6) e inferência estatística sobre o modelo; Inferência sobre um parâmetro beta. Exemplos

**9ª semana (14/11 e 16/11)**

T16 - Modelo de Regressão Linear

Mais exemplos de inferência sobre um parâmetro beta; Inferência sobre uma combinação linear de betas; exemplos.

T17 - Modelo de Regressão Linear

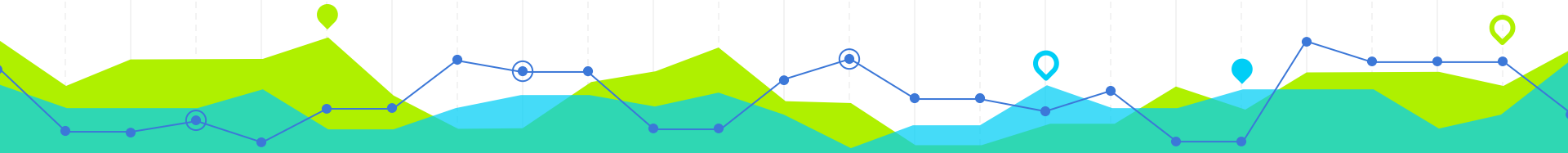
Teste de nulidade conjunta de vários coeficientes; exemplo; Teste F à significância global da regressão; Teste de um conjunto de restrições lineares; exemplo.

**10ª semana (21/11 e 23/11)**

T18 - Complementos ao MRL

Variáveis artificiais: Introdução à modelação de fatores qualitativos, conceito de variável artificial, estimação e interpretação do modelo com variáveis artificiais; exemplos.

T19 - Complementos ao MRL



# Modelo de Regressão Linear Múltipla

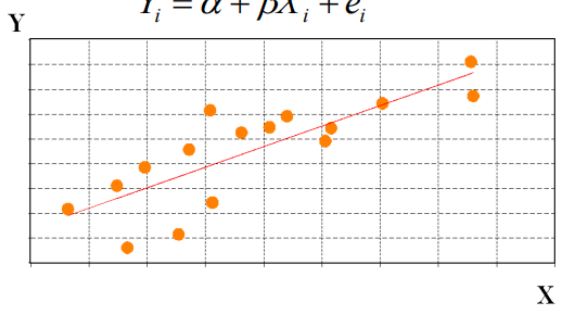
1

Interpretação dos Coeficientes da Reta de Regressão

# MRLM: Interpretação dos Coeficientes

**Regressão Linear Simples:**

$$Y_i = \alpha + \beta X_i + e_i$$



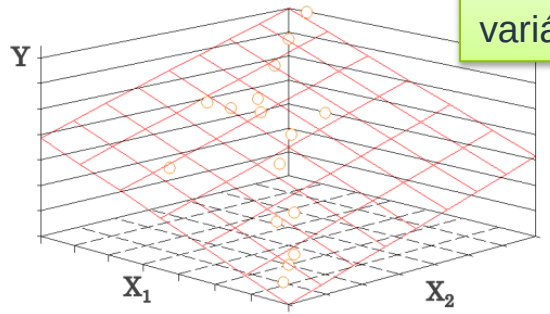

**Temos que:**

$E[Y / X = 0] = \alpha$ Valor esperado de *Y* quando *X* é nulo.

$\frac{dY}{dX} = \beta$ Variação marginal esperada em *Y* para cada variação unitária em *X*.

**Regressão Linear Múltipla:**

$$Y_i = \alpha + \beta_1 X_{1_i} + \beta_2 X_{2_i} + e_i$$




**Temos que:**

$E[Y / X_1 = 0, X_2 = 0] = \alpha$ Valor esperado de *Y* quando ambos $X_1$ e $X_2$ são nulos.

$\frac{\partial Y}{\partial X_1} = \beta_1$ Variação marginal esperada em *Y* para cada variação unitária em $X_1$, *mantendo* $X_2$ *constante*.

$\frac{\partial Y}{\partial X_2} = \beta_2$ Variação marginal esperada em *Y* para cada variação unitária em $X_2$, *mantendo* $X_1$ *constante*.

**Caso particular**: MRLM com apenas duas variáveis regressoras

Maia, Alexandre Gori (2017). Econometria: conceitos e aplicações. Cap. 6.

# MRLM: Interpretação dos Coeficientes...

**Caso geral**: MRLM com k variáveis regressoras

**Regressão Múltipla**

Em um modelo de regressão múltipla, a variável dependente ($Y$) será determinada por mais de uma variável independente ($X$). Genericamente, um modelo de regressão linear múltipla com $k$ variáveis independentes e $p$ parâmetros ($p=k+1$) pode ser representado por:

$$Y_i = \alpha + \beta_1 X_{1_i} + \beta_2 X_{2_i} + \ldots + \beta_k X_{k_i} + e_i$$

Onde:

$\alpha$ é o valor esperado de $Y$ quando todos as variáveis independentes forem nulas;

$\beta_1$ é a variação esperada em $Y$ dado um incremento unitário em $X_1$, mantendo-se constantes todas as demais variáveis independentes;

...

$\beta_k$ é a variação esperada em $Y$ dado um incremento unitário em $X_k$, mantendo-se constantes todas as demais variáveis independentes;

$e_i$ é o erro não explicado pelo modelo;

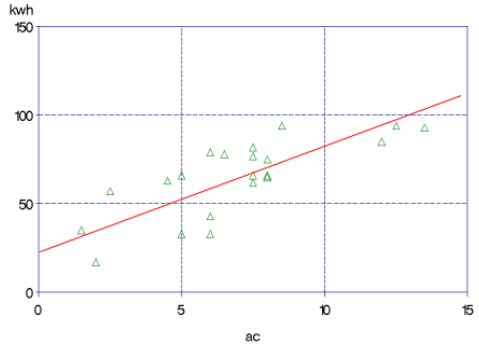
Maia, Alexandre Gori (2017). Econometria: conceitos e aplicações. Cap. 6.

# MRLM: Exemplo 1 - Interpretação dos Coeficientes

**Seja a relação para consumo de energia (*Kwh*), horas de ar condicionado ligado (*AC*) e horas de secadora ligada (*SEC*):**

Regressão Linear Simples

$$Kwh_i = \alpha + \beta AC_i + e_i$$

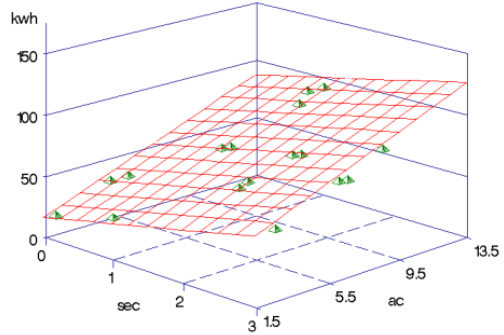



O coeficiente $\alpha$ indicará o consumo esperado de energia quando o ar condicionado permanecer desligado.

O coeficiente $\beta$ indicará o consumo de energia adicional esperado para cada hora adicional com ar condicionado ligado.

Regressão Linear Múltipla

$$Kwh_i = \alpha + \beta_1 AC_i + \beta_2 SEC_i + e_i$$




O coeficiente $\alpha$ indicará o consumo esperado de energia quando ambos ar condicionado e secadora permanecerem desligados.

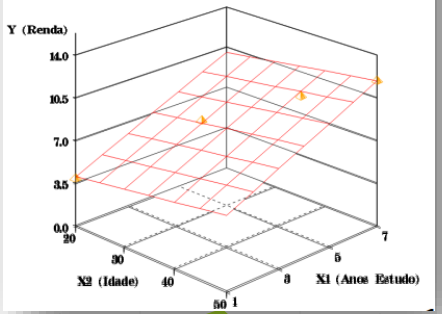
O coeficiente $\beta_1$ indicará o aumento no consumo de energia esperado para cada hora adicional com ar condicionado ligado, mantendo-se constante o tempo de uso da secadora. Analogamente, O coeficiente $\beta_2$ indicará efeito isolado de uma hora adicional com a secadora ligada sobre o consumo esperado de energia.

Maia, Alexandre Gori (2017). Econometria: conceitos e aplicações. Cap. 6.

# MRLM: Exemplo 2 - Estimação dos Coeficientes e Interpretação

**Seja a relação entre renda familiar em SM ($Y$), anos de estudo ($X_1$) e idade ($X_2$) do responsável pela família:**

| Y (Renda) | $X_1$ (Anos Estudo) | $X_2$ (Idade) |
|---|---|---|
| 4 | 1 | 20 |
| 8 | 4 | 30 |
| 10 | 6 | 40 |
| 12 | 7 | 50 |




$$Y_i = \alpha + \beta_1 X_{1_i} + \beta_2 X_{2_i} + e_i \Rightarrow \mathbf{y} = \mathbf{X\beta} + \mathbf{e}$$

A função de regressão amostral será dada por:

$$\mathbf{y} = \mathbf{X}\hat{\boldsymbol{\beta}} + \hat{\mathbf{e}} \Rightarrow \underbrace{\begin{pmatrix} 4 \\ 8 \\ 10 \\ 12 \end{pmatrix}}_{\mathbf{y}_{4\times1}} = \underbrace{\begin{pmatrix} 1 & 1 & 20 \\ 1 & 4 & 30 \\ 1 & 6 & 40 \\ 1 & 7 & 50 \end{pmatrix}}_{\mathbf{X}_{4\times3}} \underbrace{\begin{pmatrix} \hat{\alpha} \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix}}_{\hat{\boldsymbol{\beta}}_{3\times1}} + \underbrace{\begin{pmatrix} \hat{e}_1 \\ \hat{e}_2 \\ \hat{e}_3 \\ \hat{e}_4 \end{pmatrix}}_{\hat{\mathbf{e}}_{4\times1}}$$

E as estimativas de MQO:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}(\mathbf{X}^T\mathbf{y}) = \begin{pmatrix} 4 & 18 & 140 \\ 18 & 102 & 730 \\ 140 & 730 & 5400 \end{pmatrix}^{-1} \begin{pmatrix} 34 \\ 180 \\ 1320 \end{pmatrix} = \begin{pmatrix} 1{,}9 \\ 1 \\ 0{,}06 \end{pmatrix}$$

Maia, Alexandre Gori (2017). Econometria: conceitos e aplicações. Cap. 6.

Espera-se, para cada ano adicional de estudo do responsável pela família, um aumento de 1 SM e, para cada ano de idade adicional, um aumento de 0,06 SM na renda familiar.

O departamento de RH da empresa TEMCO objetiva estudar o comportamento dos salários dos funcionários dos mais diversos setores da empresa.

Para tanto, o gerente de RH, baseando-se numa amostra aleatória de 46 empregados, coletou informações sobre as seguintes variáveis:

**id –** número cadastral do funcionário;

**salario –** anual, em dólares;

**anosemp –** tempo (em anos) na empresa;

**expprev –** experiência anterior (em anos);

**educ –** anos de estudo após o segundo grau;

**sexo –** (feminino = 0, masculino = 1);

**dept –** departamento no qual atua (Compras = 1, Engenharia = 2, Propaganda = 3, Vendas = 4);

**super –** número de empregados sob responsabilidade do empregado.

EViews - [Group: UNTITLED Workfile: TEMCO::Temco\]

File Edit Object View Proc Quick Options Window Help

View | Proc | Object | Print | Name | Freeze | Default | Sort | Transpose | Edit+/- | Smpl+/- | Title | Sample

| obs | ID | SALARIO | ANOSEMP | EXPPREV | EDUC | SEXO | DEPT | SUPER |
|---|---|---|---|---|---|---|---|---|
| 1 | 972.0000 | 47536.00 | 15.00000 | 5.000000 | 6.000000 | 0.000000 | 3.000000 | 4.000000 |
| 2 | 539.0000 | 23654.00 | 0.000000 | 0.000000 | 0.000000 | 1.000000 | 3.000000 | 2.000000 |
| 3 | 649.0000 | 37548.00 | 19.00000 | 9.000000 | 4.000000 | 0.000000 | 3.000000 | 6.000000 |
| 4 | 824.0000 | 36578.00 | 4.000000 | 4.000000 | 8.000000 | 0.000000 | 3.000000 | 8.000000 |
| 5 | 649.0000 | 54679.00 | 20.00000 | 3.000000 | 6.000000 | 1.000000 | 3.000000 | 4.000000 |
| 6 | 624.0000 | 53234.00 | 25.00000 | 0.000000 | 6.000000 | 0.000000 | 3.000000 | 3.000000 |
| 7 | 891.0000 | 31425.00 | 7.000000 | 6.000000 | 5.000000 | 1.000000 | 3.000000 | 6.000000 |
| 8 | 974.0000 | 39743.00 | 9.000000 | 6.000000 | 5.000000 | 1.000000 | 2.000000 | 1.000000 |
| 9 | 648.0000 | 26452.00 | 1.000000 | 3.000000 | 2.000000 | 1.000000 | 2.000000 | 0.000000 |
| 10 | 321.0000 | 34632.00 | 5.000000 | 4.000000 | 4.000000 | 0.000000 | 2.000000 | 0.000000 |
| 11 | 264.0000 | 35631.00 | 6.000000 | 4.000000 | 4.000000 | 0.000000 | 2.000000 | 2.000000 |
| 12 | 291.0000 | 46211.00 | 14.00000 | 5.000000 | 6.000000 | 1.000000 | 2.000000 | 5.000000 |
| 13 | 267.0000 | 34231.00 | 6.000000 | 2.000000 | 6.000000 | 0.000000 | 2.000000 | 3.000000 |
| 14 | 548.0000 | 26548.00 | 5.000000 | 1.000000 | 0.000000 | 0.000000 | 2.000000 | 2.000000 |
| 15 | 555.0000 | 36512.00 | 6.000000 | 6.000000 | 4.000000 | 1.000000 | 2.000000 | 2.000000 |
| 16 | 366.0000 | 34869.00 | 7.000000 | 5.000000 | 4.000000 | 1.000000 | 2.000000 | 1.000000 |
| 17 | 246.0000 | 41255.00 | 9.000000 | 4.000000 | 6.000000 | 0.000000 | 2.000000 | 4.000000 |
| 18 | 215.0000 | 39331.00 | 9.000000 | 3.000000 | 6.000000 | 1.000000 | 2.000000 | 1.000000 |
| 19 | 814.0000 | 35487.00 | 8.000000 | 2.000000 | 2.000000 | 1.000000 | 2.000000 | 2.000000 |
| 20 | 212.0000 | 36487.00 | 6.000000 | 5.000000 | 2.000000 | 0.000000 | 2.000000 | 3.000000 |
| 21 | 526.0000 | 68425.00 | 25.00000 | 2.000000 | 12.00000 | 0.000000 | 2.000000 | 1.000000 |
| 22 | 778.0000 | 69246.00 | 22.00000 | 3.000000 | 10.00000 | 0.000000 | 2.000000 | 45.00000 |

**Quadro 1 - Parte de uma planilha que contem informações sobre os empregados da empresa TEMCO.**

Como parte do estudo, a gerente de RH propôs a estimação dos parâmetros do seguinte modelo de regressão múltipla:

$$salario = \beta_0 + \beta_1 educ + \beta_2 anosemp + \varepsilon$$

a) Em termos do problema, $\beta_0$ apresenta algum significado prático?

b) Qual o sinal esperado para $\beta_1$? E para $\beta_2$?

c) Encontre as estimativas dos parâmetros, via mínimos quadrados ordinários, escreva a equação estimada e interprete os resultados obtidos, em termos do problema de interesse.



# Exercícios (a) e (b): Interpretação dos Parâmetros

**Interpretação dos parâmetros do modelo proposto, em termos do problema:**

- $\beta_0$ – salário médio dos funcionários da empresa TEMCO, que acabaram de entrar na empresa (ou que ainda não completaram um ano) e que não apresentam nenhum ano de escolaridade após o segundo grau;
- $\beta_1$ – efeito no salário médio dos funcionários da empresa TEMCO, dada a variação de um ano no tempo de escolaridade após o segundo grau, mantendo constante a variável *anosemp*; e
- $\beta_2$ – efeito no salário médio dos funcionários da empresa TEMCO, dada a variação de um ano no tempo de empresa, mantendo constante a variável *educ*.

# Exercícios (a) e (b): Interpretação dos Parâmetros

Dependent Variable: SALARIO
Method: Least Squares
Date: 08/26/12 Time: 15:45
Sample: 1 46
Included observations: 46
SALARIO=C(1)+C(2)*EDUC+C(3)*ANOSEMP

| | Coefficient | Std. Error | t-Statistic | Prob. |
|---|---|---|---|---|
| C(1) | 23177.47 | 1769.732 | 13.09660 | 0.0000 |
| C(2) | 1916.489 | 379.2670 | 5.053139 | 0.0000 |
| C(3) | 672.3250 | 141.6725 | 4.745629 | 0.0000 |

| | | | |
|---|---|---|---|
| R-squared | 0.739927 | Mean dependent var | 39827.39 |
| Adjusted R-squared | 0.727830 | S.D. dependent var | 10999.24 |
| S.E. of regression | 5738.291 | Akaike info criterion | 20.21070 |
| Sum squared resid | 1.42E+09 | Schwarz criterion | 20.32996 |
| Log likelihood | -461.8462 | Hannan-Quinn criter. | 20.25538 |
| F-statistic | 61.16907 | Durbin-Watson stat | 1.229794 |
| Prob(F-statistic) | 0.000000 | | |

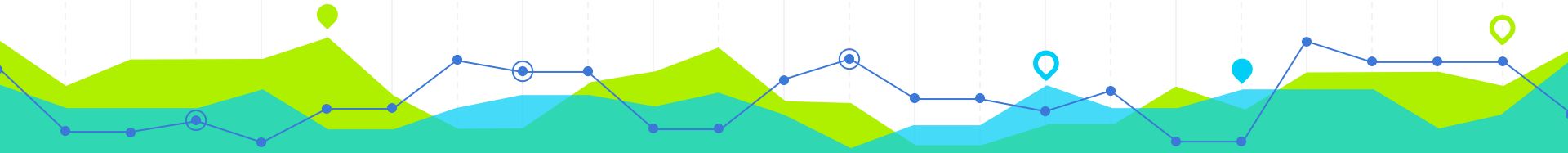
# Exercício (c): Modelo Estimado

$$\hat{salário} = 23177{,}47 + 1916{,}49\, educ + 672{,}32\, anosemp$$

**Pergunta**: qual o salário médio estimado para pessoas com 3 anos de escolaridade após o 2º grau e com 5 anos na empresa?

$$\hat{salario} = 23.177{,}47 + 1.916{,}49 * 3 + 672{,}33 * 5$$

$$\hat{salario} = 32288{,}54$$



# Qualidade do Modelo de Regressão Linear Múltipla

## Coeficiente de Determinação

2

# Coeficiente de Determinação ($R^2$): Medida do Grau de Ajuste

- O **coeficiente de determinação** $R^2$ é uma medida que diz quão bem a reta de regressão da amostra se ajusta ao dados.

# $R^2$: Medida do Grau de Ajuste

- Nós podemos dividir a variação em Y em dois componentes, uma parte explicada pelo modelo de regressão e uma parte não explicada.

- O **coeficiente de determinação** $R^2$ mede a proporção ou a percentagem da variação total em Y explicada pelo modelo de regressão.

# Como Calcular $R^2$?

Definindo :

$\sum (Y_i - \bar{Y})^2$ como Soma dos Quadrados Total (**SST**)

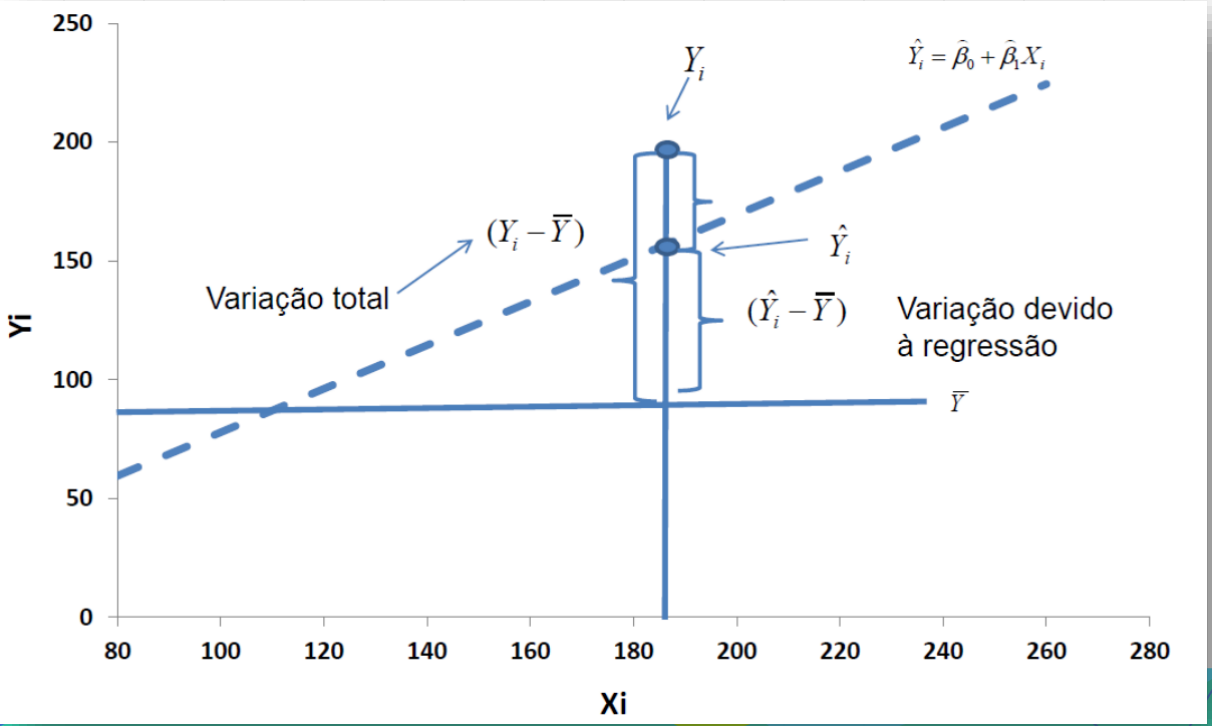
$\sum (\hat{Y}_i - \bar{Y})^2$ como Soma dos Quadrados Explicada (**SSR**)

$\sum \hat{u}_i^2$ como Soma dos Quadrados dos Resíduos (**SSE**)

É possível mostrar que

$$SST = SSR + SSE$$

# Ilustração da Regressão Linear Simples



# Coeficiente de Determinação

Resultado: **SST = SSR + SSE**

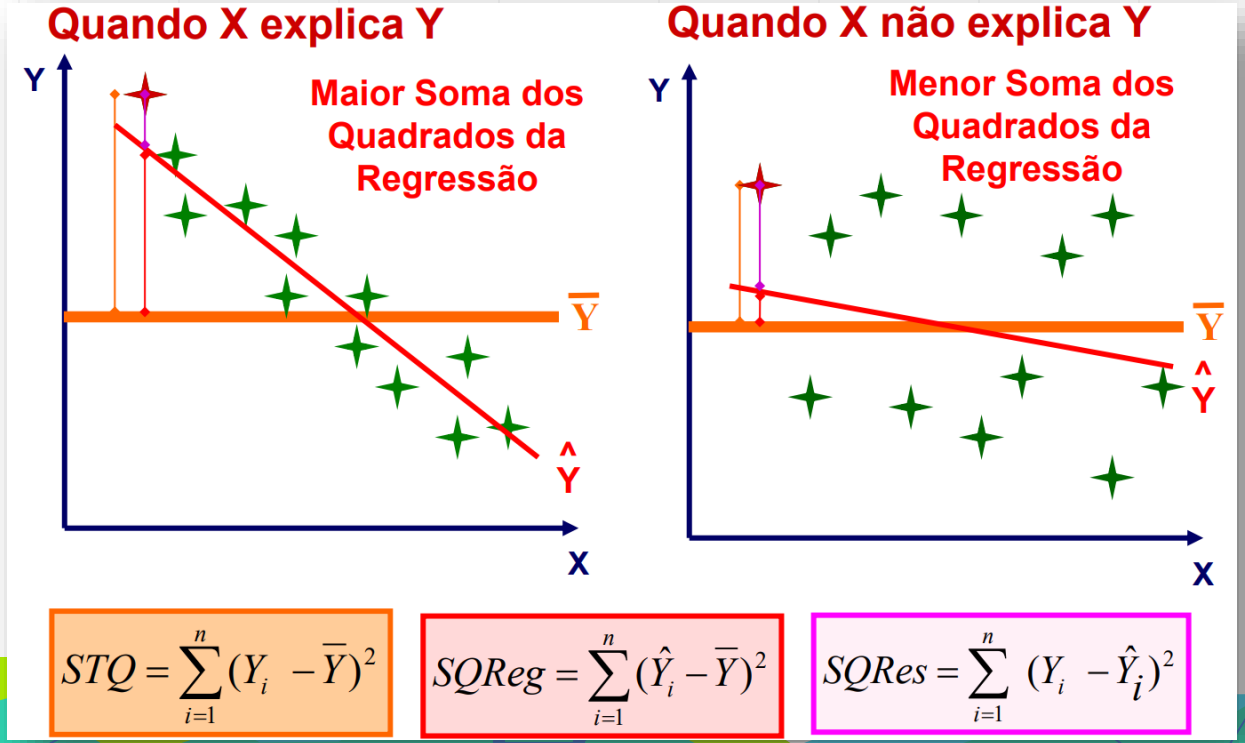
Parcela da variabilidade de y que é explicada pelas variáveis do modelo

Parcela da variabilidade de y que **não** é explicada pelas variáveis do modelo

$$R^2 = \frac{SSR}{SST}$$

**Proporção da variabilidade de y que é explicada pelo conjunto de variáveis explicativas.**

# Soma dos Quadrados: Conceitos




$$STQ = \sum_{i=1}^{n} (Y_i - \overline{Y})^2$$

$$SQReg = \sum_{i=1}^{n} (\hat{Y}_i - \overline{Y})^2$$

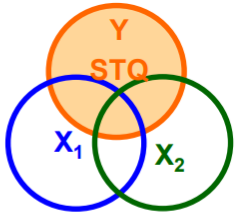
$$SQRes = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

# Soma dos Quadrados: Definição

**Soma Total dos Quadrados (STQ):**

$$STQ = \sum_{i=1}^{n} (Y_i - \overline{Y})^2 = \sum_{i=1}^{n} y_i^{\,2} = \mathbf{y}^T\mathbf{y} - n\overline{Y}^2$$

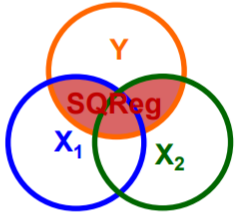
Variabilidade total da variável dependente. Representa as distâncias quadráticas dos valores de *Y* em relação à média aritmética.




**Soma dos Quadrados da Regressão (SQReg):**

$$SQReg = \sum_{i=1}^{n} (\hat{Y}_i - \overline{Y})^2 = \hat{\boldsymbol{\beta}}^T\mathbf{X}^T\mathbf{y} - n\overline{Y}^2$$

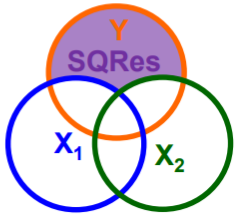
Variabilidade da variável dependente explicada pelo conjunto de variáveis independentes. Representa as distâncias quadráticas dos valores ajustados pelo modelo em relação à média aritmética.




**Soma dos Quadrados dos Resíduos (SQRes):**

$$SQRes = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \hat{\mathbf{e}}^T\hat{\mathbf{e}} = \mathbf{y}^T\mathbf{y} - \hat{\boldsymbol{\beta}}^T\mathbf{X}^T\mathbf{y}$$

Variabilidade da variável dependente não explicada pelo conjunto de variáveis independentes. Representa as distâncias quadráticas entre os valores observados de Y e seus valores ajustados pelo modelo.

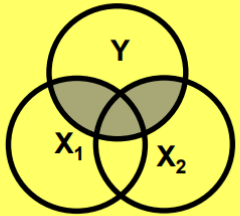


# Coeficiente de Determinação ($R^2$)

**Coeficiente de Determinação ($R^2$):**

Definição: Estima a proporção da variabilidade da variável dependente ($Y$) que é explicada pelo conjunto das $k$ variáveis independentes do modelo de regressão ($X$).

$$R^2 = \frac{SQReg}{STQ} = 1 - \frac{SQRes}{STQ}$$




**Escala para $R^2$:**

**0**
**Independência linear**

**1**
**Relação linear exata**

# Coeficiente de Determinação: Exemplo

**Seja a relação entre renda familiar em salários mínimos (*Y*), anos de estudo ($X_1$) e idade ($X_2$) do responsável pela família:** $Y_i = 1{,}9 + 1X_{1_i} + 0{,}06X_{2_i} + \hat{e}_i$

| Y (Renda) | $X_1$ (Anos Estudo) | $X_2$ (Idade) |
|---|---|---|
| 4 | 1 | 20 |
| 8 | 4 | 30 |
| 10 | 6 | 40 |
| 12 | 7 | 50 |

| Fonte | gl | Soma dos Quadrados | Quadrados Médios |
|---|---|---|---|
| Regressão | 2 | 34,8 | 17,4 |
| Resíduos | 1 | 0,2 | 0,2 |
| Total | 3 | 35,0 | |

$$STQ = \mathbf{y}^T\mathbf{y} - n\bar{Y}^2 = \begin{pmatrix}4 & 8 & 10 & 12\end{pmatrix}\begin{pmatrix}4\\8\\10\\12\end{pmatrix} - 4(8{,}5)^2 = 324 - 289 = 35$$

$$SQReg = \hat{\boldsymbol{\beta}}^T\mathbf{X}^T\mathbf{y} - n\bar{Y}^2 = \begin{pmatrix}1{,}9 & 1 & 0{,}06\end{pmatrix}\begin{pmatrix}34\\180\\1320\end{pmatrix} - 4(8{,}5)^2 = 323{,}8 - 289 = 34{,}8$$

$$SQRes = STQ - SQReg = 35 - 34{,}8 = 0{,}2$$

$$R^2 = \frac{SQReg}{STQ} = \frac{34{,}8}{35} = 0{,}994$$

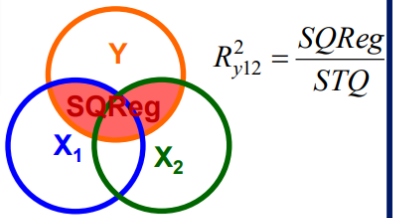
As variáveis anos de estudo e idade explicam, conjuntamente, quase a totatilidade (99,4%) da variabilidade observada para a renda familiar na amostra.

# Coeficiente de Determinação Ajustado

Regressão Linear Múltipla com duas variáveis independentes

**Seja o ajuste:**

$$Y_i = \alpha + \hat{\beta}_1 X_{1_i} + \hat{\beta}_2 X_{2_i} + \hat{e}_i$$

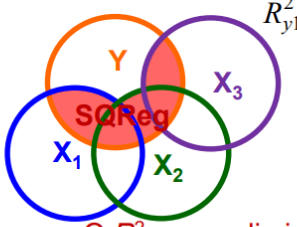



$$R^2_{y12} = \frac{SQReg}{STQ}$$

**Incorporando uma variável independente adicional ($X_3$):**

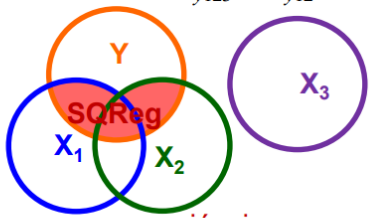
$$Y_i = \alpha + \hat{\beta}_1 X_{1_i} + \hat{\beta}_2 X_{2_i} + \hat{\beta}_3 X_{3_i} + \hat{e}_i$$

**Poderemos ter:**




$$R^2_{y123} > R^2_{y12}$$

ou




$$R^2_{y123} = R^2_{y12}$$

O $R^2$ nunca diminui quando incorporamos variáveis independentes adicionais no modelo.

**Coeficiente de Determinação Ajustado ($\bar{R}^2$):**

O $R^2$ ajustado ($\bar{R}^2$) pondera o coeficiente de determinação ($R^2$) pelo número de variáveis explicativas e pelo número de observações da amostra. É particularmente útil quando desejamos comparar modelos de regressão múltipla que prevêem a mesma variável dependente, pois penaliza aquele modelo com maior número de variáveis independentes.

Será dado por:

$$\bar{R}^2 = 1 - \frac{SQRes/[n-(k+1)]}{STQ/(n-1)} = 1 - (1-R^2)\frac{n-1}{n-(k+1)}$$

Regressão Linear Múltipla com três variáveis independentes

# Coeficiente de Determinação Ajustado: Exemplo

**Seja a relação entre renda familiar em salários mínimos (*Y*), anos de estudo ($X_1$) e idade ($X_2$) do responsável pela família:** $Y_i = 1{,}9 + 1X_{1_i} + 0{,}06X_{2_i} + \hat{e}_i$

| Fonte | gl | Soma dos Quadrados | Quadrados Médios | F |
|---|---|---|---|---|
| Regressão | 2 | 34,8 | 17,4 | 87,0 |
| Resíduos | 1 | 0,2 | 0,2 | |
| Total | 3 | 35,0 | | |

$$R^2 = 0{,}994$$

$$\overline{R}^2 = 1 - (1 - 0{,}994)\frac{4-1}{4-(2+1)} = 0{,}982$$

Não há mudanças expressivas no coeficiente de determinação ajustado pelo número de observações e variáveis do modelo é expressivamente inferior ao $R^2$. Reflexo, sobretudo, do elevadíssimo valor encontrado para o $R^2$.

Slide 1 (unicamp.br)

# Variação e Coeficiente de Determinação

Formulário

**No modelo com termo independente:**

$y_t = \beta_1 + \beta_2 x_{t2} + ... + \beta_k x_{tk} + u_t$, $t = 1,2,...,$

**SST** $VT = \sum_{t=1}^{n} (y_t - \bar{y})^2$;

**SSR** $VE = \sum_{t=1}^{n} (\hat{y}_t - \bar{y})^2$;

**SSE** $VR = \sum_{t=1}^{n} \hat{u}_t^2$

$VT = VE + VR$;

$R^2 = \frac{VE}{VT} = 1 - \frac{VR}{VT}$;

$\overline{R}^2 = 1 - \frac{VR/(n-k)}{VT/(n-1)} = 1 - (1 - R^2)\frac{n-1}{n-k}$.

Coeficiente de Determinação

Coeficiente de Determinação Ajustado

**Nota**:
k = nº de parâmetros a estimar
k-1 = nº de variáveis

**A senhorita Jolie, gerente do departamento de RH da empresa TEMCO, objetiva estudar o comportamento médio dos salários dos funcionários dos mais diversos setores da empresa. Para tanto, baseando-se numa amostra aleatória de 46 funcionários da empresa, ela propôs os seguintes modelos de regressão:**

$$\text{salario} = \beta_0 + \beta_1 \text{educ} + \varepsilon \quad (1)$$

$$\text{salario} = \alpha_0 + \alpha_1 \text{anosemp} + \nu \quad (2)$$

$$\text{salario} = \delta_0 + \delta_1 \text{educ} + \delta_2 \text{anosemp} + \xi \quad (3)$$

**Como a gerente pode avaliar a qualidade de ajuste dos modelos?**

Index of /wp-content/uploads/2014/02 (hedibert.org)

# Resolução do Exercício

$$salario = \beta_0 + \beta_1\, educ + \varepsilon$$

Dependent Variable: SALARIO
Method: Least Squares
Date: 08/26/12 Time: 14:31
Sample: 1 46
Included observations: 46
SALARIO=C(1)+C(2)*EDUC

| | Coefficient | Std. Error | t-Statistic | Prob. |
|---|---|---|---|---|
| C(1) | 24581.70 | 2129.189 | 11.54510 | 0.0000 |
| C(2) | 3009.878 | 367.6294 | 8.187262 | 0.0000 |

| | | | |
|---|---|---|---|
| R-squared | 0.603715 | Mean dependent var | 39827.39 |
| Adjusted R-squared | 0.594709 | S.D. dependent var | 10999.24 |
| S.E. of regression | 7002.393 | Akaike info criterion | 20.58840 |
| Sum squared resid | 2.16E+09 | Schwarz criterion | 20.66790 |
| Log likelihood | -471.5331 | Hannan-Quinn criter. | 20.61818 |
| F-statistic | 67.03125 | Durbin-Watson stat | 1.334781 |
| Prob(F-statistic) | 0.000000 | | |

$$salario = \alpha_0 + \alpha_1\, anosemp + \nu$$

Dependent Variable: SALARIO
Method: Least Squares
Date: 08/26/12 Time: 14:23
Sample: 1 46
Included observations: 46
SALARIO=C(1)+C(2)*ANOSEMP

| | Coefficient | Std. Error | t-Statistic | Prob. |
|---|---|---|---|---|
| C(1) | 28394.16 | 1793.951 | 15.82772 | 0.0000 |
| C(2) | 1107.218 | 140.4476 | 7.883500 | 0.0000 |

| | | | |
|---|---|---|---|
| R-squared | 0.585491 | Mean dependent var | 39827.39 |
| Adjusted R-squared | 0.576070 | S.D. dependent var | 10999.24 |
| S.E. of regression | 7161.598 | Akaike info criterion | 20.63336 |
| Sum squared resid | 2.26E+09 | Schwarz criterion | 20.71286 |
| Log likelihood | -472.5673 | Hannan-Quinn criter. | 20.66314 |
| F-statistic | 62.14957 | Durbin-Watson stat | 1.081824 |
| Prob(F-statistic) | 0.000000 | | |

$$salario = \delta_0 + \delta_1\, educ + \delta_2\, anosemp + \xi$$

Dependent Variable: SALARIO
Method: Least Squares
Date: 08/26/12 Time: 15:45
Sample: 1 46
Included observations: 46
SALARIO=C(1)+C(2)*EDUC+C(3)*ANOSEMP

| | Coefficient | Std. Error | t-Statistic | Prob. |
|---|---|---|---|---|
| C(1) | 23177.47 | 1769.732 | 13.09660 | 0.0000 |
| C(2) | 1916.489 | 379.2670 | 5.053139 | 0.0000 |
| C(3) | 672.3250 | 141.6725 | 4.745629 | 0.0000 |

| | | | |
|---|---|---|---|
| R-squared | 0.739927 | Mean dependent var | 39827.39 |
| Adjusted R-squared | 0.727830 | S.D. dependent var | 10999.24 |
| S.E. of regression | 5738.291 | Akaike info criterion | 20.21070 |
| Sum squared resid | 1.42E+09 | Schwarz criterion | 20.32996 |
| Log likelihood | -461.8462 | Hannan-Quinn criter. | 20.25538 |
| F-statistic | 61.16907 | Durbin-Watson stat | 1.229794 |
| Prob(F-statistic) | 0.000000 | | |

Index of /wp-content/uploads/2014/02 (hedibert.org)

# Resolução do Exercício

| Variáveis explicativas no modelo | $R^2$ |
|---|---|
| Educ | 60,4% |
| Anosemp | 58,6% |
| Educ e Anosemp | 74,0% |

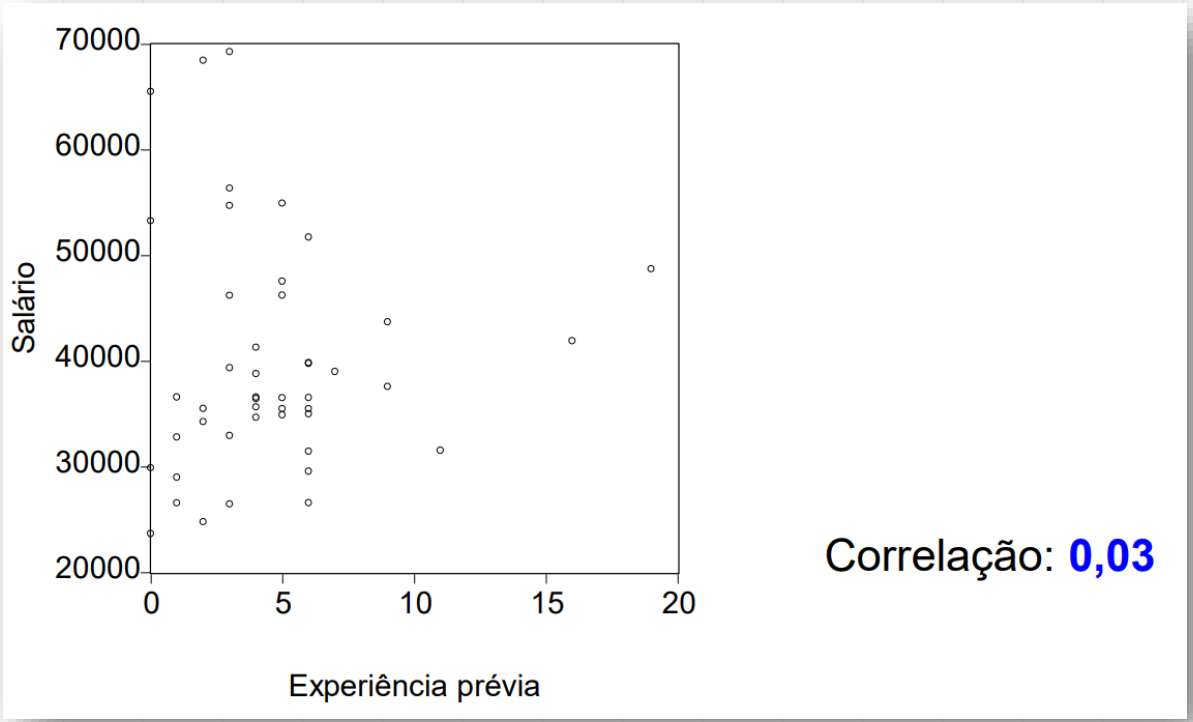
Index of /wp-content/uploads/2014/02 (hedibert.org)

# Resolução do Exercício

**O departamento de RH desconfia que a variável *EXPPREV* (experiência anterior, em anos) é importante para explicar o salário dos funcionários, uma vez que os recém-contratados passam por um treinamento antes de iniciar as atividades na empresa. Pede-se, então: acrescente a variável ao modelo de regressão linear múltipla e verifique o que acontece com o $R^2$?**

Index of /wp-content/uploads/2014/02 (hedibert.org)

# Resolução do Exercício




Index of /wp-content/uploads/2014/02 (hedibert.org)

# Resolução do Exercício

Dependent Variable: SALARIO
Method: Least Squares
Date: 08/22/03 Time: 17:42
Sample: 1 46
Included observations: 46

| Variable | Coefficient | Std. Error | t-Statistic | Prob. |
|---|---|---|---|---|
| C | 23480.46 | 2027.696 | 11.57987 | 0.0000 |
| EDUC | 1925.882 | 384.4395 | 5.009586 | 0.0000 |
| ANOSEMP | 671.3254 | 143.2125 | 4.687618 | 0.0000 |
| EXPPREV | -73.82734 | 232.7840 | -0.317150 | 0.7527 |

| | | | |
|---|---|---|---|
| R-squared | 0.740548 | Mean dependent var | 39827.39 |
| Adjusted R-squared | 0.722016 | S.D. dependent var | 10999.24 |
| S.E. of regression | 5799.262 | Akaike info criterion | 20.25179 |
| Sum squared resid | 1.41E+09 | Schwarz criterion | 20.41080 |
| Log likelihood | -461.7912 | F-statistic | 39.95994 |
| Durbin-Watson stat | 1.250596 | Prob(F-statistic) | 0.000000 |

Index of /wp-content/uploads/2014/02 (hedibert.org)

# Coeficiente de Determinação

**Fato**: **Quanto maior o número de variáveis independentes, maior será o valor de $R^2$.**

Isso pode vir a ser um problema ao se comparar modelos, já que modelos com um número maior de variáveis tenderão a ter um $R^2$ maior do que um modelo, eventualmente equivalente, em termos de qualidade, com um número menor de variáveis.

# Coeficiente de Determinação Ajustado

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

**Nota**: Se o MRLM considerado for este, o número de coeficientes de regressão é p=k+1.

**Valor ajustado pelo número de variáveis**

$$R_a^2 = 1 - \left(1 - R^2\right)\frac{n-1}{n-(k+1)}$$

Formulário

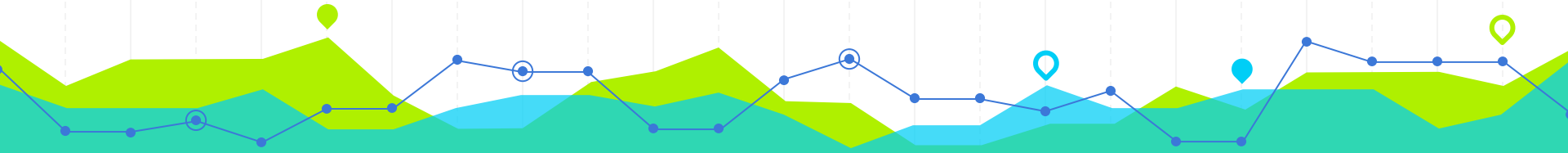
$$\overline{R}^2 = 1 - \frac{VR/(n-k)}{VT/(n-1)} = 1 - (1-R^2)\frac{n-1}{n-k}$$

**O acréscimo de variáveis não acarreta necessariamente um aumento em $R^2_a$.**

# Resolução do Exercício

| Variáveis explicativas no modelo | $R^2$ | $R_a^2$ |
|---|---|---|
| Educ | 60,3% | 59,5% |
| Anosemp | 58,6% | 57,6% |
| Educ e Anosemp | 73,9% | 72,8% |
| Educ, Anosemp e Expprev | 74,1% | 72,2% |

Index of /wp-content/uploads/2014/02 (hedibert.org)



# Modelo de Regressão Linear Múltipla

3

Suposições e Propriedades

# MRLM: Suposições

**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**

$\beta_0, \beta_1, \ldots, \beta_k$ **– são parâmetros desconhecidos (constantes);**

$\varepsilon$ **– termo de erro aleatório não observável.**

# MRLM: Suposições

**MLR.2 – Amostragem Aleatória**

**Temos uma amostra aleatória de $n$ observações**

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

**do modelo populacional descrito em MLR.1.**

# MRLM: Suposições

**MLR.3 – Ausência de Colinearidade Perfeita**

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

# MRLM: Suposições

**MLR.4 – Média Condicional Zero**

**O valor esperado do vetor de erro aleatório, $\varepsilon$, condicionado na matriz de explicação $X$, é igual a zero.**

**Ou seja,**

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

# MRLM: Suposições e Propriedades

**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.**

# MRLM: Suposições

**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 $\sigma^2$).**

**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)

# MRLM: Suposições e Propriedades

## Observação 1

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

# MRLM: Suposições e Propriedades

Curiosidade

## Sob as suposições MLR.1 a MLR.5:

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

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

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

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

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

Index of /wp-content/uploads/2014/02 (hedibert.org)

# MRLM: Variância dos Estimadores

**Variância dos estimadores de MQO:**

Seja o modelo de RLM:

$$\mathbf{y} = \mathbf{X\beta} + \mathbf{e}$$

Caso os pressupostos do Teorema de Gauss-Markov sejam válidos, o método de MQO oferecerá estimadores não viesados para os coeficientes do modelo e para suas respectivas variâncias. As variâncias dos estimadores e seus respetivos estimadores serão dados por:

$$Var(\hat{\boldsymbol{\beta}}) = (\mathbf{X}^T\mathbf{X})^{-1}\sigma^2 \quad \Rightarrow \quad S^2_{\hat{\boldsymbol{\beta}}} = (\mathbf{X}^T\mathbf{X})^{-1}\hat{\sigma}^2$$

Onde $\sigma^2$ é a variânica dos erros ou variância da regressão e $\hat{\sigma}^2$ seu respectivo estimador, dado por:

$$\hat{\sigma}^2 = \frac{\hat{\mathbf{e}}^T\hat{\mathbf{e}}}{n-(k+1)} = \frac{\mathbf{y}^T\mathbf{y} - \hat{\boldsymbol{\beta}}^T\mathbf{X}^T\mathbf{y}}{n-(k+1)}$$

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

MSR (Quadrado Médio devido aos Resíduos)

# MRLM: Suposições e Propriedades

Curiosidade

## 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**

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

# Estimador da Variância

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

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

**Observação**

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

# Eficiência dos Estimadores dos 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 $\beta_0$, $\beta_1$, ..., $\beta_k$, respectivamente.**

# Eficiência dos Estimadores dos MQO

⇒ **Restringindo 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** $\sigma^2 \mathbf{I}_n$ **.**

# 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.

Index of /wp-content/uploads/2014/02 (hedibert.org)

# Propriedades dos Estimadores

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

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

**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 $\hat{\underset{\sim}{\beta}}$, 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 $\left(\underset{\sim}{\mathbf{X}}'\underset{\sim}{\mathbf{X}}\right)^{-1}$.

# Teorema de Gauss-Markov: Resumo

**Método de Mínimos Quadrados**

Seja o modelo de regressão múltipla representado matricialmente por:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$$

O estimador de MQO para o vetor de coeficientes $\boldsymbol{\beta}$ que minimizará o erro quadrático total do modelo será dado por:

$$\hat{\boldsymbol{\beta}}_{p\times 1} = (\mathbf{X}^T\mathbf{X})^{-1}(\mathbf{X}^T\mathbf{y})$$

Caso os pressupostos do Teorema de Gauss-Markov sejam válidos, os estimadores de MQO serão os MELNV dos coeficientes de um modelo de RLM.

**1.** A v.a. $Y_i$ é uma função linear das variáveis explanatórias ($X_{ij}, j=1..k$);
**2.** Os valores de $X_j$ são fixos;
**3.** Os erros possuem esperança condicional zero, ou seja, $E(e_i)=0$;
**4.** Os erros são homocedásticos, ou seja, $E(e_i^2)=\sigma^2$;
**5.** Os erros são não-correlacionados, ou seja, $E(e_i e_j)=0$, para $i\neq j$;

} $E(\mathbf{ee}')=\mathbf{I}\sigma^2$

E, para que tenhamos um modelo clássico de regressão linear:

**6.** Os erros estão normalmente distribuídos;

Maia, Alexandre Gori (2017). Econometria: conceitos e aplicações. Cap. 6.

# Hipóteses Básicas: MRLM

Formulário

**MODELO REGRESSÃO LINEAR** $y_t = \beta_1 + \beta_2 x_{t2} + \ldots + \beta_k x_{tk} + u_t,\ t = 1,2,\ldots,n.$

**Hipóteses básicas**

- **H1 – Linearidade:** $Y = X\beta + U$ (ver definições anteriores de $Y, X, \beta, U$)
- **H2 – Exogeneidade:** $E(u_t \mid X) = 0$ $(t = 1,2,\ldots,n)$. Os **regressores** são **exógenos**.
- **H3 – Homocedasticidade condicionada:** $\mathrm{Var}(u_t \mid X) = \sigma^2 > 0$ $(t = 1,2,\ldots,n)$.
- **H4 – Ausência de autocorrelação -** $\mathrm{Cov}(u_t, u_s \mid X) = 0$ $(t, s = 1,2,\ldots,n; t \neq s)$.
- **H5 – Não existência de multicolinearidade exacta -** A característica da matriz $X$ é igual a $k$ (número de coeficientes de regressão) e $k < n$.

**Nota**: Se o MRLM considerado for este, o número de coeficientes de regressão é p=k+1

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

Murteira et al (2015)

# Propriedades dos Resíduos dos MQ

**Propriedades dos resíduos MQ** (ver livro):

a) A soma dos resíduos é igual a zero (papel do termo independente):

$$\sum_{t=1}^{n} \hat{u}_t = 0.$$

b) A soma dos produtos das observações de cada regressor pelos resíduos é igual a zero:

$$\sum_{t=1}^{n} x_{tj}\hat{u}_t = 0 \quad (j = 2,3,\ldots,k).$$

c) A soma dos produtos dos valores ajustados pelos resíduos é igual a zero:

$$\sum_{t=1}^{n} \hat{y}_t\hat{u}_t = 0.$$

d) A soma dos quadrados das observações do regressando é igual à soma dos quadrados dos respectivos valores ajustados mais a soma dos quadrados dos resíduos:

$$\sum_{t=1}^{n} y_t^2 = \sum_{t=1}^{n} \hat{y}_t^2 + \sum_{t=1}^{n} \hat{u}_t^2.$$

Murteira et al (2015)

Formulário

**Propriedades:**

$$\sum_{t=1}^{n} x_{tj}\hat{u}_t = 0 \quad (j = 1,2,\ldots,k); \quad \sum_{t=1}^{n} \hat{y}_t\hat{u}_t = 0; \quad \sum_{t=1}^{n} y_t^2 = \sum_{t=1}^{n} \hat{y}_t^2 + \sum_{t=1}^{n} \hat{u}_t^2; \quad r_{y\hat{y}}^2 = \frac{\left(\sum_t (y_t - \bar{y}_t)(\hat{y}_t - \bar{\hat{y}}_t)\right)^2}{\sum_t (y_t - \bar{y}_t)^2 \sum_t (\hat{y}_t - \bar{\hat{y}}_t)^2}$$

# Obrigada!

## Questões?