

Regressão Linear Simples

Econometria

Alexandre Gori Maia

Ementa:

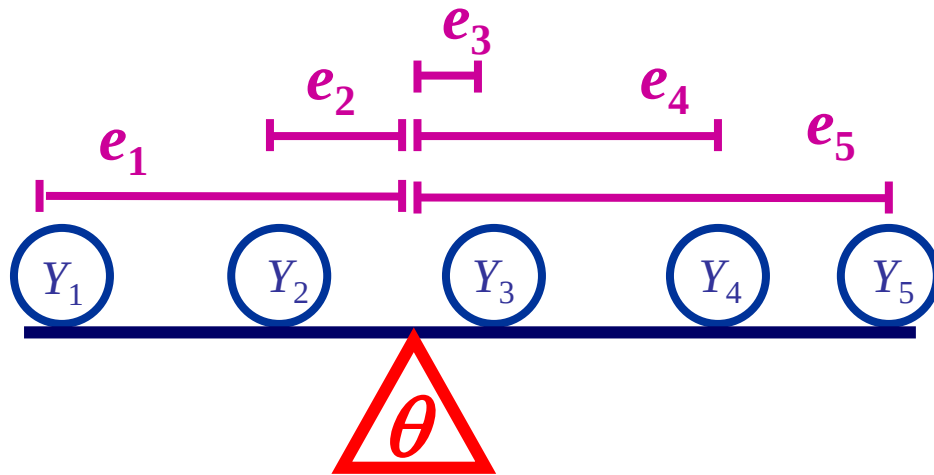
- Definição;
- Estimador de MQO para RLS;
- Teorema de Gauss-Markov;
- Variância dos Estimadores;
- Teste t para os coeficientes;
- Análise de Varância

Bibliografia Básica:

Maia, Alexandre Gori (2017). *Econometria: conceitos e aplicações*. Cap. 1-3; 5.

Mínimos Quadrados - EQT

Supondo que Y seja uma função de θ : $Y_i = \theta + e_i$



$$e_1 = Y_1 - \theta$$

$$e_2 = Y_2 - \theta$$

$$e_3 = Y_3 - \theta$$

$$e_4 = Y_4 - \theta$$

$$e_5 = Y_5 - \theta$$

Os erros e_i representam os erros de mensuração de Y a partir de θ .

1 ° Passo) Definir Erro Quadrático Total:

$$EQT = e_1^2 + e_2^2 + e_3^2 + e_4^2 + e_5^2$$

$$EQT = (Y_1 - \theta)^2 + (Y_2 - \theta)^2 + (Y_3 - \theta)^2 + (Y_4 - \theta)^2 + (Y_5 - \theta)^2$$

$$EQT(\theta) = \sum_{i=1}^5 (Y_i - \theta)^2$$

A função EQT é uma medida do erro total que cometemos ao utilizar θ para estimar Y na função $Y_i = \theta + e_i$.

EQT - Exemplo

Supondo que Y seja uma função de θ : $Y_i = \theta + e_i$

Seja o conjunto de valores de Y

$$Y_1 = 100$$

$$Y_2 = 200$$

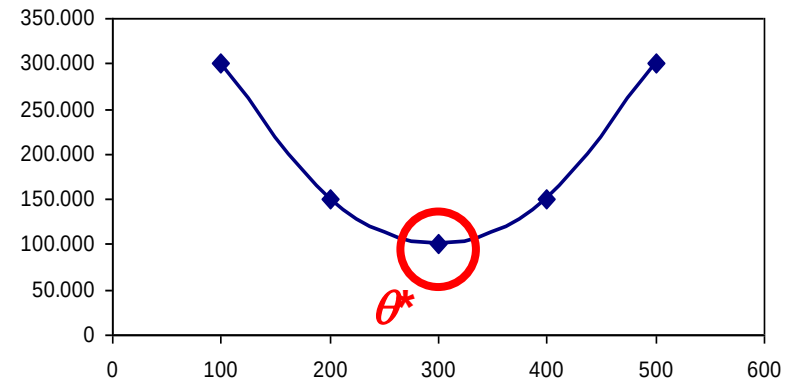
$$Y_3 = 300$$

$$Y_4 = 400$$

$$Y_5 = 500$$

$$EQT(\theta) = \sum_{i=1}^5 (Y_i - \theta)^2$$

EQT(θ)	θ
300.000	100
150.000	200
100.000	300
150.000	400
300.000	500



Para este conjunto de dados, o valor de θ que minimizaria a função $EQT(\theta^*)$ seria o 300.

Estimador de Mínimos Quadrados

Para uma amostra de tamanho n teremos:

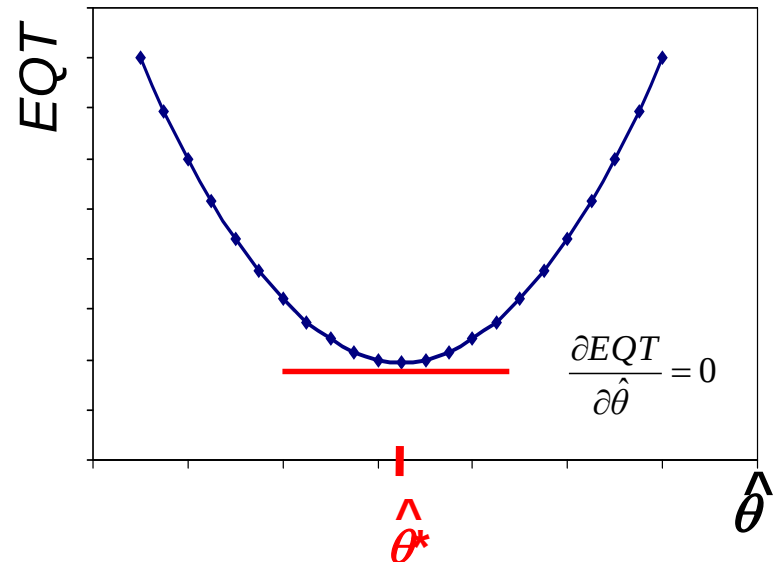
$$Y_i = \hat{\theta} + \hat{e}_i \quad \Rightarrow \quad EQT = \sum_{i=1}^n \hat{e}_i^2 = \sum_{i=1}^n (Y_i - \hat{\theta})^2$$

2 ° Passo) Encontrar $\hat{\theta}$ que minimiza EQT :

$$\frac{dEQT}{d\hat{\theta}} = 0$$

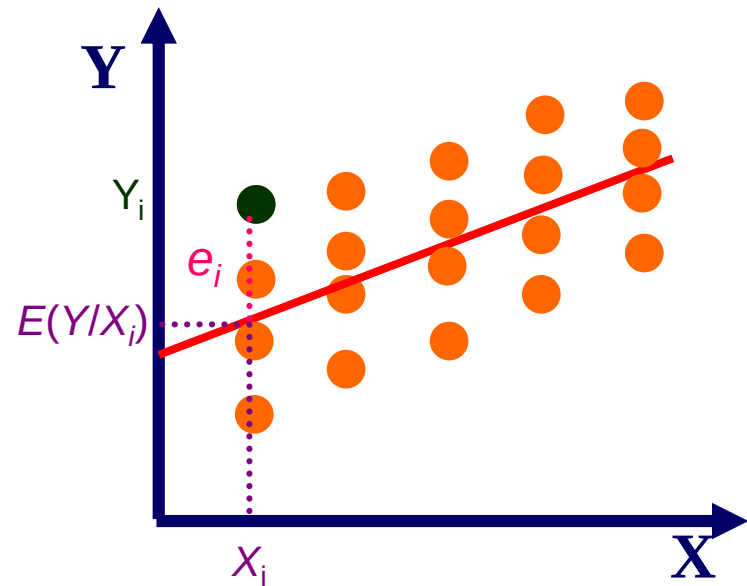
$$\frac{dEQT}{d\hat{\theta}} = \sum_{i=1}^n 2(Y_i - \hat{\theta})(-1) = -2 \sum_{i=1}^n Y_i + 2 \sum_{i=1}^n \hat{\theta}$$

$$\frac{dEQT}{d\hat{\theta}} = -2 \sum_{i=1}^n Y_i + 2n\hat{\theta} = 0 \quad \Rightarrow \quad \hat{\theta} = \frac{\sum_{i=1}^n Y_i}{n}$$



Função de Regressão Populacional

Seja a relação entre Y e X na população:



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

ou

$$E(Y/X_i) = \alpha + \beta X_i$$

Modelo de Regressão
Linear Simples para Y
na população

Onde:

Y é a variável dependente ou regressando

X é a variável independente ou explanatória

α é o intercepto ou constante do modelo

β é o coeficiente angular do modelo

Erro de previsão:

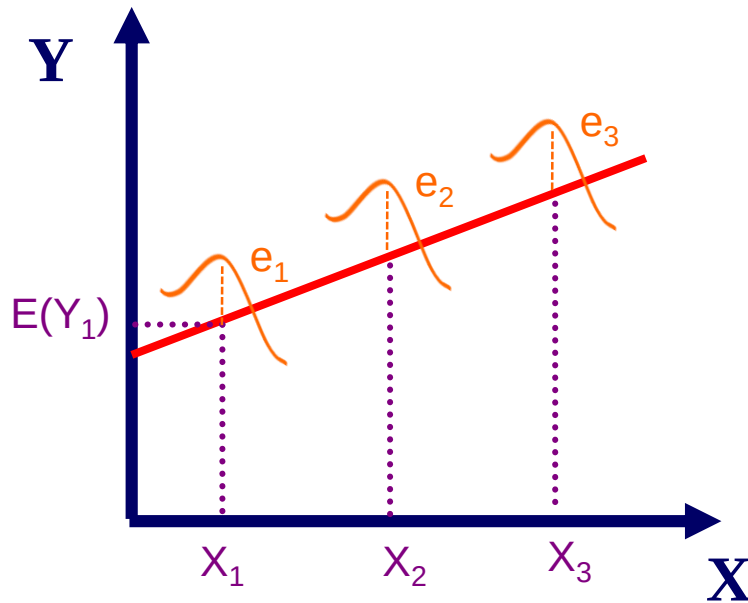
Seja X_i a i -ésima observação de X , teremos:

Y_i é o valor observado em Y para o i -ésimo valor de X

$E(Y/X_i)$ é a esperança condicional de Y e representa o valor esperado de Y para o i -ésimo valor de X

e_i é o erro, ou variação de Y_i não explicada pelo modelo

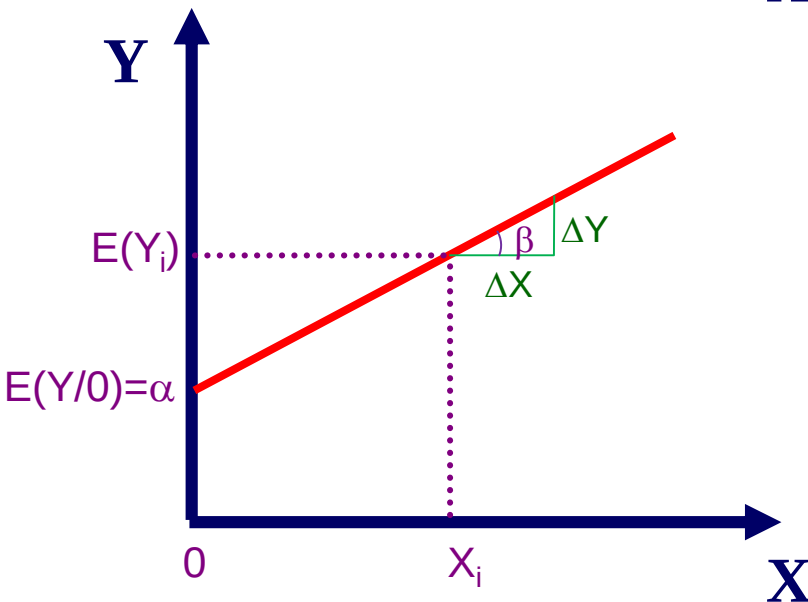
Interpretação dos Coeficientes



Para cada valor controlado de X, teremos um valor esperado de Y.

O erro representa variáveis omitidas ou mesmo dificuldades para mensurar aquelas presentes no modelo.

Pressupõe-se que o efeito do erro seja mínimo e que este esteja aleatoriamente distribuído em torno da reta de regressão.



Para compreendermos a interpretação dos coeficientes do modelo:

O intercepto α representa o valor esperado de Y quando o valor controlado de X for nulo:

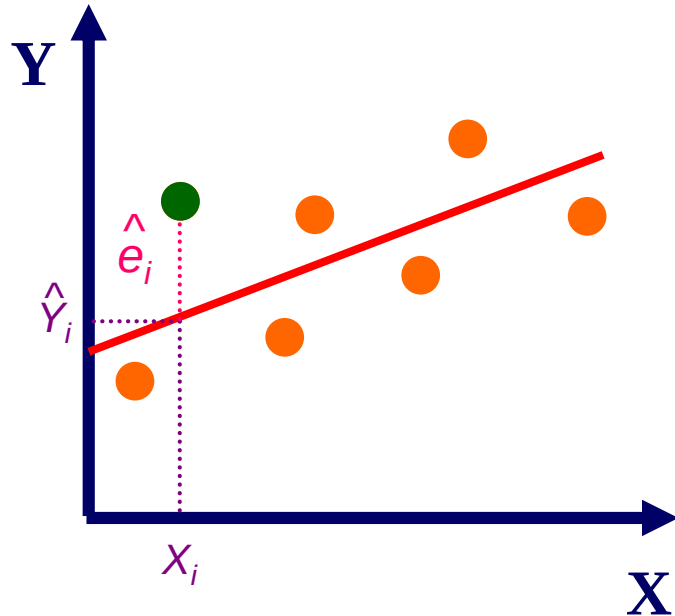
$$E(Y / 0) = \alpha + \beta(0) = \alpha$$

O coeficiente angular β , por sua vez, representa a variação marginal no valor esperado de Y dada uma variação unitária em X

$$\frac{\Delta E(Y)}{\Delta X} = \frac{dE(Y)}{dX} = \frac{d(\alpha + \beta X)}{dX} = \beta$$

Função de Regressão Amostral

Em uma amostra, a relação entre Y e X será dada por:



Função de regressão amostral:

$$Y_i = \hat{\alpha} + \hat{\beta}X_i + \hat{e}_i$$

Y previsto pelo ajuste:

$$\hat{Y}_i = \hat{\alpha} + \hat{\beta}X_i$$

Resíduo, valor não previsto pelo ajuste:

$$\hat{e}_i = Y_i - \hat{Y}_i$$

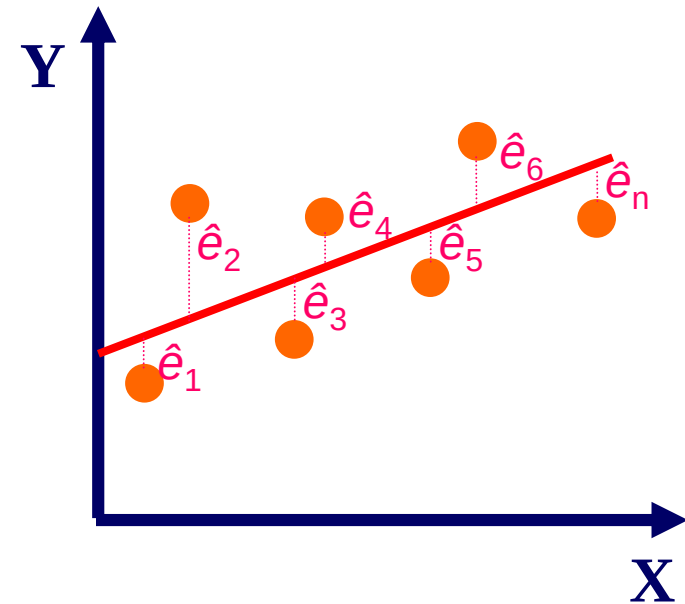
Função de Erro Quadrático Total (EQT):

$$EQT = \hat{e}_1^2 + \hat{e}_2^2 + \dots + \hat{e}_n^2$$

$$EQT = (Y_1 - \hat{Y}_1)^2 + (Y_2 - \hat{Y}_2)^2 + \dots + (Y_n - \hat{Y}_n)^2$$

$$EQT = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^n [Y_i - (\hat{\alpha} + \hat{\beta}X_i)]^2$$

Método de Mínimos Quadrados



Mínimos Quadrados Ordinários:

O método de mínimos quadrados ordinários (MQO) ajustará a reta que apresentar as menores distâncias quadráticas entre os valores observados e a reta ($\hat{e}_i$). Em outras palavras, os estimadores de MQO dos parâmetros α e β minimizarão o valor da função de Erro Quadrático Total (EQT), dada por:

$$EQT = \sum_{i=1}^n [Y_i - (\hat{\alpha} + \hat{\beta}X_i)]^2$$

Para encontrar os valores que minimizam a função EQT:

$$\frac{\partial EQT}{\partial \hat{\alpha}} = 2 \sum_{i=1}^n [Y_i - (\hat{\alpha} + \hat{\beta}X_i)](-1) = 0 \quad (1) \quad \Rightarrow \quad \hat{\alpha} = \bar{Y} - \hat{\beta}_1 \bar{X}$$

$$\frac{\partial EQT}{\partial \hat{\beta}} = 2 \sum_{i=1}^n [Y_i - (\hat{\alpha} + \hat{\beta}X_i)](-X_i) = 0 \quad (2) \quad \Rightarrow \quad \hat{\beta} = \frac{\sum_{i=1}^n X_i Y_i - n \bar{X} \bar{Y}}{\sum_{i=1}^n X_i^2 - n \bar{X}^2}$$

Estimadores de MQO

Estimadores de Mínimos Quadrados Ordinários:

Seja o modelo de Regressão Linear Simples (RLS) para o conjunto de dados populacionais e sua equivalente representação na amostra:

$$Y_i = \alpha + \beta X_i + e_i \quad \Rightarrow \quad Y_i = \hat{\alpha} + \hat{\beta} X_i + \hat{e}_i \quad (3)$$

Os estimadores de MQO para os parâmetros α e β , ou seja, os estimadores que minimizam o valor de $\sum \hat{e}_i^2$, serão dadas por:

$$\hat{\alpha} = \bar{Y} - \hat{\beta} \bar{X} \quad \text{e} \quad \hat{\beta} = \frac{\sum_{i=1}^n X_i Y_i - n \bar{X} \bar{Y}}{\sum_{i=1}^n X_i^2 - n \bar{X}^2}$$

Para simplificar os cálculos, podemos demonstrar que essas expressões podem também ser dadas como função dos desvios de X e Y em relação às suas respectivas médias amostrais:

$$\hat{\alpha} = \bar{Y} - \hat{\beta} \bar{X} \quad (4) \quad \text{e} \quad \hat{\beta} = \frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2} = \frac{\sum_{i=1}^n X_i Y_i}{\sum_{i=1}^n X_i^2} = \frac{\sum_{i=1}^n x_i Y_i}{\sum_{i=1}^n x_i^2} \quad (5)$$

$$\text{Onde: } x_i = X_i - \bar{X} \quad \text{e} \quad y_i = Y_i - \bar{Y}$$

Estimadores de MQO - Exemplo

Seja a relação entre renda familiar em salários mínimos (Y), anos de estudo (X) do responsável pela família:

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

A partir dos dados de uma amostra de 4 observações, os estimadores de MQO para a função amostral serão:

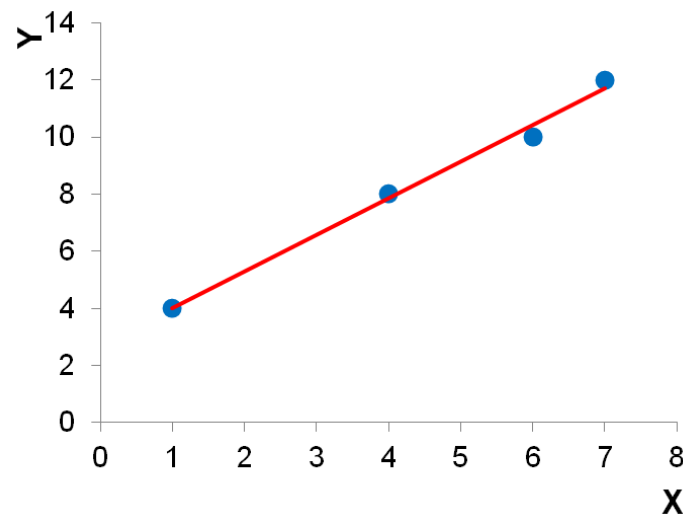
$$Y_i = \hat{\alpha} + \hat{\beta}X_i + \hat{e}_i \quad \Rightarrow \quad Y_i = 2,71 + 1,29X_i + \hat{e}_i$$

$$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X} = 8,5 - 1,29(4,5) = 2,71$$

$$\hat{\beta} = \frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2} = \frac{27}{21} = 1,29$$

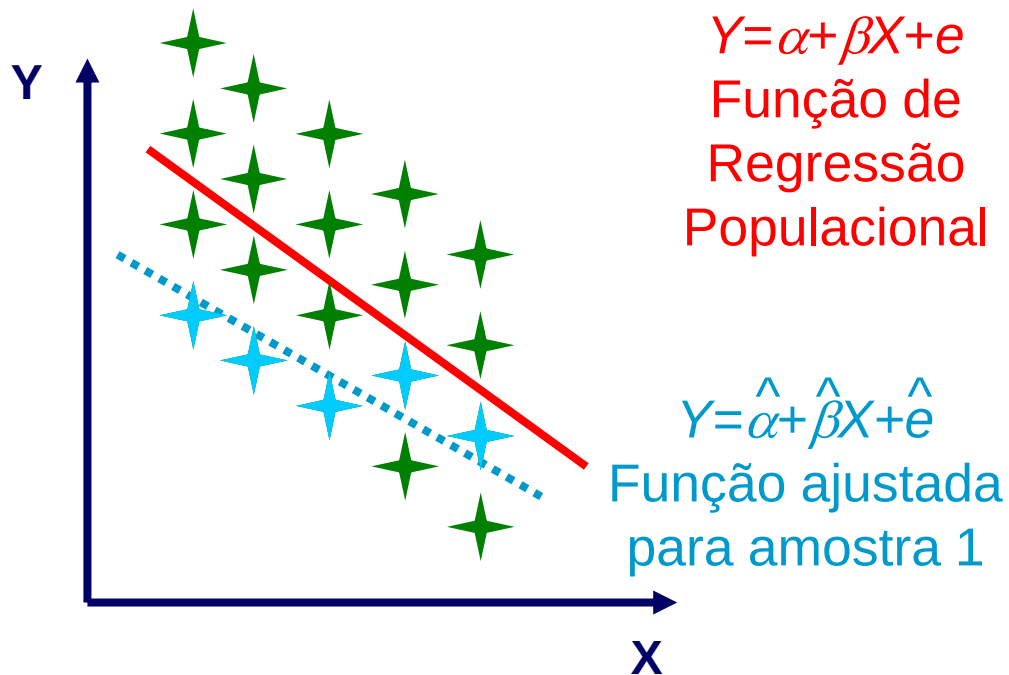
Para cada ano adicional de escolaridade da pessoa responsável, espera-se um acréscimo de 1,98 SM na renda familiar. O rendimento esperado para famílias lideradas por pessoas sem escolaridade seria de 2,71 SMs.

Y (Renda)	X (Anos Estudo)
4	1
8	4
10	6
12	7



Distribuição Amostral dos Estimadores

Seja uma população com 20 observações:



Os estimadores de MQO variarão aleatoriamente em função dos valores observados na amostra.

Teorema de Gauss-Markov

Teorema de Gauss-Markov

Sob a hipótese da veracidade de alguns pressupostos, os estimadores obtidos pelo método de mínimos quadrados ordinários serão os Melhores Estimadores Lineares Não Viesados (MELNV, ou *BLUE* - *Best Linear Unbiased Estimator*).

Em outras palavras:

$$E(\hat{\beta}) = \beta \quad \text{estimador não viesado}$$

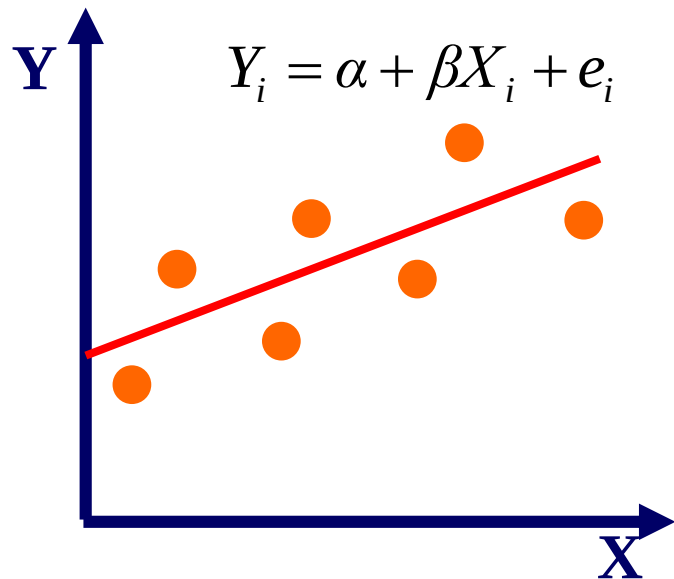
E:

$$\text{Var}(\hat{\beta}) < \text{Var}(\hat{\beta}')$$

estimador mais eficiente que $\hat{\beta}'$,
qualquer que seja $\hat{\beta}'$ outro estimador
de β

1 – Relação Linear

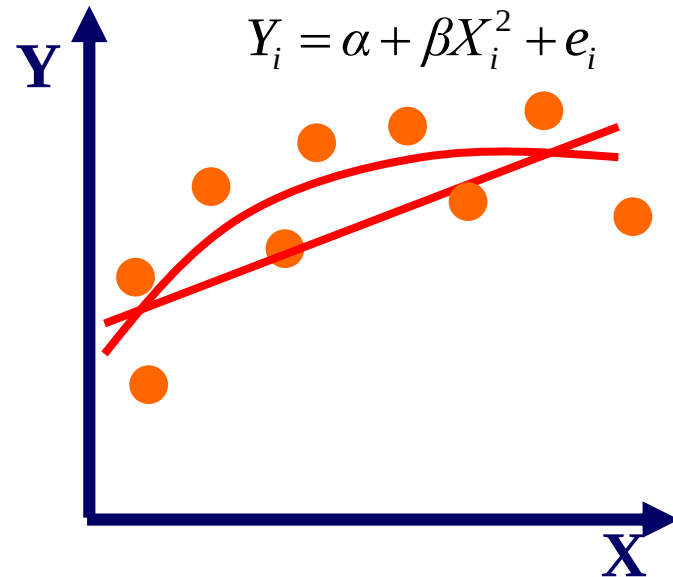
Linearidade nos parâmetros e nas variáveis



Modelo é **linear nas variáveis** pois todos os expoentes de Y e X são iguais a 1.

Modelo é **linear nos parâmetros** pois todos os expoentes dos parâmetros são iguais a 1.

Linearidade nos parâmetros

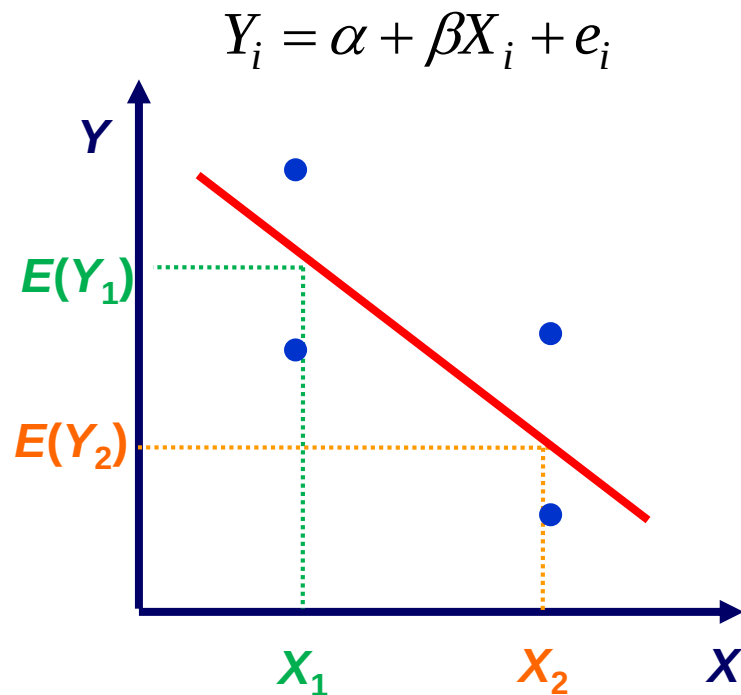


Modelo **não é linear nas variáveis** pois o expoente de X não é igual a 1.

Modelo é **linear nos parâmetros** pois todos os expoentes dos parâmetros são iguais a 1.

2- Valores de X são Fixos

Pressupõe que o valor de X seja controlado em repetidas amostras para se observar variações aleatórias de Y .

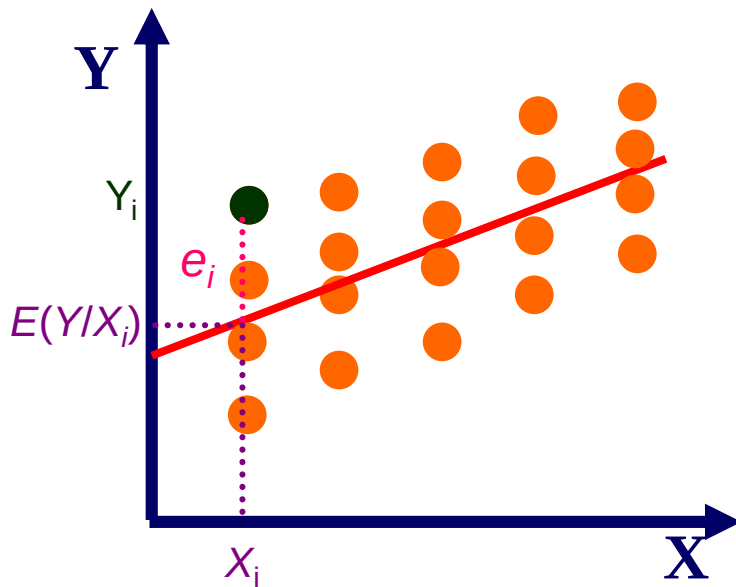


Premissa utilizada em inúmeras demonstrações, embora seja, muitas vezes, pouco factível. Nessas circunstâncias, cuidados adicionais deverão ser tomados para que as estimativas de MQO não sejam viesadas.

3 - Esperança Condicional Zero

Os erros apresentam esperança condicional zero, ou seja, esses não podem estar associados aos regressores.

$$E(e | X_i) = E(e_i) = 0 \quad \Rightarrow \quad E(Y / X_i) = \alpha + \beta X_i$$



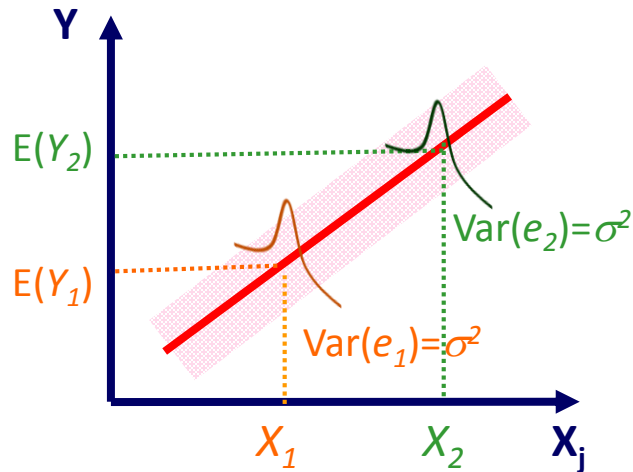
A média dos erros associados a cada valor controlado de X será sempre 0, não importa qual seja o valor de X .

4 - Homocedasticidade

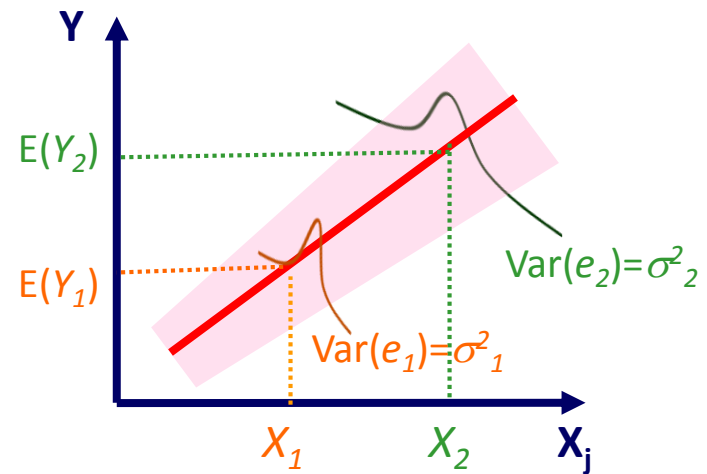
A variabilidade dos erros deve ser constante, qualquer que seja o valor de X

$$\text{Var}(e_i / X_i) = \sigma^2$$

Homocedasticidade



Heterocedasticidade

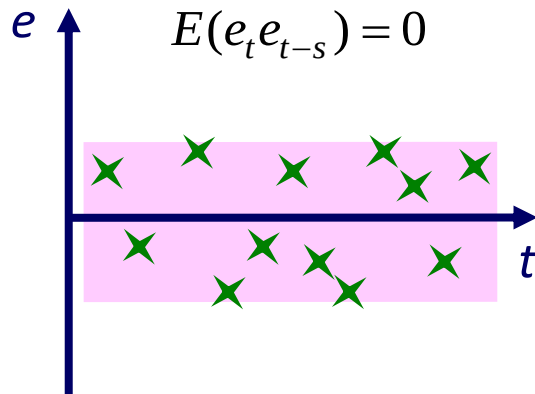


5 – Ausência de Autocorrelação

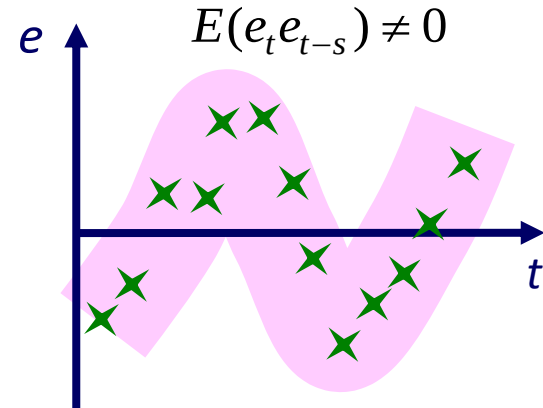
Não há correlação entre os erros do modelo regressão:

$$\text{Cov}(e_t, e_{t+s}) = E(e_t e_{t-s}) \neq 0$$

Não Autocorrelacionado



Autocorrelacionado



A autocorrelação é frequente em análise de séries temporais de dados espaciais.

Modelo Clássico de Regressão Linear

1) Relação Linear entre X e Y :

O ajuste só é válido para relações lineares.

2) Os valores de X são fixos em repetidas amostras, não aleatórios:

Quem varia é o regressando, o regressor é fixo e dado, qualquer que seja a amostra. Fazemos a pressuposição que dado um valor de X , Y irá variar segundo uma distribuição de probabilidade com valor esperado dado por $E(Y|X_i)$.

3) Esperança condicional dos erros igual a zero, ou seja, $E(e|X_i)=0$:

É a mesma coisa afirmar que $E(Y|X_i)=\alpha+\beta X_i$.

4) A variabilidade dos erros é constante, ou seja, $E(e_i^2)=\sigma^2$

Os erros são homocedásticos, ou seja, sua variância é a mesma para qualquer X .

5) Os erros são não autocorrelacionados, ou seja, $E(e_i e_j)=0$:

Não há relação entre valores ordenados dos erros segundo tempo ou espaço.

6) Os erros apresentam distribuição normal:

Não é um pressuposto necessário para que os estimadores de MQO sejam MELNV, mas necessário para que as inferências sejam válidas.

Variância dos Estimadores

Variância dos estimadores de MQO:

Seja o modelo de RLS:

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

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 respectivos estimadores serão dados por:

$$\text{Var}(\hat{\alpha}) = E(\hat{\alpha} - \alpha)^2 = \frac{\sum X_i^2}{n \sum x_i^2} \sigma^2 \quad \Rightarrow \quad S_{\hat{\alpha}}^2 = \frac{\sum X_i^2}{n \sum x_i^2} \hat{\sigma}^2 = \left[\frac{1}{n} + \frac{\bar{X}^2}{\sum x_i^2} \right] \hat{\sigma}^2$$

$$\text{Var}(\hat{\beta}) = E(\hat{\beta} - \beta)^2 = \frac{\sigma^2}{\sum x_i^2} \quad \Rightarrow \quad S_{\hat{\beta}}^2 = \frac{\hat{\sigma}^2}{\sum x_i^2}$$

Onde $\sigma^2 = \text{Var}(e_i)$ é a variância dos erros ou variância da regressão e $\hat{\sigma}^2$ seu respectivo estimador, dado por:

$$\hat{\sigma}^2 = \frac{\sum \hat{e}_i^2}{n-2} \quad \text{onde} \quad \sum \hat{e}_i^2 = \sum y_i^2 - \hat{\beta} \sum x_i y_i$$

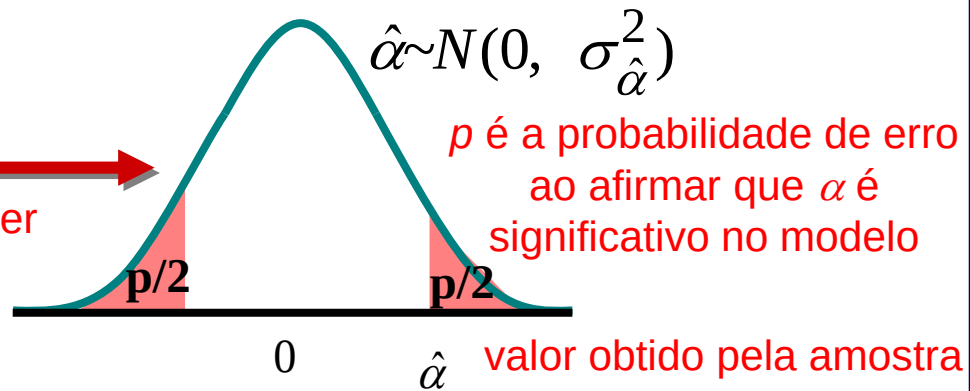
Teste t para os Estimadores

Então, para verificar que os estimadores de MQO são significativos:

Teste de hipótese para α :

$$\begin{cases} H_0: \alpha=0 \\ H_1: \alpha \neq 0 \end{cases}$$

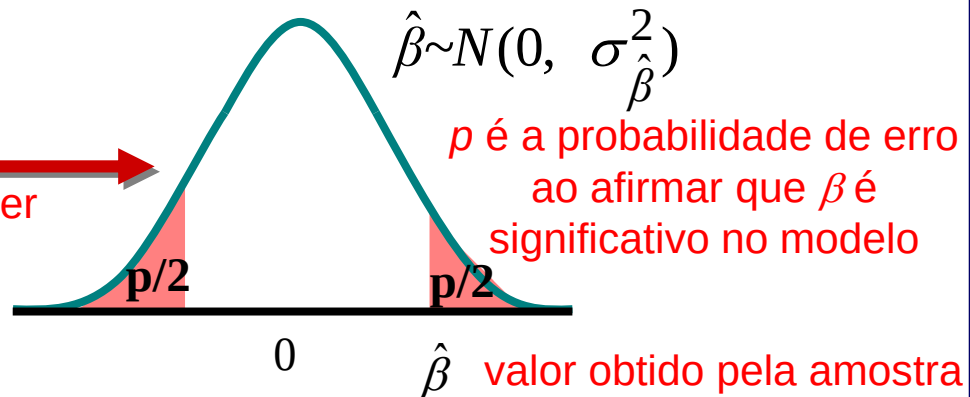
Sob a hipótese de H_0 ser verdadeiro temos...



Teste de hipótese para β_1 :

$$\begin{cases} H_0: \beta=0 \\ H_1: \beta \neq 0 \end{cases}$$

Sob a hipótese de H_0 ser verdadeiro temos...



Variância dos Estimadores - Exemplo

Seja a relação entre renda familiar em salários mínimos (Y), anos de estudo (X) do responsável pela família:

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

A partir dos dados de uma amostra de 4 observações, as estimativas de MQO para a função amostral serão:

$$Y_i = \hat{\alpha} + \hat{\beta}X_i + \hat{e}_i \quad \Rightarrow \quad Y_i = 2,71 + 1,29X_i + \hat{e}_i$$

As estimativas das variâncias dos erros e dos estimadores de MQO serão:

$$\hat{\sigma}^2 = \frac{\sum \hat{e}_i^2}{n-2} = \frac{0,2857}{4-2} = 0,1429 = 0,3780^2$$

$$S_{\hat{\alpha}}^2 = \left[\frac{1}{n} + \frac{\bar{X}^2}{\sum x_i^2} \right] \hat{\sigma}^2 = \left[\frac{1}{4} + \frac{4,5^2}{21} \right] 0,1429 = 0,1735 = 0,4165^2$$

$$S_{\hat{\beta}}^2 = \frac{\hat{\sigma}^2}{\sum x_i^2} = \frac{0,1429}{21} = 0,0068 = 0,0825^2$$

Y (Renda)	X (Anos Estudo)
4	1
8	4
10	6
12	7

Teste t - Exemplo

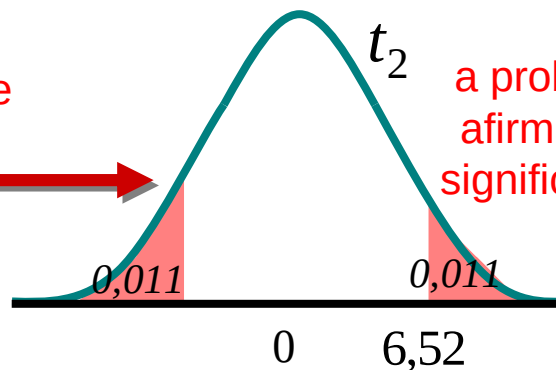
Conhecendo os erros padrões e a fdp dos estimadores, temos que:

Teste de hipótese para β_0 :

$$\begin{cases} H_0: \alpha=0 \\ H_1: \alpha \neq 0 \end{cases}$$

Aplicando a estatística de teste t_{n-2}

$$t = \frac{\hat{\alpha} - 0}{S_{\hat{\alpha}}} = \frac{2,71}{0,42}$$



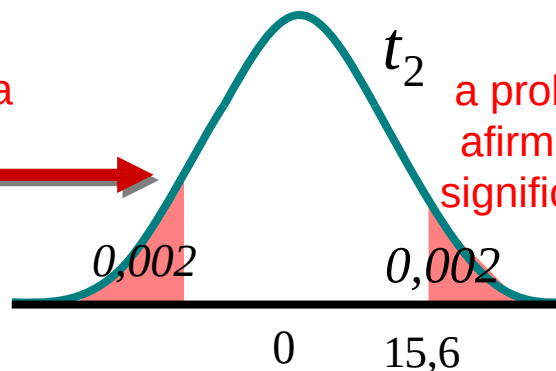
a probabilidade de erro ao afirmar que o intercepto é significativo no modelo é de 0,2%

Teste de hipótese para β_1 :

$$\begin{cases} H_0: \beta=0 \\ H_1: \beta \neq 0 \end{cases}$$

Aplicando a estatística de teste t_{n-2}

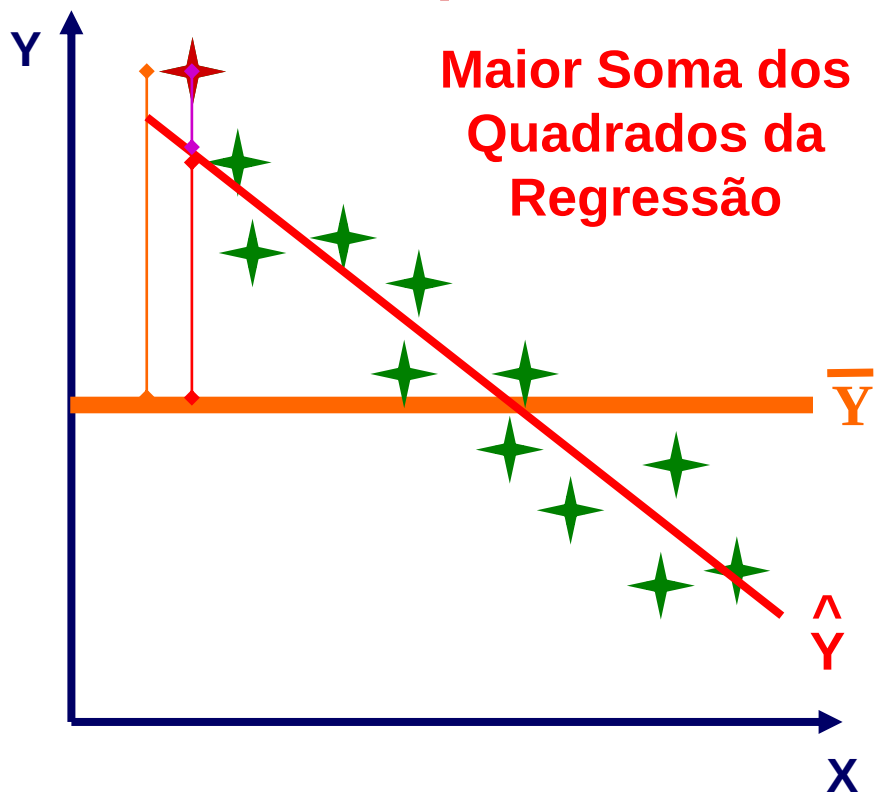
$$t = \frac{\hat{\beta} - 0}{S_{\hat{\beta}}} = \frac{1,29}{0,08}$$



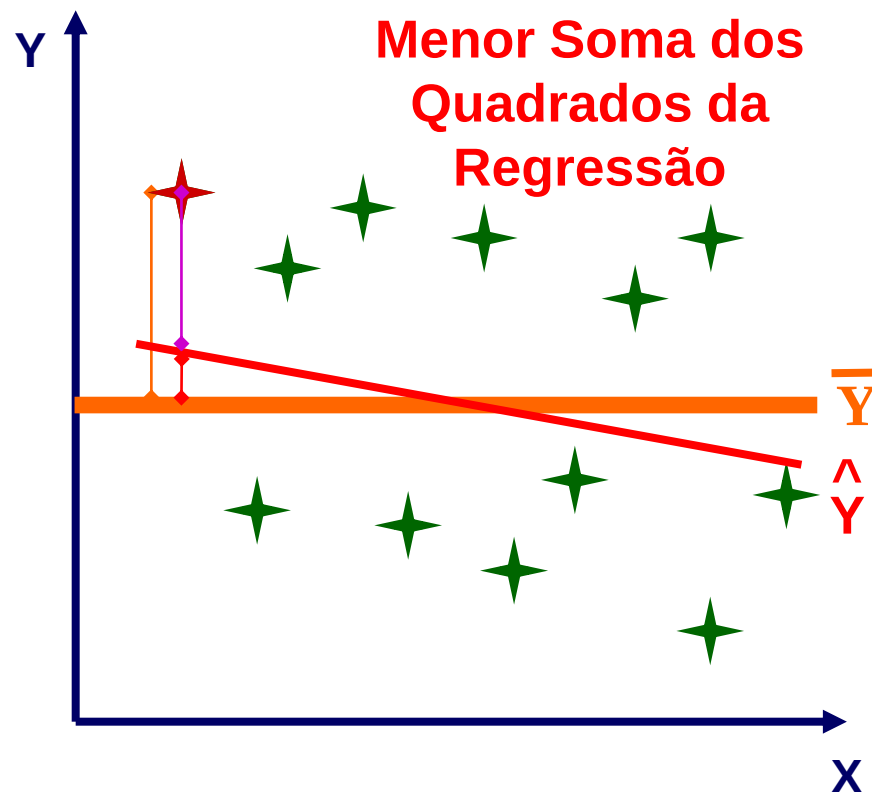
a probabilidade de erro ao afirmar que a variável X é significativa no modelo é de 0,4%

Análise de Variância

Quando X explica Y



Quando X não explica Y



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

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

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

Soma dos Quadrados

Soma Total dos Quadrados (STQ)

Mensura a variabilidade total da variável dependente.

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

Soma dos Quadrados da Regressão (SQReg)

Mensura a variabilidade da variável dependente explicada pelo modelo.

$$SQReg = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 = \hat{\beta}_1^2 \sum_{i=1}^n x_i^2 = \hat{\beta}_1 \sum_{i=1}^n x_i y_i$$

Soma dos Quadrados da Regressão (SQReg)

Mensura a variabilidade residual, não explicada pelo modelo.

$$SQRes = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^n \hat{e}_i^2 = \sum_{i=1}^n y_i^2 - \hat{\beta}_1 \sum_{i=1}^n x_i y_i$$

Coeficiente de Determinação

Coeficiente de Determinação (R^2):

Definição: Estima a proporção da variabilidade da variável dependente (Y) que é explicada pela(s) variável(eis) independente(s) do modelo de regressão.

$$R^2 = \frac{SQReg}{STQ} = \frac{\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2} = 1 - \frac{\sum_{i=1}^n \hat{e}_i^2}{\sum_{i=1}^n y_i^2} = \hat{\beta}_1^2 \frac{\sum_{i=1}^n x_i^2}{\sum_{i=1}^n y_i^2}$$

Escala de R^2 :



0
independência
linear

A significância das escalas depende muito da
natureza da variável dependente

1
relação
linear exata

Análise de Variância

Problema: O modelo contribui para explicar Y?

$$\begin{cases} H_0: \beta_1 = 0 \text{ (o modelo não contribui)} \\ H_1: \beta_1 \neq 0 \text{ (o modelo contribui)} \end{cases}$$

Dados:

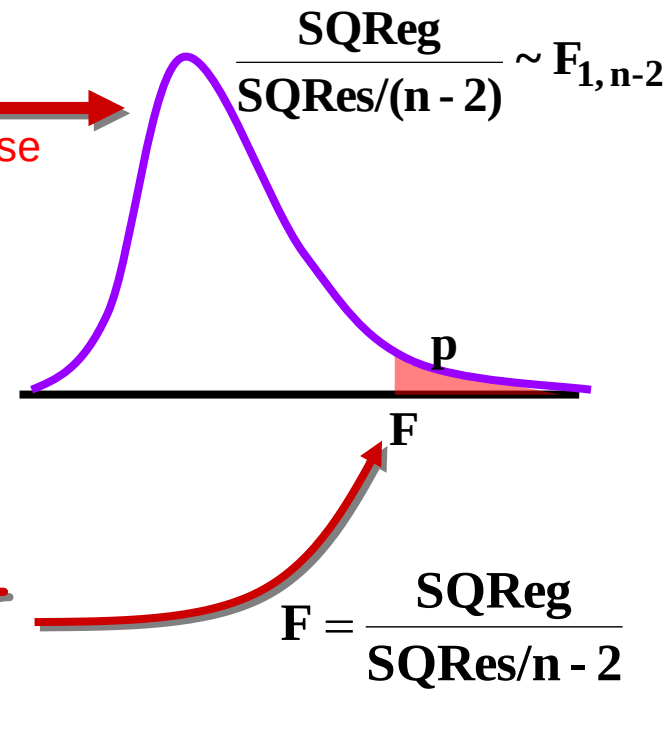
α = nível de significância

SQReg = Soma dos Quadrados da Regressão

SQRes = Soma dos Quadrados dos Resíduos

n = número de observações

Sob a hipótese
de H_0 ser
verdadeiro
temos...



Conclusão:

Rejeito H_0 se valor $p \leq \alpha$.

Análise de Variância

Tabela ANOVA (Analysis Of Variance)

Fonte	GL	Soma dos Quadrados	Quadrados Médios	F
Regressão	1	$SQReg = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2$	$\frac{SQReg}{1}$	$F = \frac{SQReg/1}{SQRes/(n-2)}$
Resíduos	n - 2	$SQRes = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$	$\frac{SQRes}{n - 2}$	
Total	n - 1	$STQ = \sum_{i=1}^n (Y_i - \bar{Y})^2$		

Rejeito H_0 quando valor p – prob ($F > F_{1, n-2}$) - for menor que nível de significância esperado (α).