Análise de Variância

Econometria

Alexandre Gori Maia

Ementa:

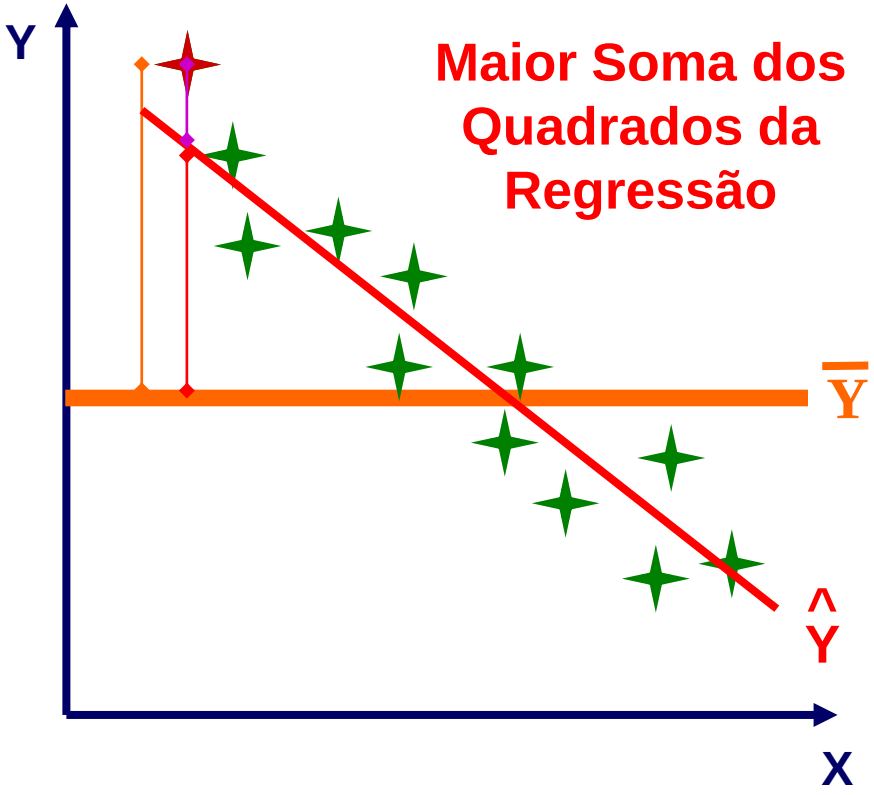
- Soma dos Quadrados;
- Coeficiente de Determinação;
- Teste *F* para Tabela ANOVA;
- Coeficiente de Determinação Ajustado;

Bibliografia Básica:

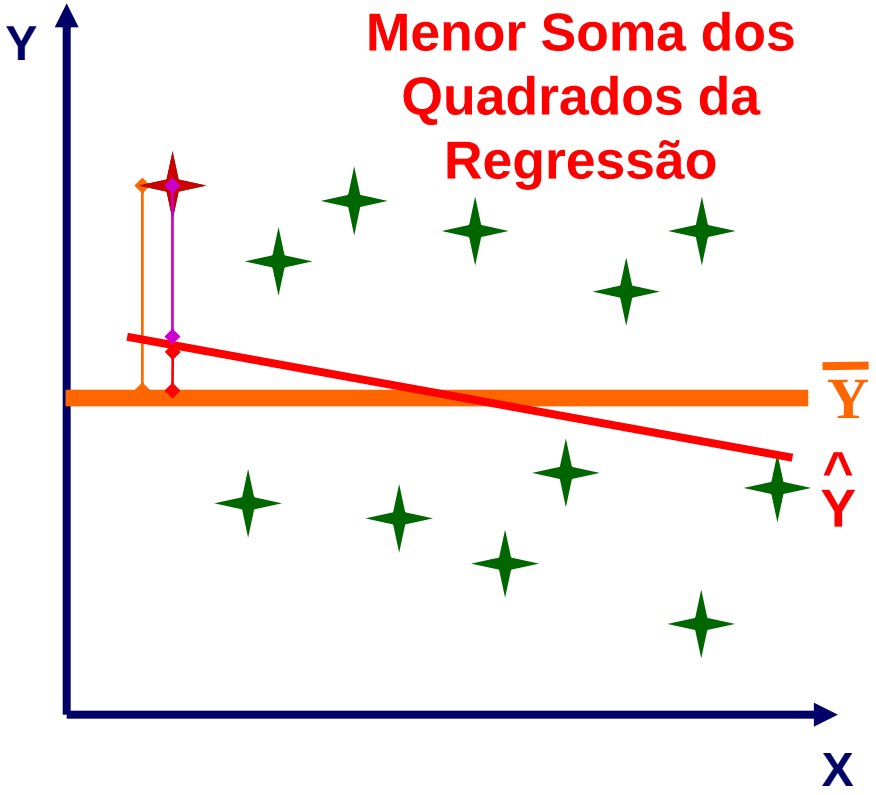
Maia, Alexandre Gori (2017). *Econometria: conceitos e aplicações*. Cap. 7.

Soma dos Quadrados - Conceito

Quando X explica Y



Quando X não explica Y



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

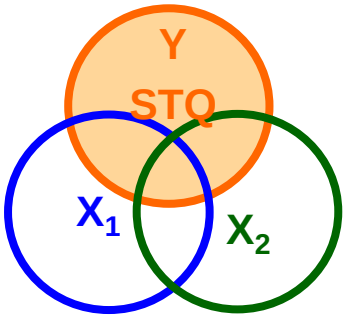
$$SQReg = \sum_{i=1}^{n} (\hat{Y}_i - \overline{Y})^2$$

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

Soma dos Quadrados - Definição

Soma Total dos Quadrados (STQ):

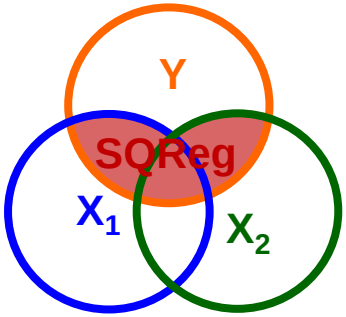
$$STQ = \sum_{i=1}^{n} (Y_i - \bar{Y})^2 = \sum_{i=1}^{n} y_i^{\,2} = \mathbf{y}^T\mathbf{y} - n\bar{Y}^2$$



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

Soma dos Quadrados da Regressão (SQReg):

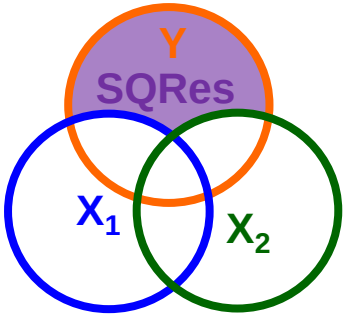
$$SQReg = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2 = \hat{\boldsymbol{\beta}}^T\mathbf{X}^T\mathbf{y} - n\bar{Y}^2$$



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.

Análise de Variância

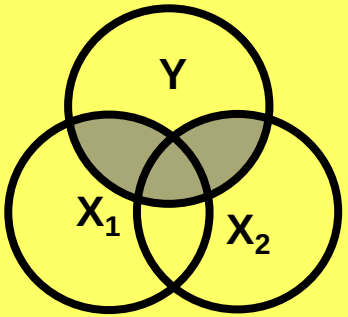
Fonte	GL	Soma dos Quadrados	Quadrados Médios	F
Regressão	k	$\hat{\boldsymbol{\beta}}^T \mathbf{X}^T \mathbf{y} - n\bar{Y}^2$	$\frac{SQReg}{k}$	$F = \frac{SQReg/k}{SQRes/(n-k-1)}$
Resíduos	$n-(k+1)$	$\mathbf{y}^T \mathbf{y} - \hat{\boldsymbol{\beta}}^T \mathbf{X}^T \mathbf{y}$	$\frac{SQRes}{n-(k+1)}$	
Total	$n-1$	$\mathbf{y}^T \mathbf{y} - n\bar{Y}^2$		

Coeficiente de Determinação

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

Soma dos Quadrados - 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	

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

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

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

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

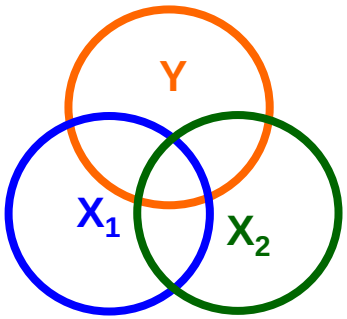
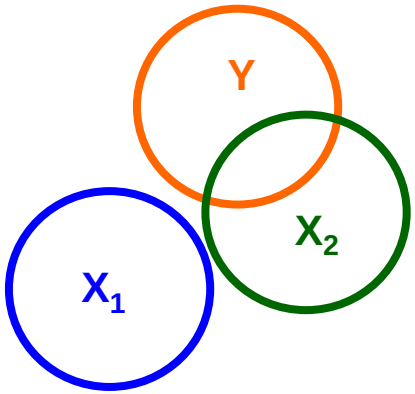
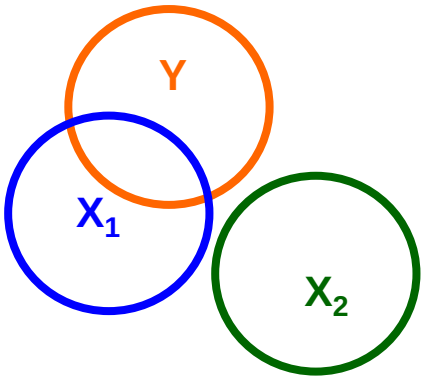
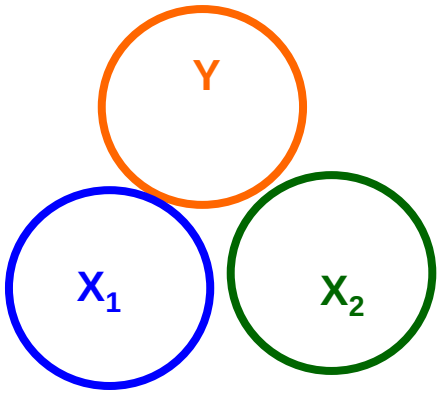
$$SQRes = STQ - SQReg = 35 - 34{,}8 = 0{,}2$$

Teste *F* para ANOVA- Exemplo

Seja o modelo de RLM com duas variáveis: $Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + e$

E as hipóteses:
$$\begin{cases} H_0: \ \beta_1 = \beta_2 = 0 \\ H_1: \ Pelo\ menos\ um\ \beta_j \neq 0 \end{cases}$$

Possíveis resultados do modelo:

			
$\beta_1 \neq 0$ $\beta_2 \neq 0$	$\beta_1 = 0$ $\beta_2 \neq 0$	$\beta_1 \neq 0$ $\beta_2 = 0$	$\beta_1 = 0$ $\beta_2 = 0$
X_1 e X_2 contribuem para explicar Y. H_0 deveria ser rejeitado	Apenas X_2 contribui para explicar Y. H_0 deveria ser rejeitado	Apenas X_1 contribui para explicar Y. H_0 deveria ser rejeitado	Nenhuma variável contribui para explicar Y. H_0 não deveria ser rejeitado

Teste *F* para ANOVA

Seja o modelo de RLM:

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k + e$$

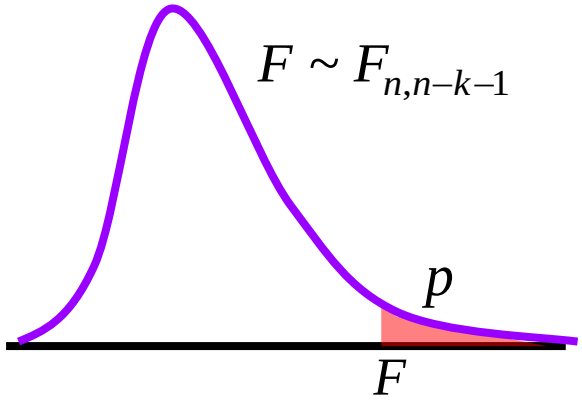
Para testarmos a contribuição do conjunto de *k* variáveis independentes do modelo, teremos as hipóteses:

- **H_0:** β_1=...=β_k=0 (não contribui)
- **H_1:** Pelo menos um $\beta_j \neq 0$ (contribui)

A estatística de teste será

$$F = \frac{SQReg/k}{SQRes/(n-k-1)}$$

Considerando H_0 verdadeiro, a fdp de *F* será...

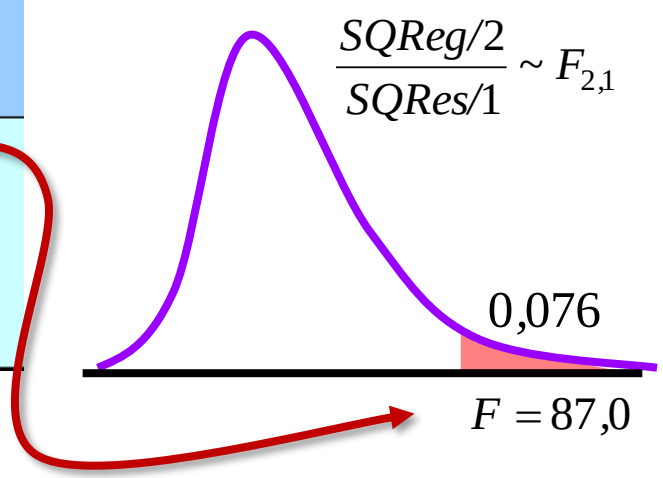


Rejeitar H_0 significa afirmar que o modelo contribui para explicar *Y*, ou seja, há relação significativa entre pelo menos uma variável explicativa e a variável dependente.

Teste *F* para ANOVA - 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		



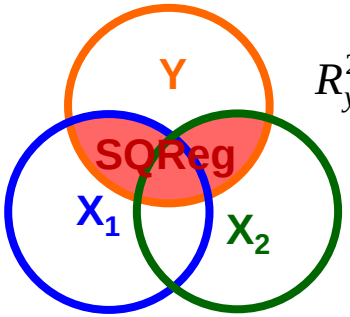
valor *p* = 0,076

Há evidências moderadas para afirmar que o modelo contribui para explicar a variabilidade da renda familiar. A probabilidade de erro ao fazermos tal afirmação é de aproximadamente 7,6%.

R^2 Ajustado - Definição

Seja o ajuste:

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



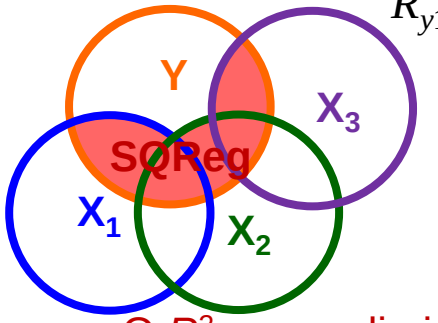
$$R^2_{y12} = \frac{SQReg}{STQ}$$

Incorporando uma variável independente adicional (X_3):

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

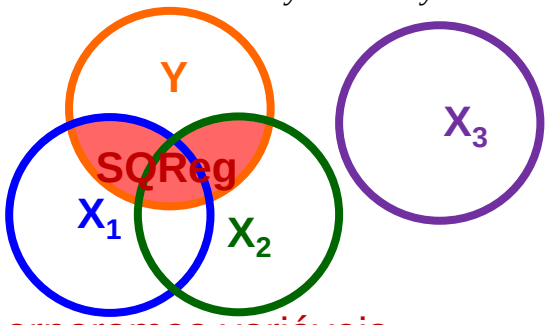
Poderemos ter:

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



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:

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

R^2 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.