Revisão de Regressão Múltipla

Prof. Alexandre Gori Maia

Universidade Estadual de Campinas

Disciplina: Econometria Aplicada II

Universidade Nacional Agraria La Molina

Ementa

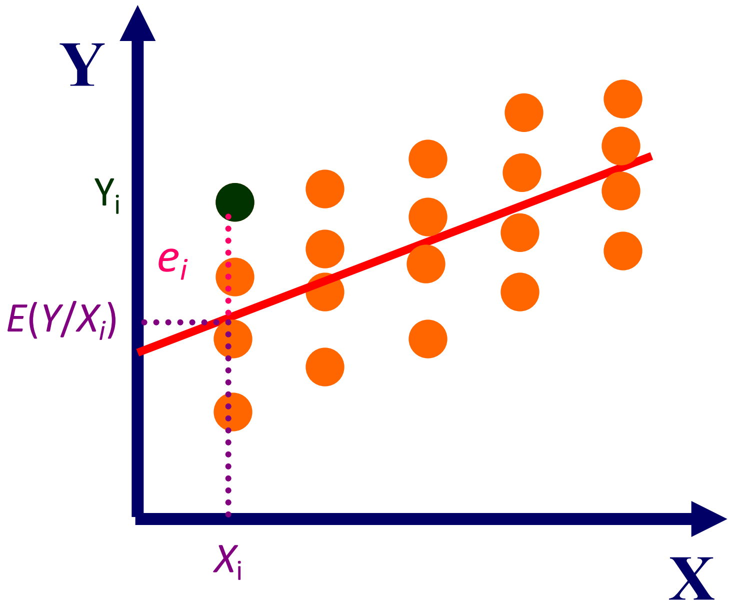
- Regressão Linear – Definição e Interpretação
- Método de Mínimos Quadrados Ordinários
- Análise de Variabilidade
- Teste de Hipóteses para os Coeficientes

Bibliografia

Maia, A. G. 2014. Econometria: conceitos e aplicações. Insittuto de Economica. Caps. 1 a 8.

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 regressor

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:

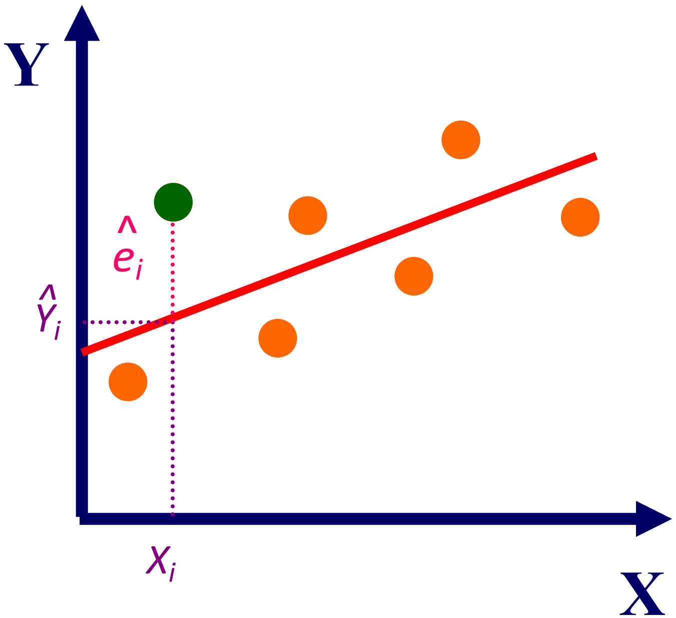
Y_i é o valor observado em Y para a i-ésima observação da amostra

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

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

Função de Regressão Amostral

A relação entre *Y* e *X* estimada na amostra 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):

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

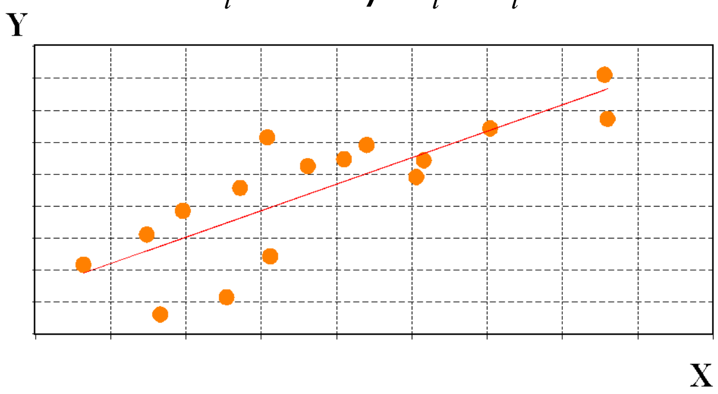
$$EQT = (Y_1 - \hat{Y}_1)^2 + (Y_2 - \hat{Y}_2)^2 + \ldots + (Y_n - \hat{Y}_n)^2$$

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

Interpretação dos Coeficientes

Regressão Linear Simples:

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



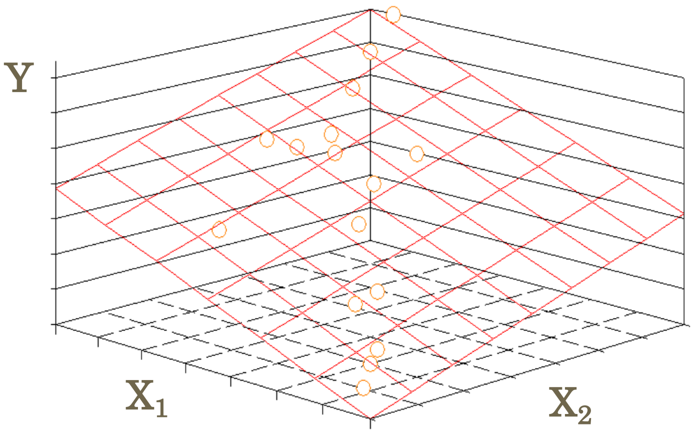
Temos que:

$$E[Y / X = 0] = \alpha$$ 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*.

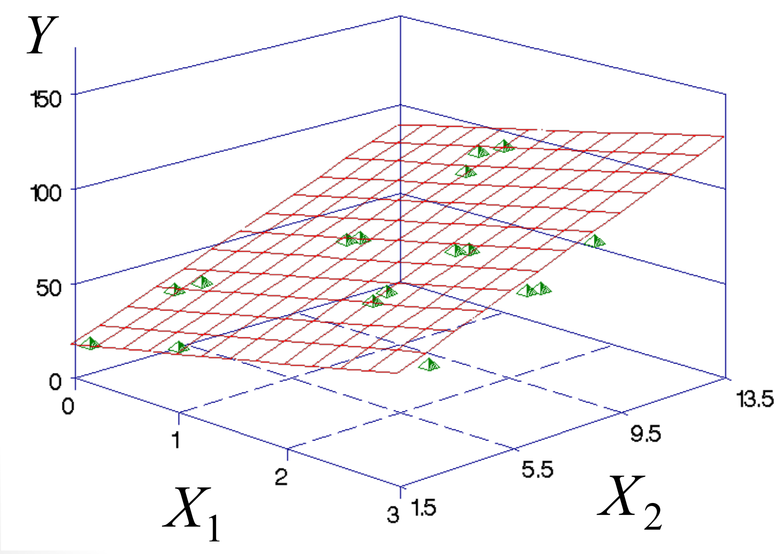
$$\frac{\partial Y}{\partial X_2} = \beta_2$$ Variação marginal esperada em *Y* para cada variação unitária em X_2, *mantendo* X_1 *constante*.

Regressão Múltipla - Definição

Seja a função de regressão: $Y_i = \alpha + \beta_1 X_{1_i} + \beta_1 X_{2_i} + \ldots + \beta_k X_{k_i} + e_i$

Pressupomos que a variável dependente *Y* será determinada por *k* variáveis independentes *X* mais um erro aleatória *e*.

O modelo de regressão estabelece a relação esperada na população (**FRP - função de regressão populacional**). Como usualmente estimamos a relação com base em dados da amostra, trabalhos com a **função de regressão amostral (FRA)**.



$Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + e_i$ **(FRP)**

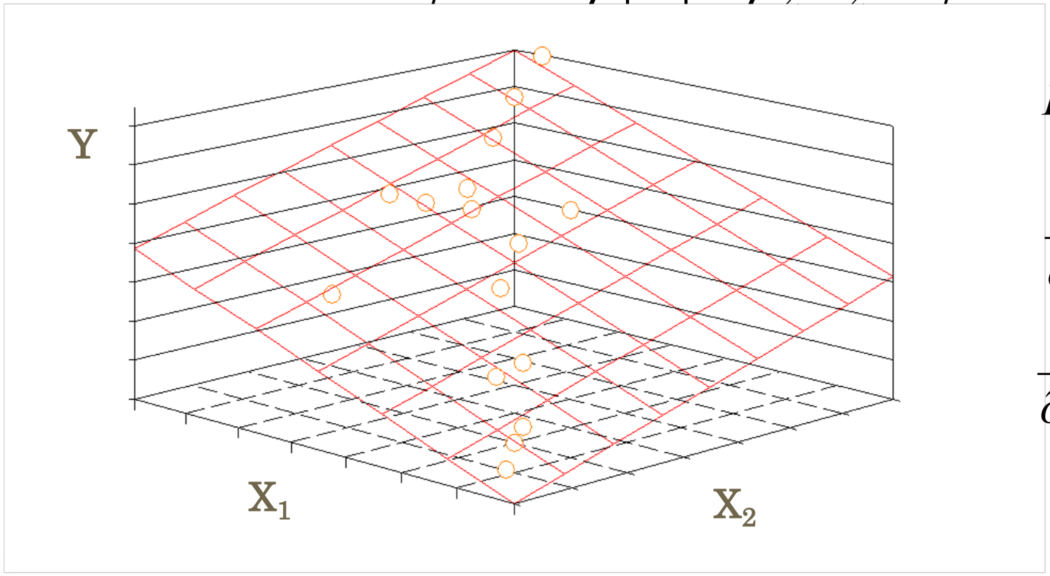
$Y_i = \hat{\alpha} + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \hat{e}_i$ **(FRA)**

Os parâmetros α e β estabelecem a relação esperada na população e os estimadores $\hat{\alpha}$ e $\hat{\beta}$ estimam valores com base em observações de uma amostra. Analogamente, *e* representa os **erros** na população e *ê* os **resíduos** estimados com base nos valores da amostra.

Regressão Múltipla - Interpretação

- Em um modelo de regressão linear, os coeficientes angulares β's captam o **efeito parcial** de uma variável independente sobre a variável dependente. Em outras palavras, qual seria a variação marginal em *Y* para uma variação unitária (pressupostamente marginal) em X_j, mantendo-se constante as demais variáveis independentes;

Seja a função: $$Y_i = \alpha + \beta_1 X_1 + \beta_2 X_2 + e_i$$



$$E[Y / X_1 = 0, X_2 = 0] = \alpha$$

$$\frac{\partial Y}{\partial X_1} = \beta_1$$

$$\frac{\partial Y}{\partial X_2} = \beta_2$$

MQO – Notação Matricial

- Obtém os estimadores da função de regressão de tal forma que os erros sejam mínimos ;

Seja a função: $Y_i = \alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + e_i$

E a equivalente matricial: $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$

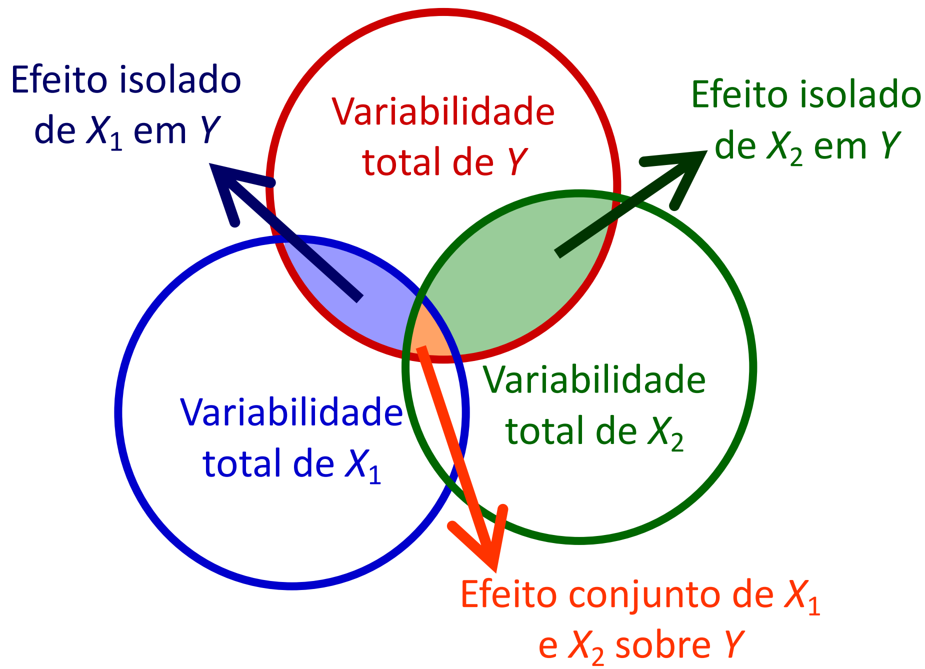
Que representa o sistema:

$$\underbrace{\begin{pmatrix} Y_1 \\ Y_2 \\ \ldots \\ Y_n \end{pmatrix}}_{\mathbf{y}_{n\times 1}} = \underbrace{\begin{pmatrix} 1 & X_{1_1} & X_{2_1} & \ldots & X_{k_1} \\ 1 & X_{1_2} & X_{2_2} & \ldots & X_{k_2} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 1 & X_{1_n} & X_{2_n} & \ldots & X_{k_n} \end{pmatrix}}_{\mathbf{X}_{n\times p}} \underbrace{\begin{pmatrix} \alpha \\ \beta_1 \\ \ldots \\ \beta_k \end{pmatrix}}_{\boldsymbol{\beta}_{p\times 1}} + \underbrace{\begin{pmatrix} e_1 \\ e_2 \\ \ldots \\ e_n \end{pmatrix}}_{\mathbf{e}_{n\times 1}}$$

Minimizando, chegaremos a: $\dfrac{\partial EQT}{\partial \hat{\boldsymbol{\beta}}} = 0 \;\Rightarrow\; \hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}(\mathbf{X}^T\mathbf{y})$

Análise de Variabilidade

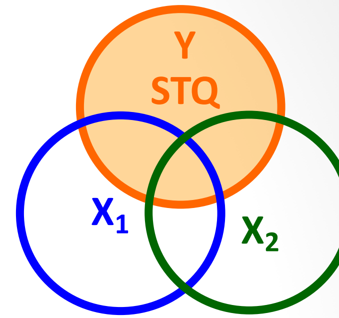
- A variabilidade total de *Y* representa a diversidade de valores que *Y* pode assumir;
- Uma parcela da variabilidade de *Y* pode ser explicada isoladamente pela variável independente X_1, outra explicada isoladamente por X_2 e outra explicada conjuntamente por X_1 e X_2;
- A variabilidade não explicada por *X* será refletida nos erros do modelo de regressão;



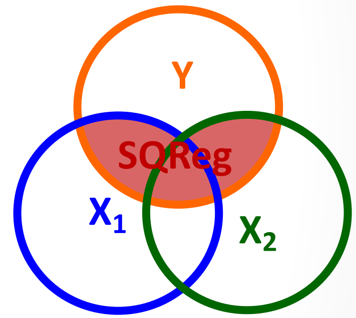
Soma dos Quadrados

- Permitem estimar a qualidade do ajuste;
- Bons modelos implicam variabilidade relativamente baixa dos resíduos (*SQRes*) e variabilidade relativamente alta do ajuste de regressão (*SQReg*);

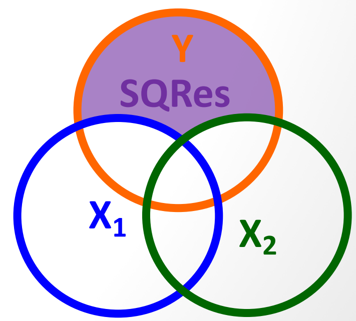
$$STQ = \sum_{i=1}^{n} (Y_i - \bar{Y})^2 = \mathbf{y}^T \mathbf{y} - n\bar{Y}^2$$

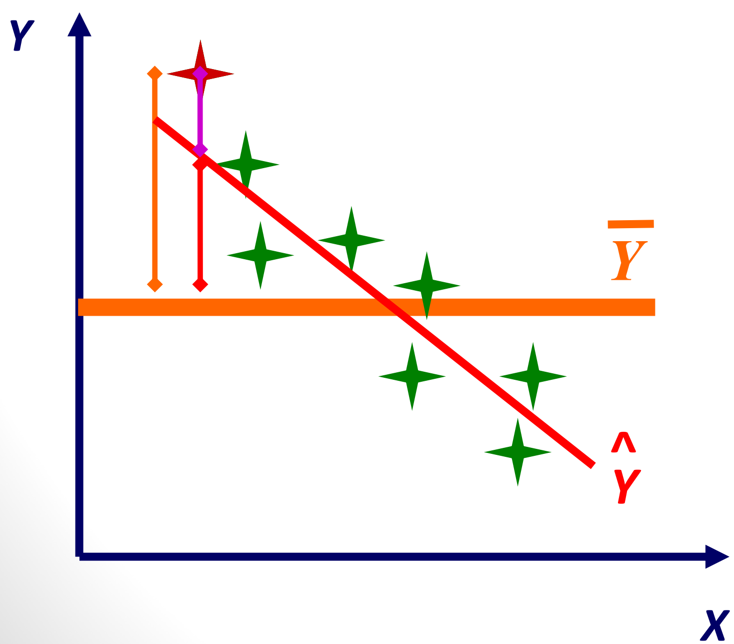


$$SQReg = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2 = \hat{\boldsymbol{\beta}}^T \mathbf{X}^T \mathbf{y} - n\bar{Y}^2$$



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

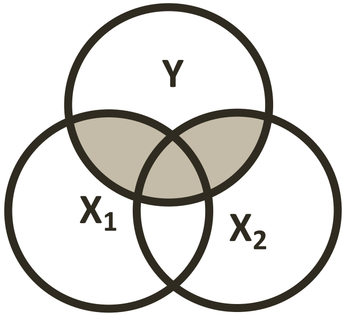




Coeficiente de Determinação

- Estima a proporção da variabilidade da variável dependente *Y* que é explicada pelo conjunto das *k* variáveis independentes do modelo de regressão *X*.

$$R^2 = \frac{SQ\,\mathrm{Re}\,g}{STQ} = 1 - \frac{SQ\,\mathrm{Re}\,s}{STQ}$$



Escala de R^2:

0
Independência linear

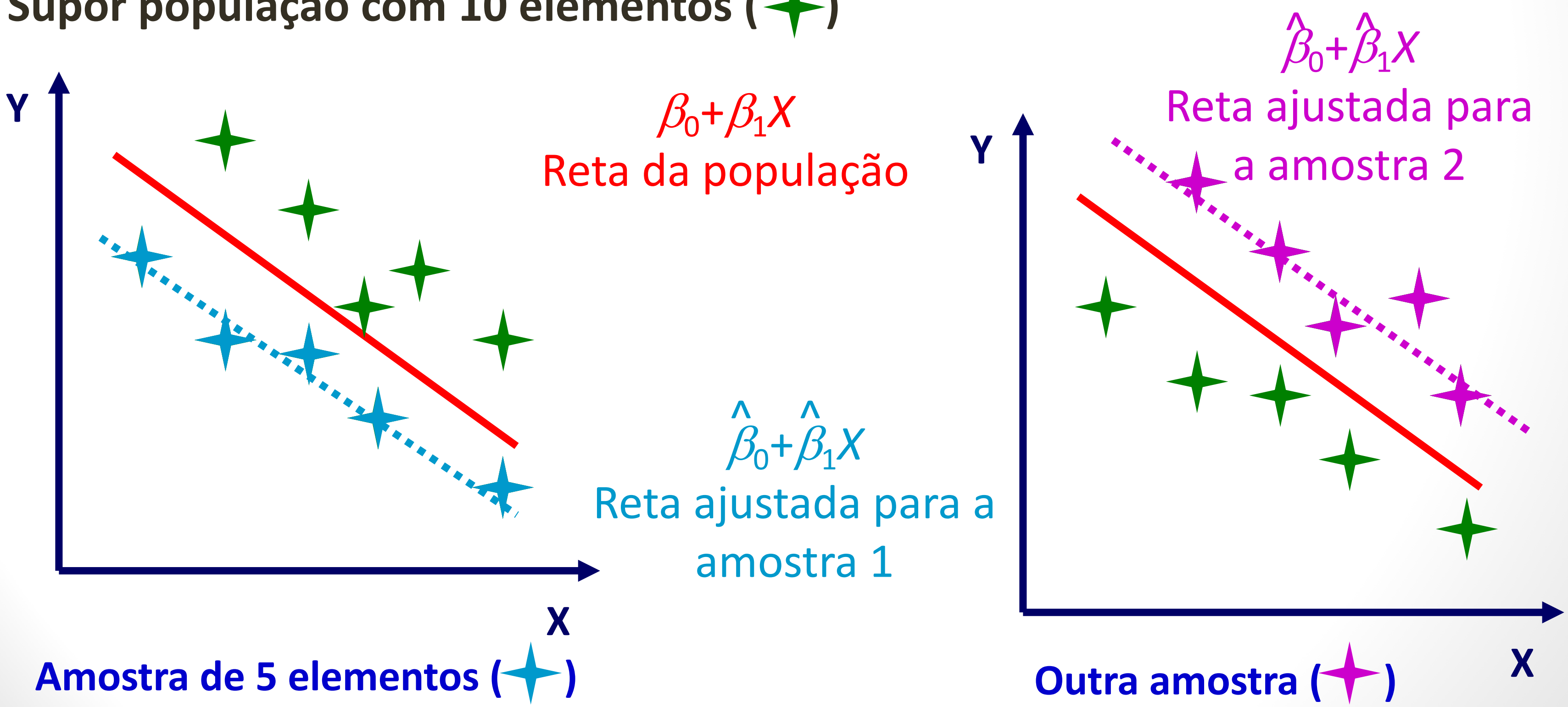
A relevância do R^2 depende muito do tipo de variável dependente (*Y*) sendo analisada

1
Relação linear exata

Distribuição Amostral

- Os valores estimados em uma amostra podem diferir daqueles existentes na população;
- Ademais, os valores irão variar em função da combinação de observações selecionadas na amostra;

Supor população com 10 elementos ( )



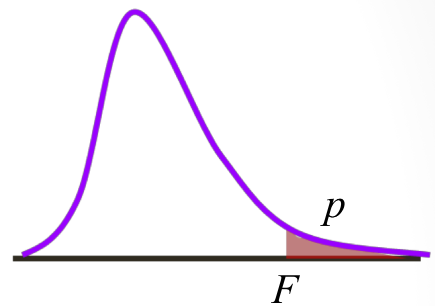
Teste *F*

- Estima o nível de significância do ajuste, ou seja, qual a probabilidade de erro (*p*) se afirmarmos que o modelo contribui para explicar a variabilidade da variável dependente (rejeitar H_0).

Dado o modelo: $Y = \alpha + \beta_1 X_1 + \ldots + \beta_k X_k + e$

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

$$F = \frac{SQ\,\mathrm{Reg}/k}{SQ\,\mathrm{Res}/[n-(k+1)]}$$



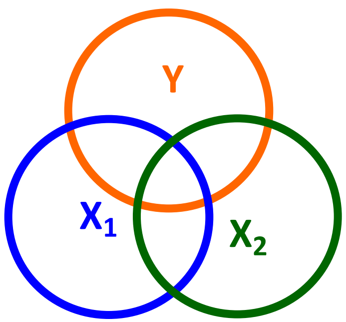
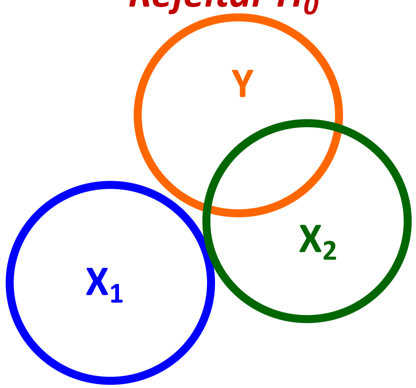
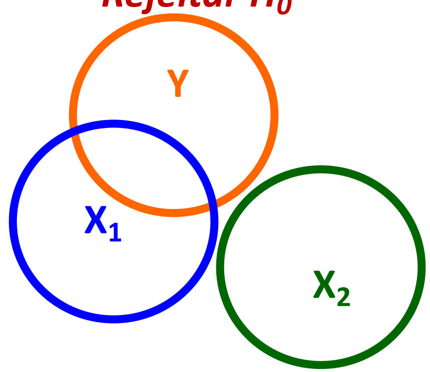
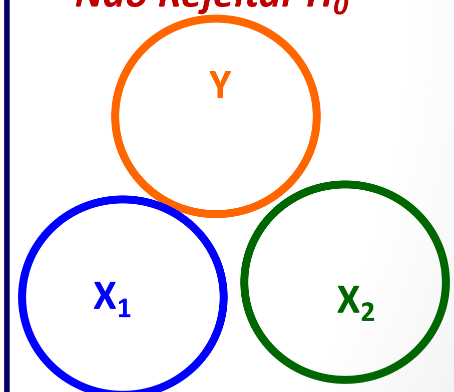
Rejeitar H_0	**Rejeitar H_0**	**Rejeitar H_0**	**Não Rejeitar H_0**
			
$\beta_1 \neq 0$ $\beta_2 \neq 0$	$\beta_1 = 0$ $\beta_2 \neq 0$	$\beta_1 \neq 0$ $\beta_2 = 0$	$\beta_1 = 0$ $\beta_2 = 0$
X_1 e X_2 contribuem para explicar *Y*. H_0 deveria ser rejeitado	Apenas X_2 contribui para explicar *Y*. 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

Tabela Anova - Definição

- Resume os resultados da Análise de Variância do modelo.
- Valores de *p* pequenos (usualmente menores que 5%) indicam que o modelo contribui significativamente para explicar a variabilidade da variável dependente ($R^2 > 0$);

Fonte	**gl**	**SQ**	**QM**	***F***	***p***
Regressão	k	$\hat{\boldsymbol{\beta}}^T \mathbf{X}^T \mathbf{y} - n\overline{Y}^2$	$\frac{\text{SQReg}}{k}$	$\frac{\text{QMReg}}{\text{QMRes}}$	valor *p*
Resíduos	$n-(k+1)$	$\mathbf{y}^T \mathbf{y} - \hat{\boldsymbol{\beta}}^T \mathbf{X}^T \mathbf{y}$	$\frac{\text{SQRes}}{n-(k+1)}$		
Total	$n-1$	$\mathbf{y}^T \mathbf{y} - n\overline{Y}^2$			

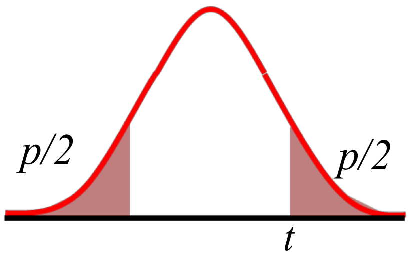
Teste *t*

- Estima a significância de cada coeficiente do modelo, ou seja, qual a probabilidade de erro (*p*) se afirmarmos que a *j*-ésima variável independente contribui isoladamente para explicar a variabilidade da variável dependente (rejeitar H_0).

Dado o modelo: $Y = \alpha + \beta_1 X_1 + \ldots + \beta_k X_k + e$

E as hipóteses: $\begin{cases} H_0: \ \beta_j = 0 \\ H_1: \ \beta_j \neq 0 \end{cases}$

$t = \hat{\beta}_j / S_{\hat{\beta}_j}$



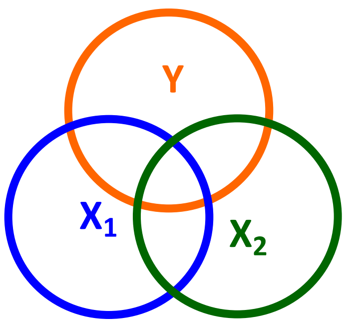
Onde:

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

e:

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

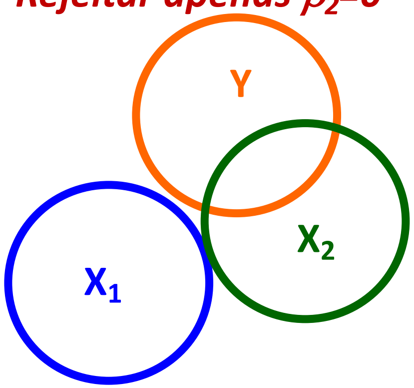
Rejeitar β_1=0 e β_2=0



$\beta_1 \neq 0$ $\beta_2 \neq 0$

X_1 e X_2 contribuem para explicar *Y*. Os dois testes t deveriam ser rejeitados

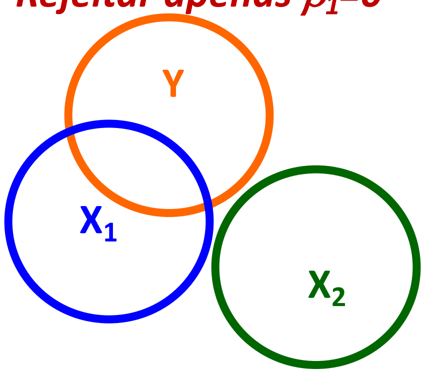
Rejeitar apenas β_2=0



$\beta_1 = 0$ $\beta_2 \neq 0$

Apenas X_2 contribui para explicar *Y*. $H_0 : \beta_2 = 0$ deveria ser rejeitado

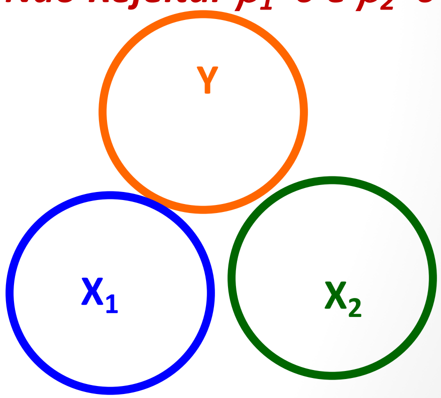
Rejeitar apenas β_1=0



$\beta_1 \neq 0$ $\beta_2 = 0$

Apenas X_1 contribui para explicar *Y*. $H_0 : \beta_1 = 0$ deveria ser rejeitado

Não Rejeitar β_1=0 e β_2=0



$\beta_1 = 0$ $\beta_2 = 0$

Nenhuma variável contribui para explicar *Y*. Nenhum dos testes *t* deveria ser rejeitado.

Exercícios

1) O arquivo Data_TravelCosts.xls contém informações sobre o custos médios de deslocamentos de municípios a um parque nacional (MAIA, A. G. , ROMEIRO, A. . Validade e confiabilidade do método de custo de viagem: um estudo aplicado ao Parque Nacional da Serra Geral. Revista de Economia Aplicada, v. 12, p. 103-123, 2008):
   a) Analise a relação entre o custo de viagem (variável independente) e a taxa de visitação (variável dependente);
   b) Incorpore variáveis de controle;
   c) Os resultados são consistentes com a teórica microeconômica?