Regressão Múltipla

HO 231 – Econometria

Profa. Rosangela Ballini

Instituto de Economia - UNICAMP

Ementa

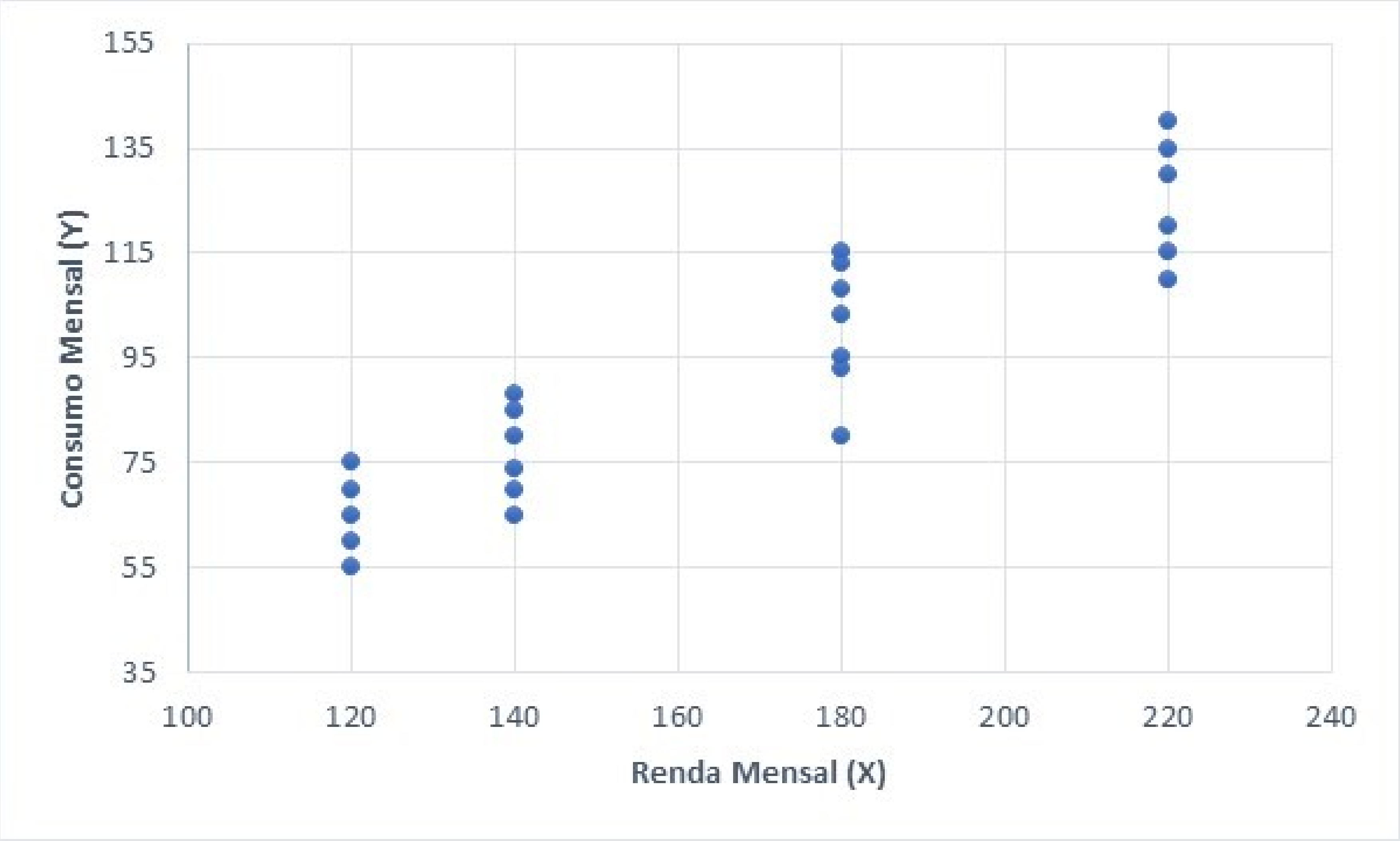
- Regressão Linear Múltipla – MQO
- Escalas de Medidas
- Linearidade dos Coeficientes
- Viés de Omissão
- Análise de Variabilidade
- Inferência para os Coeficientes

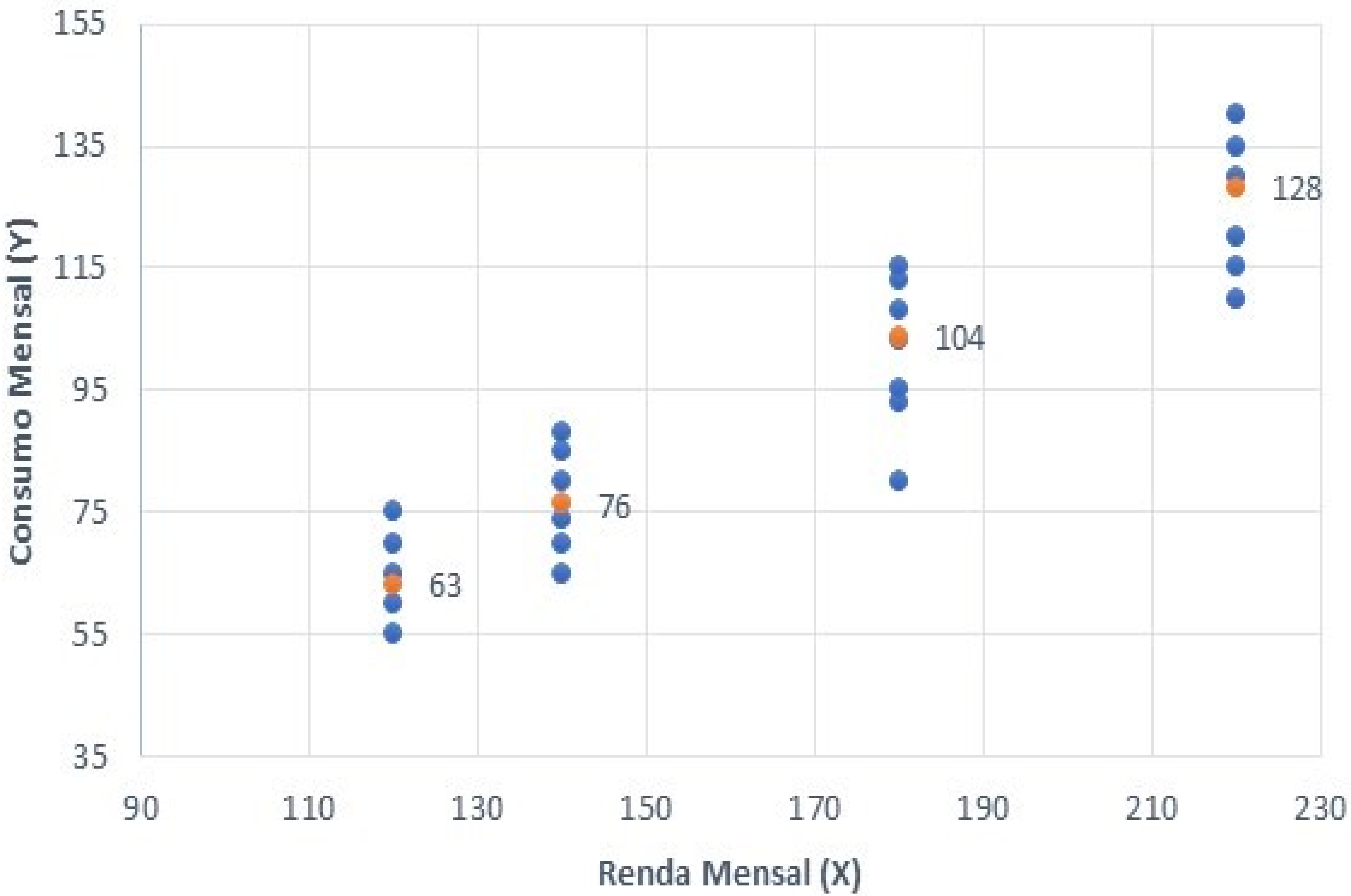
Bibliografia

Wooldridge, J. M. 2001. Introductory Econometric. Caps. 1-4.

Associação Linear

Gráfico de Dispersão




155
135
115
95
75
55
35
90
110
130
150
170
190
210
230
Consumo Mensal (Y)
Renda Mensal (X)
63
76
104
128

Modelo de Regressão Linear

Função Linear de Regressão Populacional (LRP)

Sejam os dados de uma população relacionados por:

$$\boldsymbol{Y_i = \beta_0 + \beta_1 X_i + u_i, \qquad i = 1, \dots, N}$$

sendo

$\boldsymbol{Y}$: a variável dependente

$\boldsymbol{X}$: variável independente

$\boldsymbol{\beta_0}$ e $\boldsymbol{\beta_1}$: parâmetros

$\boldsymbol{u}$: termo de erro

$\boldsymbol{N}$: tamanho da população

PRESSUPOSIÇÕES DO MODELO

1) Lineariedade nos parâmetros;

2) Amostragem aleatória;

3) *Valor médio do erro é igual a zero:*

$$E(u_i|X) = 0$$

4) *Variância do erro é constante:*

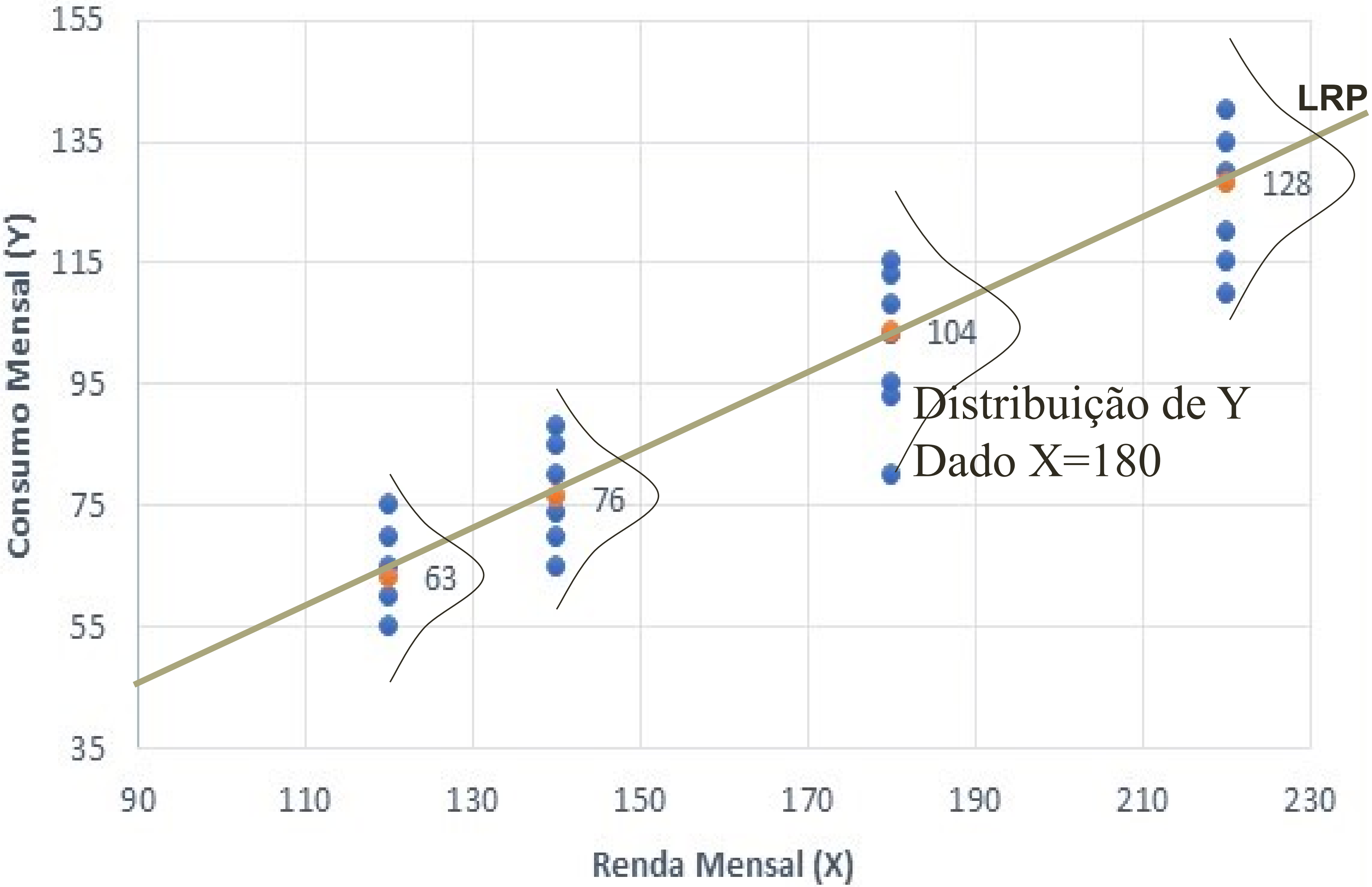
$$Var(u_i|X) = E[u_i - E(u_i)]^2 = \sigma^2$$

5) *Os erros não são correlacionados:*

$$corr(u_i, u_j) = 0, \text{para } i \neq j$$

6) *A distribuição dos erros em torno do valor médio é Normal:*

$$u_i \sim N(0, \sigma^2)$$



LRP
155
135
115
95
75
55
35
63
76
104
128
Distribuição de Y
Dado X=180
90
110
130
150
170
190
210
230
Consumo Mensal (Y)
Renda Mensal (X)

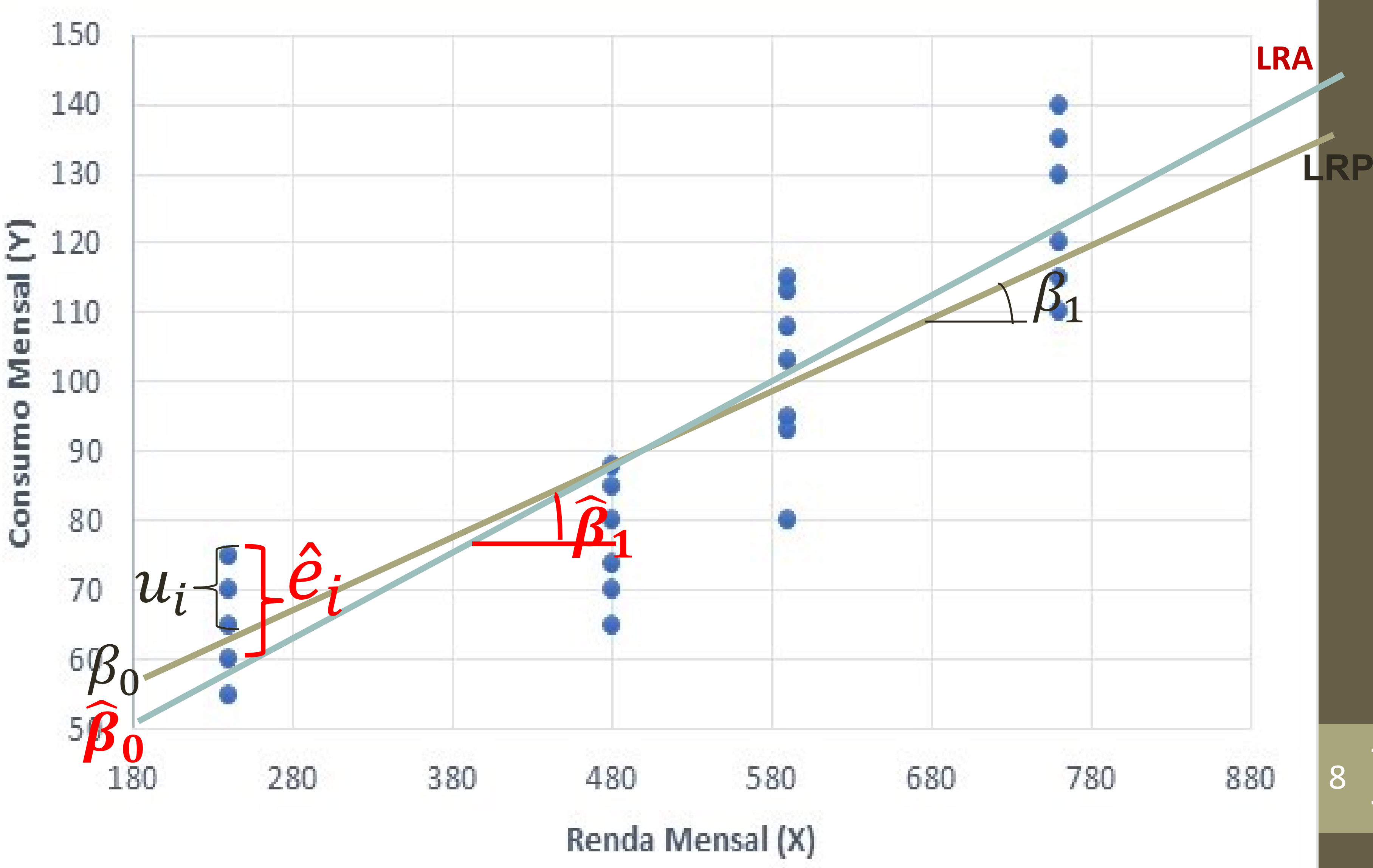
Estimativas dos Parâmetros

Como o modelo de regressão populacional

$$\boldsymbol{Y_i = \beta_0 + \beta_1 X_i + u_i, \qquad i = 1, \dots, N}$$

não é diretamente observável, nós devemos estimá-lo a partir do modelo de regressão amostral:

$$\boldsymbol{\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_i + \hat{e}_i, \qquad i = 1, \dots, n}$$



LRA
LRP
Consumo Mensal (Y)
150
140
130
120
110
100
90
80
70
60
50
180
280
380
480
580
680
780
880
Renda Mensal (X)
β_1
$\hat{\beta}_1$
u_i
$\hat{e}_i$
β_0
$\hat{\beta}_0$

Função de Regressão Amostral

Função de regressão amostral: $Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + \hat{e}_i$

***Y* previsto pelo ajuste:** $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$

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

Função de Erro Quadrático Total (EQT):

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

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

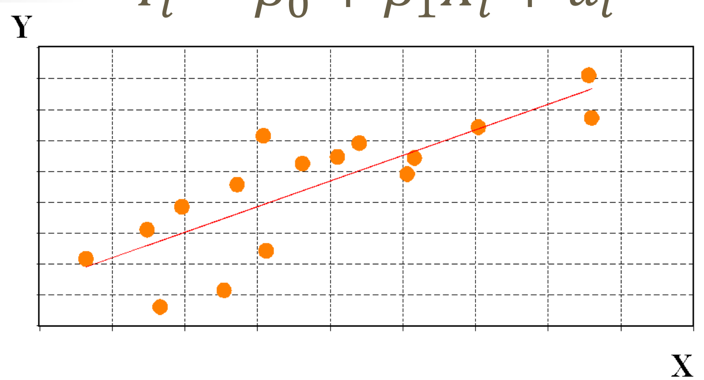
$$EQT = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

$$EQT = \sum_{i=1}^{n} \left(Y_i - \left(\hat{\beta}_0 - \hat{\beta}_1 X_i\right)\right)^2$$

Mínimos Quadrados Ordinários

Regressão Linear Simples:

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$



$$EQT = \sum_{i=1}^{n} \left(Y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 X_i\right)\right)^2$$

Minimizando EQT:

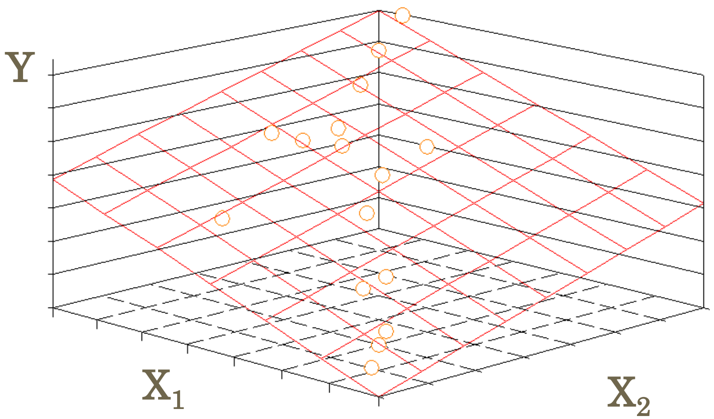
$$\frac{\partial EQT}{\partial \hat{\beta}_0} = 0 \Rightarrow \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

$$\frac{\partial EQT}{\partial \hat{\beta}_1} = 0 \Rightarrow \hat{\beta}_1 = \frac{\sum x_i y_i}{\sum x_i^2}$$

Em que $x_i = (X_i - \bar{X})$ e $y_i = (Y_i - \bar{Y})$

Regressão Linear Múltipla:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$$



$$EQT = \sum_{i=1}^{n} \left(Y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i}\right)\right)^2$$

Minimizando EQT:

$$\frac{\partial EQT}{\partial \hat{\beta}_0} = 0 \Rightarrow \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}_1 - \hat{\beta}_2 \bar{X}_2$$

$$\frac{\partial EQT}{\partial \hat{\beta}_1} = 0 \Rightarrow \hat{\beta}_1 = \frac{(\sum y_i x_{1i})(\sum x_{2i}^2) - (\sum y_i x_{2i})(\sum x_{1i} x_{2i})}{(\sum x_{1i}^2)(\sum x_{2i}^2) - (\sum x_{1i} x_{2i})^2}$$

$$\frac{\partial EQT}{\partial \hat{\beta}_2} = 0 \Rightarrow \hat{\beta}_2 = \frac{(\sum y_i x_{2i})(\sum x_{1i}^2) - (\sum y_i x_{1i})(\sum x_{1i} x_{2i})}{(\sum x_{1i}^2)(\sum x_{2i}^2) - (\sum x_{1i} x_{2i})^2}$$

Modelo de Regressão Múltipla

Em um modelo de regressão múltipla, uma variável dependente Y_i está relacionada com duas ou mais variáveis independentes X_{ji} :

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + u_i$$

sendo:

$\boldsymbol{\beta_0}$ é o valor esperado de Y quando todos as variáveis independentes forem nulas;

$\boldsymbol{\beta_1}$ é a variação esperada em Y dado um incremento em X_1, mantendo-se constante todas as demais variáveis independentes;

...

$\boldsymbol{\beta_k}$ é a variação esperada em Y dado um incremento em X_k, mantendo-se constante todas as demais variáveis independentes;

$\mathbf{u_i}$ é o erro não explicado pelo modelo;

Utilizando a forma matricial, temos:

$$\mathbf{y}=\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix};\quad \mathbf{X}=\begin{bmatrix} 1 & X_{11} & X_{21} & \cdots & X_{k1} \\ 1 & X_{12} & X_{22} & \cdots & X_{k2} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 1 & X_{1n} & X_{2n} & \cdots & X_{kn} \end{bmatrix};\ \boldsymbol{\beta}=\begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix};\mathbf{u}=\begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix}$$

Ou,

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

Estimação dos Parâmetros do Modelo de Regressão Múltipla

Princípio dos mínimos quadrados: minimizar a soma dos quadrados das diferenças entre os valores observados Y_i e os valores esperados $E(Y_i)$.

Os valores esperados são estimados pelo modelo:

$$\hat{Y}_j = b_0 + b_1 X_{1j} + \cdots + b_k X_{kj}$$

Ou, na forma matricial:

$$\hat{\mathbf{y}} = \mathbf{Xb}$$

Temos que: $$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - \mathbf{X}\mathbf{b}$$

A soma dos quadrados dos desvio é:

$$Z = \sum e_j^2 = \mathbf{e}'\mathbf{e} = (\mathbf{y}' - \mathbf{b}'\mathbf{X}')(\mathbf{y} - \mathbf{X}\mathbf{b}) = \mathbf{y}'\mathbf{y} - 2\mathbf{b}'\mathbf{X}'\mathbf{y} + \mathbf{b}'\mathbf{X}'\mathbf{X}\mathbf{b}$$

Diferenciando a função Z em relação aos parâmetros e igualando a zero, temos:

$$\mathbf{X}'\mathbf{X}\mathbf{b} - \mathbf{X}'\mathbf{y} = \mathbf{0} \Leftrightarrow \mathbf{X}'\mathbf{X}\mathbf{b} = \mathbf{X}'\mathbf{y}$$

Se **X’X** é não singular (det(**X’X**)$\neq$0) temos:

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

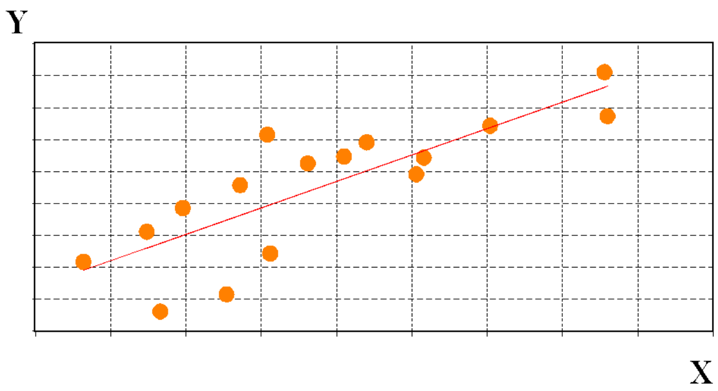
Pressupostos do Modelo

1. Linearidade do Modelo de Regressão: $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$
2. Amostragem Aleatória.
3. Ausência de Colineariedade Perfeita;
4. Erros possuem média zero:
$$E(u_j) = 0 \Leftrightarrow E(Y_j) = \beta_0 + \beta_1 X_{1j} + \cdots + \beta_k X_{kj}$$
5. Erros são homocedásticos:
$$Var(u_j) = \sigma^2 \Leftrightarrow Var(Y_j) = \sigma^2$$
6. Erros são não correlacionados:
$$\operatorname{cov}(u_j, u_h) = \mathbf{0} \Leftrightarrow \operatorname{cov}(Y_j, Y_h) = \mathbf{0},\ j \neq h$$
7. Erros são normalmente distribuídos:
$$u_j : N(0, \sigma^2) \Leftrightarrow Y_j : N(\mathbf{Xb}, \sigma^2)$$

Interpretação dos Coeficientes

Regressão Linear Simples:

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$



Tem-se que:

$$E(Y|X=0) = \beta_0$$

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:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$$

Tem-se que:

$$E(Y|X_1 = 0, X_2 = 0) = \beta_0$$

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*.

- Coeficientes β's captam o **efeito parcial** de uma variável independente sobre a variável dependente

Conceito de Efeito Parcial

Seja o modelo de RLM: $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$

- Pode-se demonstrar que o estimador de MQO para β_1 será:

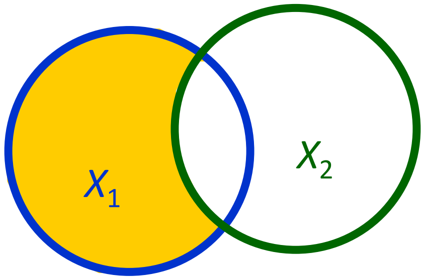
$$\beta_1 = \sum_{i=1}^{n} \hat{r}_{i1} Y_i \Big/ \sum_{i=1}^{n} \hat{r}_{i1}^2$$

sendo $\hat{r}_{i1}$ o resíduo do ajuste de MQO de X_1 em função de X_2;

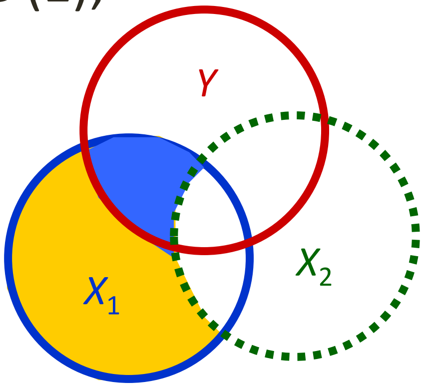
- Processo similar a dois estágios de estimação:

1) X1 em função de X2;

2) Y em função dos resíduos do ajuste (1);



$$X_1 = \hat{\phi}_0 + \hat{\phi}_1 X_2 + \hat{r}_1$$



$$Y = \hat{\alpha}_1 + \hat{\beta}_1 \hat{r}_1 + \hat{e}_1$$

Estimação por MQO – *no R*

- O comando **lm** ajusta uma função de regressão por MQO;
- A especificação deve conter o regressando seguido de todas as variáveis explanatórias;

```
> modelo.reg = lm(co2 ~ pib + setor2, data=dados)
```

Obtém as estimativas de MQO para o modelo das emissões de CO_2 (*co2*) em função do PIB (*pib*) e da participação da setor secundário no PIB (*setor2*)

- As estimativas são apresentadas no R:

```
> coefficients(modelo.reg)
 (Intercept)       pib        setor2
-1.1262519087 0.0003209484 0.1050829300
```

Intercepto negativo: não tem interpretação econômica (não existe país com PIB nulo!).
Estimativas dos parâmetros:

- mantendo-se constante a participação do setor secundário, cada aumento de 1 US$ no PIB per capita implicará um acréscimo médio de 0,0003 ton nas emissões per capitas de CO_2 (ou seja, 0,3 kg).
- Analogamente, cada variação percentual na participação do setor secundário implicará um acréscimo médio de 0,1 ton per capita de CO_2, *ceteris paribus*.

Propriedades dos Estimadores de MQO

1. A reta de regressão passa pelas médias de X e Y:

$$\bar{Y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{X} + e_i$$

2. A média dos resíduos é igual a zero:

$$\sum_{i=1}^{n} e_i = 0$$

3. A soma do produto dos resíduos pelos valores de X_i é igual a zero:

$$\sum_{i=1}^{n} e_i X_i = 0$$

4. A soma do produto dos resíduos pelos valores estimados de Y_i é igual a zero:

$$\sum_{i=1}^{n} e_i \hat{Y}_i = 0$$

Escalas de Medidas

- Mudanças nas escalas de medidas irão modificar os coeficientes;

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

$$Y_i = \beta_0 + \beta_1 X_i + u_i \Rightarrow \hat{\beta}_1 = \frac{\sum x_i y_i}{\sum x_i^2}$$

- Se, por exemplo, multiplicarmos uma das variáveis pela constante *c*:

$$Y_i = \beta_0 + \beta_1 (cX_i) + u_i \Rightarrow \hat{\beta}_0 = \bar{Y} - \frac{1}{c}\hat{\beta}_1^*(c\bar{X}) = \bar{Y} - \hat{\beta}_1^* \bar{X}$$

$$\hat{\beta}_1^* = \frac{\sum c x_i y_i}{\sum (c x_i)^2} = \frac{1}{c}\frac{\sum x_i y_i}{\sum x_i^2} = \frac{1}{c}\beta_1$$

Neste exemplo, o intercepto mantém-se o mesmo e o coeficiente angular é dividido por *c*.

Escalas de Medidas - Exemplo

- Se o PIB for medido em 1000 US$ (*c*=1/1000):

```
> pib1000=(dados$pib)/1000
```

Ajustar novo modelo, utilizando a variável independente *pib*1000, que informa o PIB do país em 1000 US$.

- As novas estimativas serão:

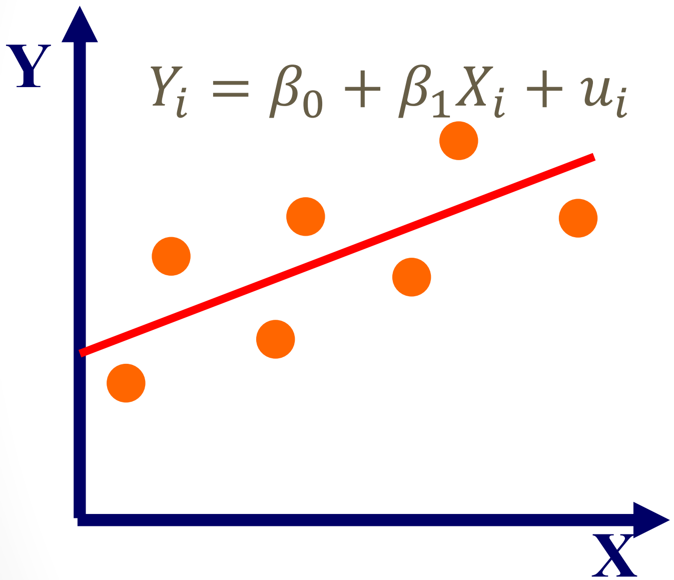
```
> coefficients(modelo2.reg)

(Intercept) pib1000 setor2
-1.1262519 0.3209484 0.1050829
```

Intercepto é o mesmo, assim como o ceficiente angular associado ao regressor *setor*2.

Para cada acréscimo de 1000 dólares no PIB per capita, há um acréscimo esperado de 320,9 kg nas emissões per capita de CO^2.

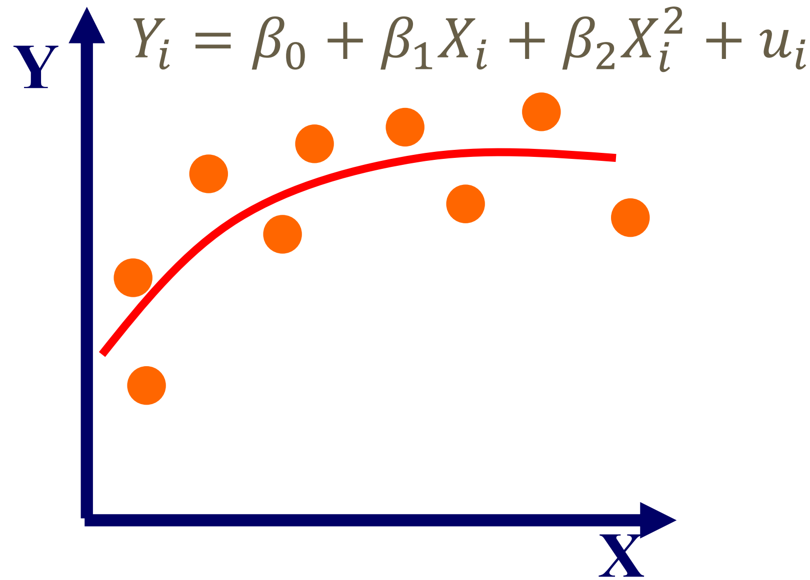
Linearidade nos Coeficientes

- Para que os estimadores de MQO sejam não tendenciosos, as relações devem ser lineares nos coeficientes;



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

Modelo é **linear nos parâmetros:** os expoentes de β_0 e β_1 são iguais a 1.



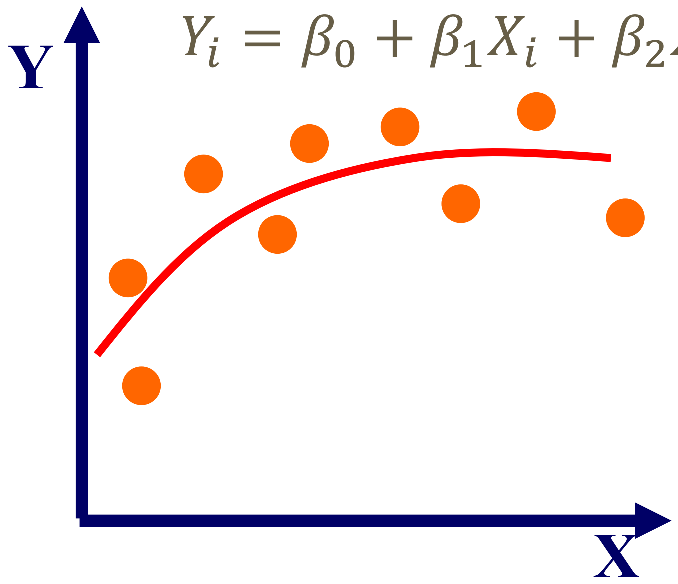
Modelo **não é linear nas variáveis:** possui um expoente de X igual a 2.

Modelo é **linear nos parâmetros:** os expoentes de β_0, β_1 e β_2 são iguais a 1.

Interpretação dos Coeficientes

- O conceito de *ceteris paribus* (tudo mais constante) nem sempre é válido em modelos não lineares nas variáveis;

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + u_i$$



- No modelo quadrático, o efeito marginal de *X* em *Y* dependerá do valor de X:

$$\frac{dY}{dX} = \beta_1 + 2\beta_2 X$$

Modelo Quadrático - Exemplo

- Pressupondo que a relação entre emissões de CO^2 e PIB seja quadrática;

```
> modelo3.reg=lm(co2 ~ pib1000 + I(pib1000^2)+setor2, data=dados)
> coefficients(modelo3.reg)
(Intercept) pib1000 I(pib1000^2) setor2
-1.389918411 0.629310698 -0.008685163 0.088556973
```

- Efeito parcial dependerá do valor anterior do PIB:

$$\frac{dY}{dX} = \beta_1 + 2\beta_2 X = 0{,}629 - 2 \times 0{,}009\, pib1000$$

A interpretação não é mais algo muito trivial.

- Podemos ainda estimar o impacto máximo (ponto de inflexão):

$$\frac{dY}{dX} = \beta_1 + 2\beta_2 X = 0 \quad \Rightarrow \quad X = -\frac{\beta_1}{2\beta_2}$$

Impacto do PIB nas emissões de CO2: cresce até o PIB per capita alcançar 36,2 mil dólares, quando passa a decair.

Esperança Condicional Zero

- Para que os estimadores de MQO sejam não tendenciosos, as experanças condicionais dos erros devem ser iguais a zero:

$$E(e / X_i) = 0$$

- É a mesma coisa que pressupor:

$$E(Y|X_i) = \beta_0 + \beta_1 X_i$$

- É um pressuposto mais forte que *Cov*(*e*,*X*)=0, pois este considera apenas relações lineares entre erros e regressores;

Viés de Omissão - Definição

- Suponha que a real relação entre as variáveis na população seja:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$$

- Mas, erroneamente, ajusta-se o modelo:

$$Y_i = \tilde{\beta}_0 + \tilde{\beta}_1 X_{1i} + u_i$$

- A omissão indevida do regressor X_2 no modelo causará viés na estimativa de $\tilde{\beta}_1$. Pode-se demonstrar que:

$$E(\tilde{\beta}_1) = \beta_1 + \beta_2 \tilde{\delta}_1 \qquad \text{sendo} \quad X_2 = \tilde{\delta}_0 + \tilde{\delta}_1 X_1$$

- O viés em β_1 dependerá de β_2 e do sentido da relação entre X_1 e X_2. De maneira geral:

	Corr (X_1, X_2) > 0	Corr (X_1, X_2) < 0
$\beta_2 > 0$	Viés Positivo	Viés Negativo
$\beta_2 < 0$	Viés Negativo	Viés Positivo

Viés de Omissão - Exemplo

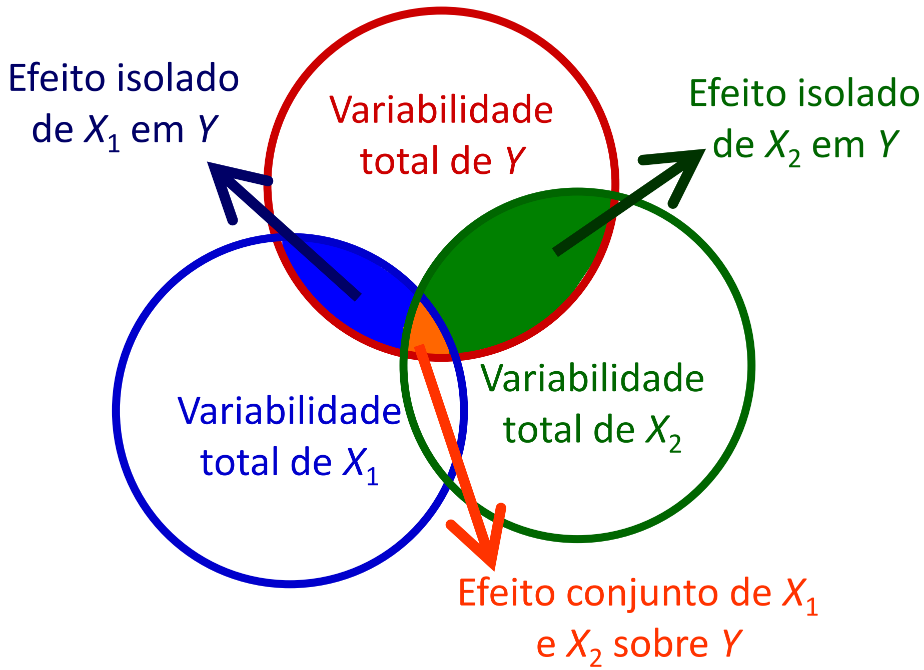
- Viés de omissão pode ser identificado teoricamente, a partir de possíveis variáveis omitidas que estejam relacionadas aos regressores;
- Análise de sensibilidade: verifica em que media a inclusão de regressores no modelo interferem nos coeficientes angulares;

```
#Análise de sensibilidade
>regressao1=lm(co2 ~ pib1000, data=dados)
>coefficients(regressao1)
(Intercept) pib1000
2.0127597 0.3178294

>modelo2.reg=lm(co2 ~ pib1000 + setor2, data=dados)
> coefficients(modelo2.reg)
(Intercept)  pib1000    setor2
-1.1262519 0.3209484 0.1050829
```

Inclusão do regressor *setor*2 no modelo: não modificou expressivamente a estimativa do coeficiente associado ao PIB. Apesar de *setor*2 ter relação com CO^2, sua omissão no modelo não causou um viés expressivo pois este teria baixa relação linear com o PIB.

Análise de Variabilidade

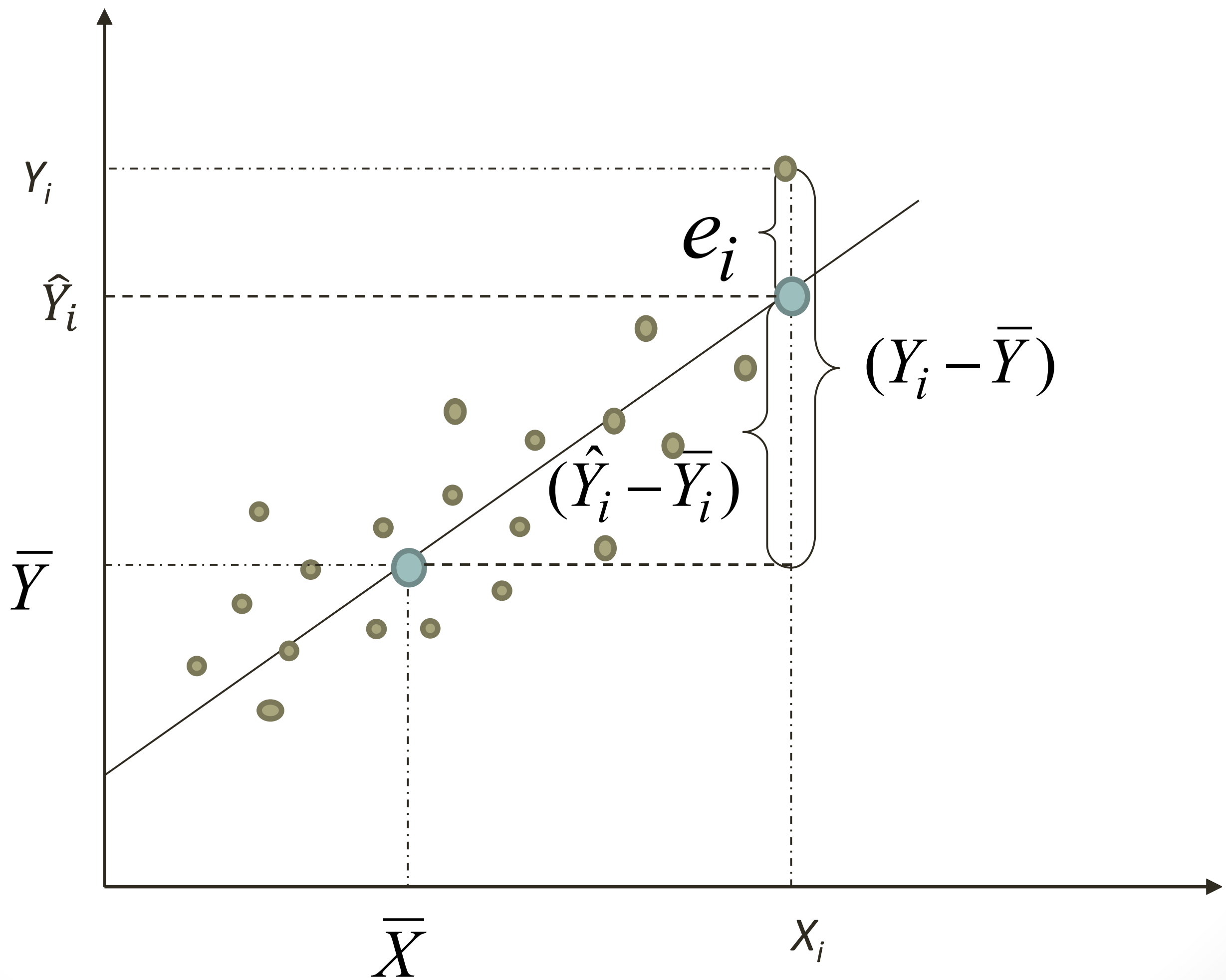
- Variabilidade total de *Y* representa os valores que *Y* pode assumir;
- Parcela da variabilidade de *Y* pode ser explicada isoladamente pela variável independente X_1, outra explicada isoladamente por X_2 e outra explicada conjuntamente por X_1 e X_2;
- Variabilidade não explicada por *X* será refletida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



Sabemos que $$Y_i = \hat{Y}_i + e_i \qquad (1)$$

sendo

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i \text{ e } e_i = Y_i - \hat{Y}_i$$



Especificamente, essas somas são:

$$\text{SQTotal} = \sum (Y_i - \bar{Y})^2 = \sum y_i^2$$

$$\text{SQReg} = \sum (\hat{Y}_i - \bar{Y})^2 = \sum \hat{y}_i^2 = \hat{\beta}_1^2 \sum x_i^2 = \hat{\beta}_1 \sum x_i y_i$$

$$\text{SQRes} = \sum e_i^{\mathbf{2}} = \sum \left(Y_i - \hat{Y}_i\right)^{\mathbf{2}} = \sum y_i^{\mathbf{2}} - \hat{\beta}_{\mathbf{1}} \sum x_i y_i$$

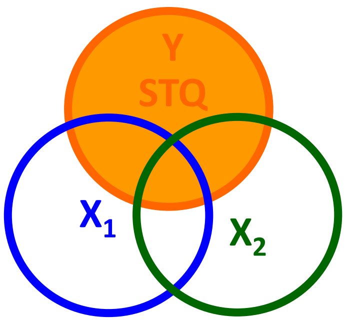
Das equações acima:

SQTotal = SQReg + SQRes

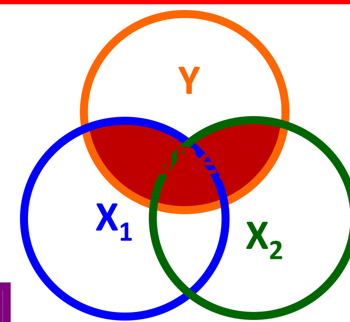
Soma dos Quadrados

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

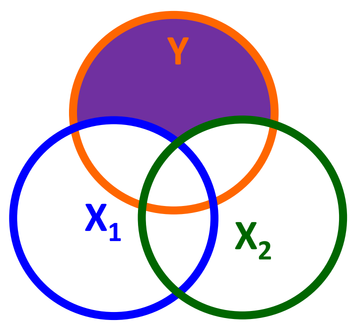
$$STQ = \sum_{i=1}^{n} (Y_i - \overline{Y})^2 = \mathbf{y}^T\mathbf{y} - n\overline{Y}^2$$



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



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



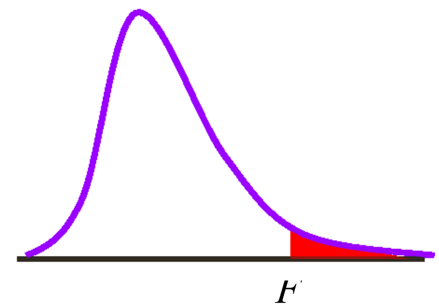
Teste *F*

- Estima a 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).

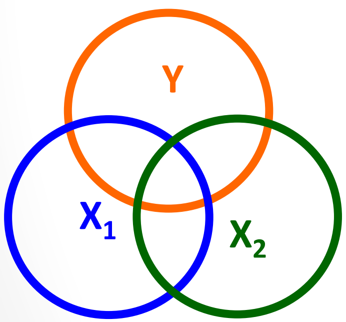
Dado o modelo: $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + u_i$

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

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



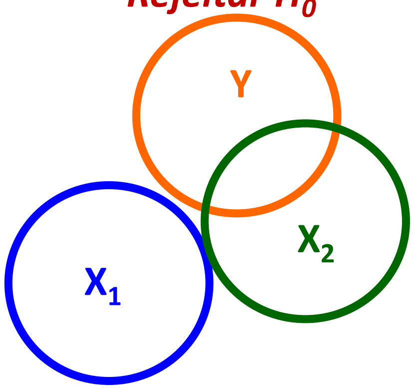
Rejeitar H_0



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

X_1 e X_2 contribuem para explicar *Y*.

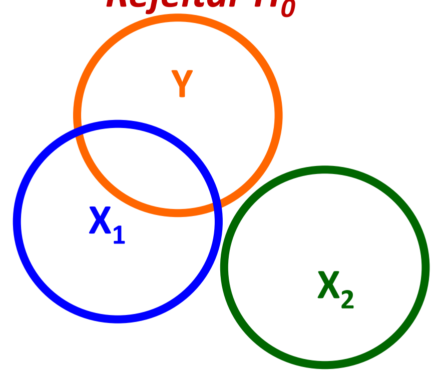
Rejeitar H_0



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

Apenas X_2 contribui para explicar *Y*.

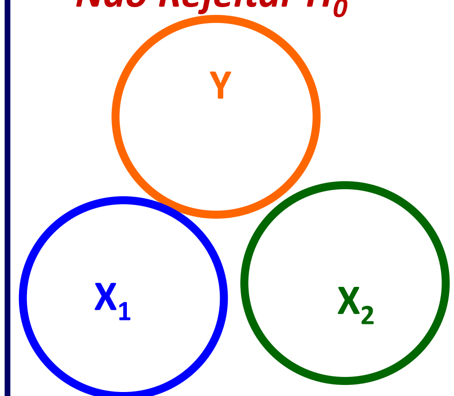
Rejeitar H_0



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

Apenas X_1 contribui para explicar *Y*.

Não Rejeitar H_0



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

Variáveis não contribuem para explicar *Y*.

Tabela Anova - Definição

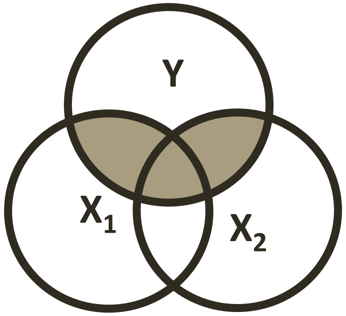
- Resume os resultados da Análise de Variância do modelo.
- Valores de *p* pequenos indicam que o modelo contribui significativamente para explicar a variabilidade da variável dependente ($R^2 > 0$);

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

Coeficiente de Determinação

Coeficiente de determinação R^2 :

$$R^2 = \frac{\text{SQReg}}{\text{SQTotal}} = 1 - \frac{SQ\,\text{Re}\,s}{SQTotal}$$



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

$$0 \leq R^2 \leq 1.$$

Coeficiente de Determinação Corrigido

Considerando os graus de liberdade das *SQRes* e *SQTotal*, o coeficiente de determinação corrigido é definido por:

$$\bar{R}^2 = 1 - \frac{(n-1)}{(n-2)}(1 - R^2)$$

Tabela ANOVA – Exemplo

- Tabela ANOVA apresentada com a execução do comando **anova**;

```
> anova(modelo3.reg)
Analysis of Variance Table

Response: co2
                Df   Sum Sq   Mean Sq    F       value Pr(>F)
pib1000         1    1716.71   1716.71 170.448    < 2.2e-16 ***
I(pib1000^2)    1     368.42    368.42  36.579     9.501e-09 ***
setor2          1     224.31    224.31  22.271     5.022e-06 ***
Residuals       165   1661.84    10.07

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- Comando **summary** apresenta um resumo da regressão

```
summary(modelo3.reg)
...
Residual standard error: 3.174 on 165 degrees of freedom
Multiple R-squared: 0.5815, Adjusted R-squared: 0.5739
F-statistic: 76.43 on 3 and 165 DF, p-value: < 2.2e-16
```

Propriedades Amostrais do Estimador de Mínimos Quadrados

1. Estimadores de mínimos quadrados **b** são variáveis aleatórias;
2. Admitindo que os erros sejam distribuídos normalmente, então Y também será uma variável aleatória distribuída normalmente;
3. Estimadores **b** também terão distribuições normais de probabilidade, pois são funções lineares de *Y*;
4. Se os erros não são distribuídos normalmente, então os estimadores de mínimos quadrados têm distribuição aproximadamente normal em grandes amostras.

Teorema de Gauss-Markov:

Sob os pressupostos do modelo de regressão múltipla:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + u_i$$

os estimadores de mínimos quadrados b_k são os melhores estimadores lineares não tendenciosos e de variância mínima de β_k.

Como os estimadores de mínimos quadrados são funções lineares da variável dependente, e sob o pressuposto 7, temos que os estimadores também são distribuídos normalmente:

$$b_k : N\left(E\left(b_k\right), Var(b_k)\right)$$

Valor Médio dos Estimadores

Temos que:

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\,\mathbf{X}'\mathbf{y} \qquad (1)$$

Substituindo $\mathbf{y} = \mathbf{X\beta} + \mathbf{u}$ em (1):

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\,\mathbf{X}'(\mathbf{X\beta} + \mathbf{u}) = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\,\mathbf{X}'\mathbf{u} \qquad (2)$$

Tomando o valor esperado:

$$E(\mathbf{b}) = \boldsymbol{\beta}$$

Variâncias e Covariâncias dos Estimadores

Por definição, a matriz de variâncias e covariâncias é dada por:

$$\text{Var\_Cov}(b) = E\left[\left(\mathbf{b} - E(\mathbf{b})\right)\right]^2 = E[(\mathbf{b} - \boldsymbol{\beta})(\mathbf{b} - \boldsymbol{\beta})']$$

Do resultado (2):

$$\mathbf{b} - \boldsymbol{\beta} = (\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\,\mathbf{X}'\mathbf{u}$$

Temos:

$$E[(\mathbf{b} - \boldsymbol{\beta})(\mathbf{b} - \boldsymbol{\beta})'] = E\left[(\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\,\mathbf{X}'\mathbf{u}\mathbf{u}'\mathbf{X}'(\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\right]=(\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\boldsymbol{\sigma}^{\mathbf{2}}$$

sendo σ^2 é a variância do erro, sendo sua estimativa dada por:

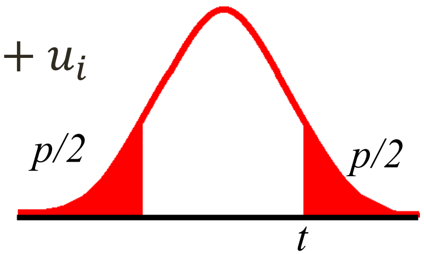
$$S^2 = \hat{\sigma}^2 = \frac{SQRes}{(n-p)} = QMRes = \frac{\mathbf{e}'\mathbf{e}}{(n-p)}$$

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: $Y_i = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki} + u_i$

E as hipóteses: $\begin{cases} H_0: \ \beta_j = 0 \\ H_1: \ \beta_j \neq 0 \end{cases}$ $\xrightarrow{t = \hat{\beta}_j / S_{\hat{\beta}_j}}$



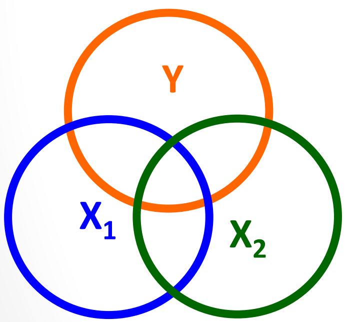
sendo:

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

e:

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

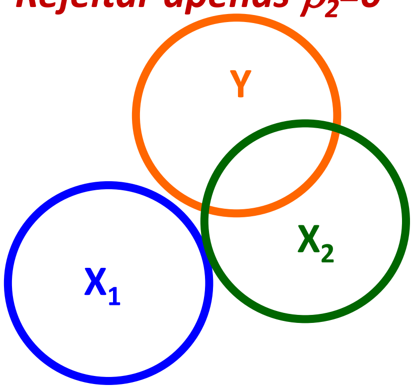
Rejeitar β_1=0 e β_2=0



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

X_1 e X_2 contribuem para explicar Y.

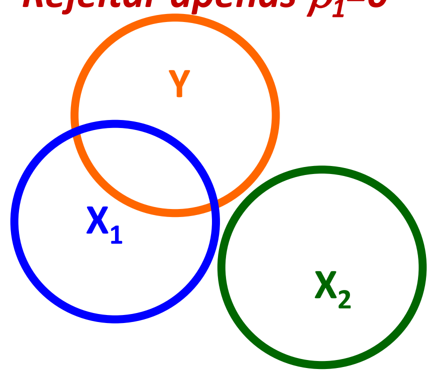
Rejeitar apenas β_2=0



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

Apenas X_2 contribui para explicar Y.

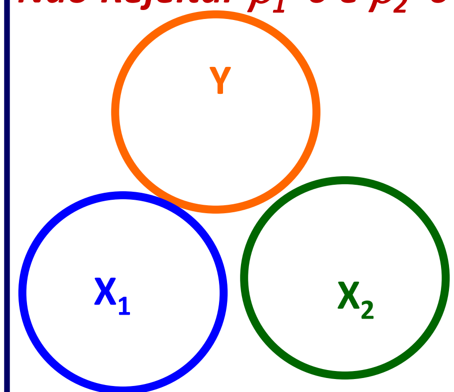
Rejeitar apenas β_1=0



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

Apenas X_1 contribui para explicar Y.

Não Rejeitar β_1=0 e β_2=0



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

Variáveis não contribuem para explicar Y.

Teste *t*– *Exemplo*

- Os testes *t* para cada coeficiente são automaticamente apresentados com a execução do comando **sumamry**;

```
> summary(modelo3.reg)

Coefficients:
               Estimate   Std. Error  t value  Pr(>|t|)
(Intercept)  -1.389918     0.626016  -2.220   0.0278 *
pib1000       0.629311     0.064612   9.740   < 2e-16 ***
I(pib1000^2) -0.008685     0.001686  -5.152   7.27e-07 ***
setor2        0.088557     0.018765   4.719   5.02e-06 ***
---
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Testes *t* para todos os coeficientes associados às variáveis independentes são significativos a 0,01% (ou até menos!).

Ou seja, todas as variáveis contribuem isoladamente para explicar a variabilidade de *co*2.

Exercícios

1) A partir do arquivo *WAGE2.xlsx*:
   a) Ajuste um modelo por MQO para a renda (*wage*) como função linear dos anos de educação (*educ*) e experiência profissional (*exper*);
   b) Interprete os coeficientes e analise a significância das estimativas;
   c) Verifique a necessidade de um termo quadrático para a experiência profissional;
   d) Analise o viés causado pela omissão da variável permanência com o mesmo empregador (*tenure*) no modelo de regressão;

Exercícios

2) A partir do arquivo Dados_PIB.XLS, que contém informações do WDB sobre sobre gastos com P&D, PIB per capita, razão de dependência para jovens e gastos com educação dos países em 2010, pede-se:
   a) Ajuste um modelo por MQO para o PIB per capita como função linear dos gastos com P&D e em educação;
   b) Interprete as estimativas dos coeficientese e suas significâncias;
   c) Haveria sentido em pressupor uma relação quadrática entre PIB e gastos com P&D?
   d) Haveria sentido em pressupor um viés na relação entre PIB per capita e gastos com educação?
   e) Analise o viés devido à omissão da variável com a razão de dependência para a população jovem;