Modelo de Regressão linear

Manuel está a pensar ingressar no Ensino Superior como meio de melhorar as suas perspectivas de emprego e remuneração

Um amigo do Manuel, o João aconselha-o a desistir da ideia dando-lhe dois exemplos:

O amigo Frederico que tem doutoramento e está desempregado

O amigo David que não fez sequer o 12º e está rico

Questões subjacentes a esta discussão:

Será que mais escolaridade está associada a melhor salário?

Quanto aumenta o salário por cada ano adicional de escolaridade?

Modelo de Regressão linear

Como já se estudou, para que a inferência estatística seja válida, a amostra tem de ser representativa da população e por isso aleatória.

Tentando encontrar uma resposta para estas questões seleccionou-se uma amostra aleatória de 100 pessoas a quem se pediu que respondessem a 2 questões:

Qual o nível de escolaridade atingido, traduzido em anos de escolaridade?

Qual o salário actual?

Com base nas respostas obtidas pretende-se saber se existe um padrão nas mesmas, isto é, se existe uma relação entre o número de anos de escolaridade e o salário.

Cada uma das respostas corresponde a um ponto $(x, y) \in \mathbb{R}^2$

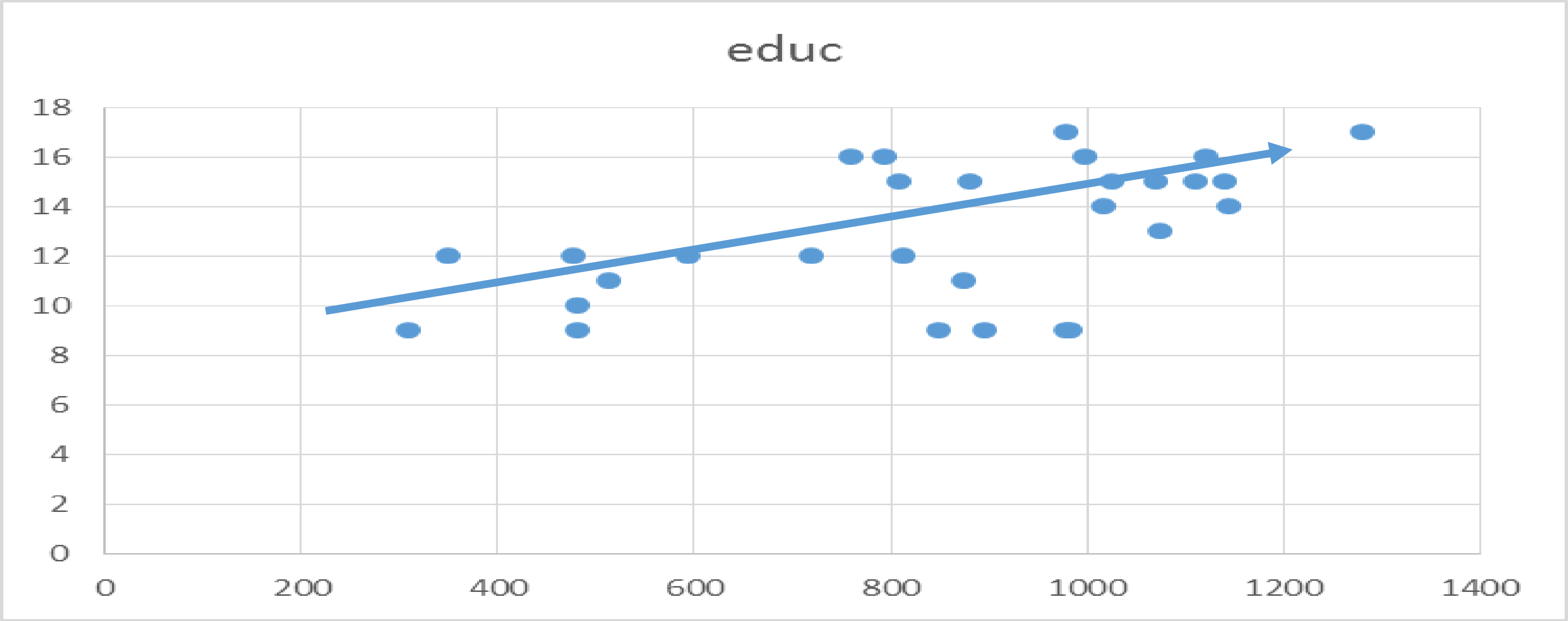
número de anos de escolaridade

Salário (u.m.)

Modelo de Regressão Linear

Qual a melhor maneira de representar a informação recolhida?   Um diagrama de dispersão

Será possível concluir algo sobre a relação entre a educação e o salário com base num diagrama de dispersão?

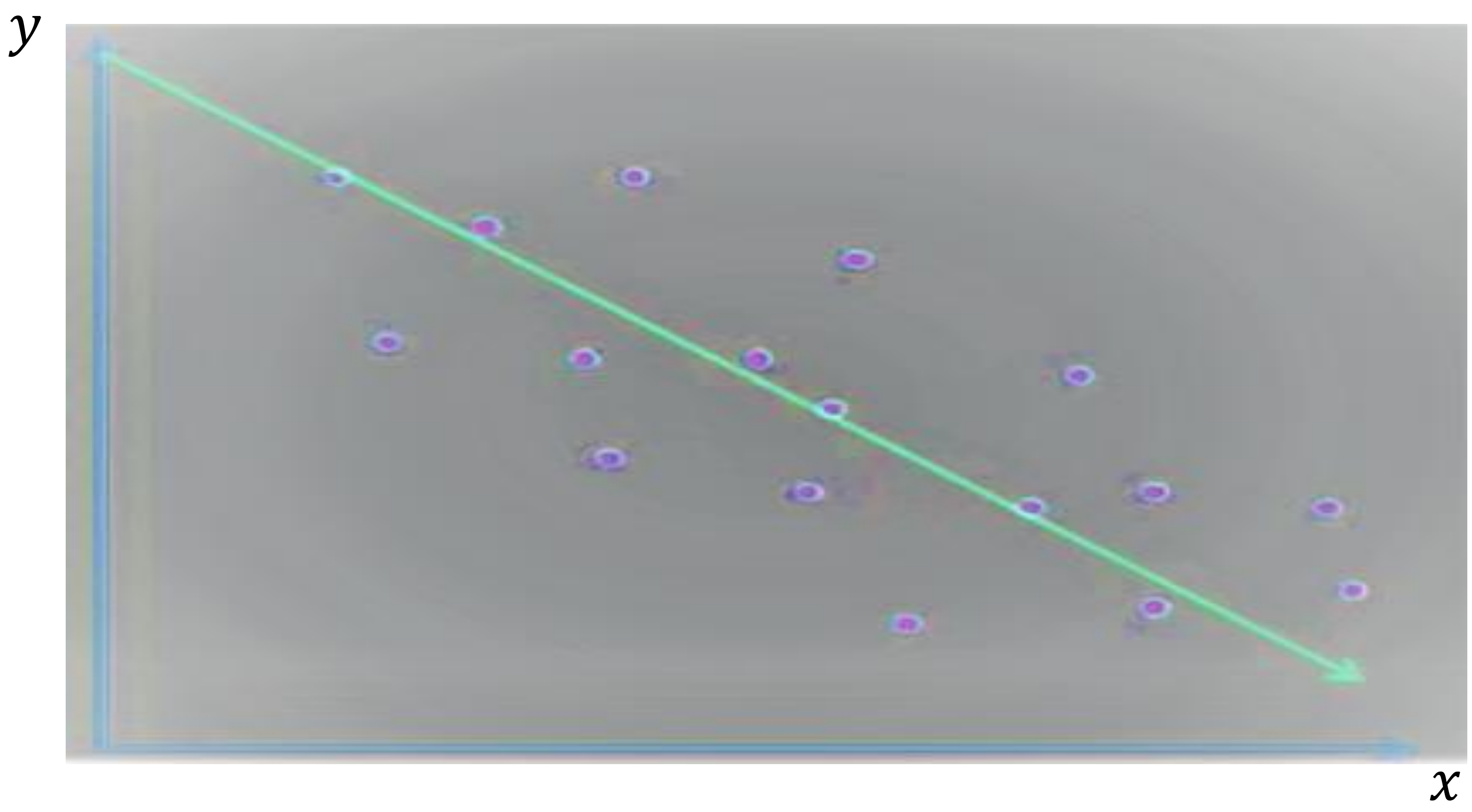


O gráfico sugere uma relação positiva entre salário e anos de escolaridade.

Esta ideia pode ser reforçada calculando o coeficiente de correlação $r_{X,Y} = 0.539$

Modelo de Regressão Linear

Com um diagrama de dispersão pode também observar-se:

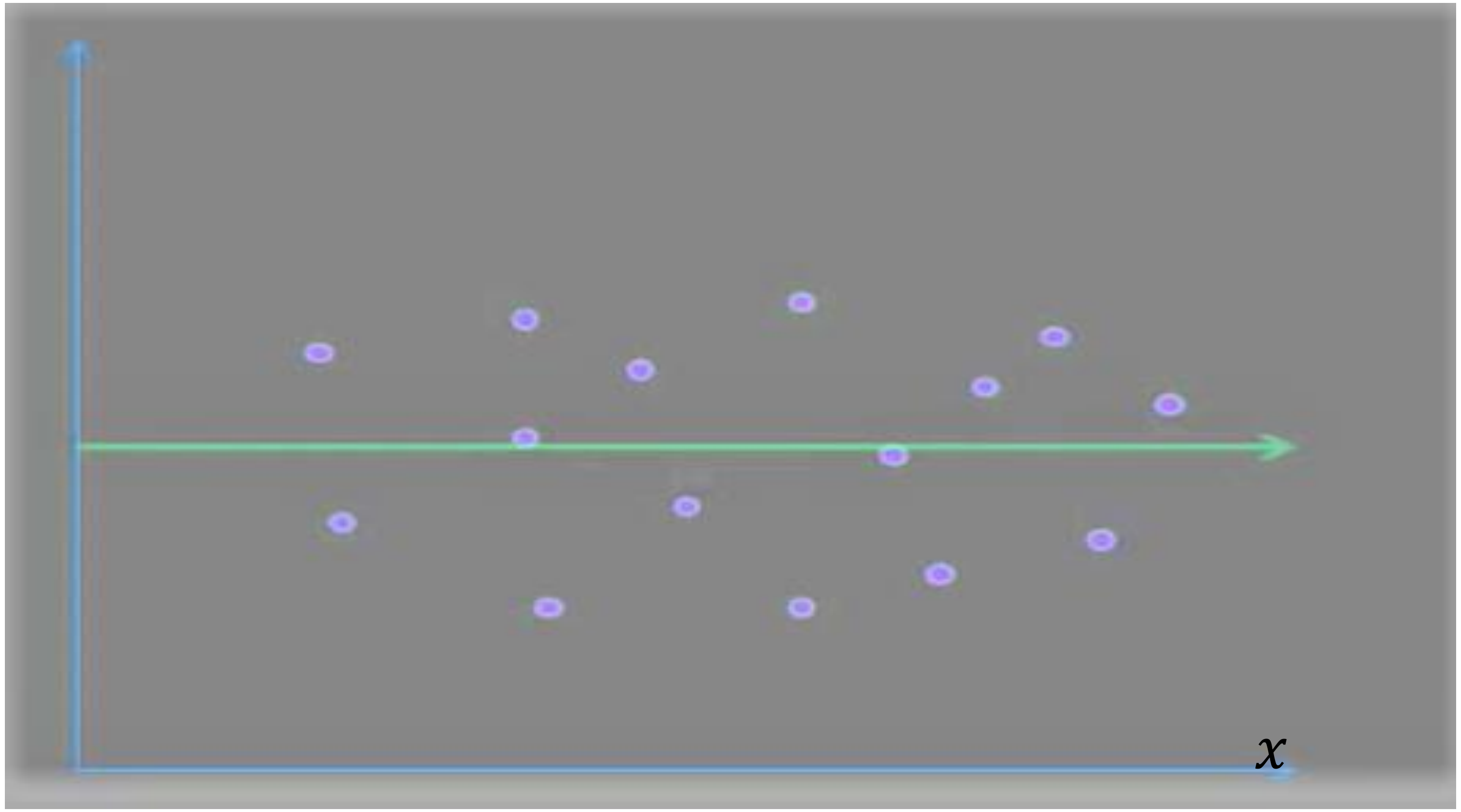


Existe uma relação negativa entre a variável X e a variável Y. $r_{X,Y} < 0$

Modelo de Regressão Linear

Ou observar-se:

y



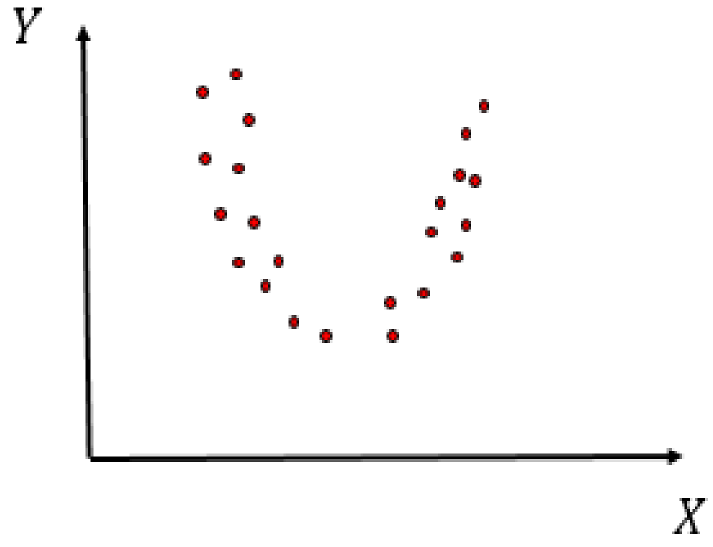
x

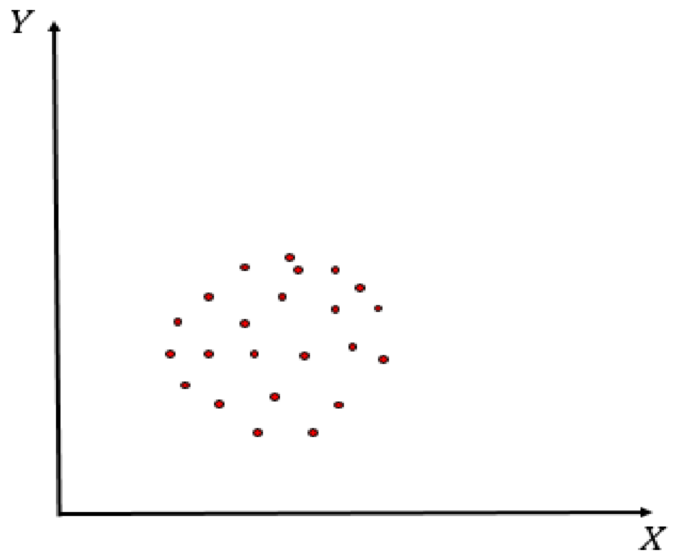
Não existe uma relação linear entre a variável X e a variável Y. $r_{X,Y} = 0$

Modelo de Regressão Linear

Importante: Sempre que se fala em relação no contexto do modelo de regressão linear está a falar-se em relação **linear**.

Os diagramas de dispersão abaixo são exemplos de casos em que embora não haja uma relação linear, pelo que $\rho_{X,Y} = 0$, existe uma relação não linear entre as variáveis.





Modelo de Regressão linear

- Objectivos

Relacionar o comportamento de uma **variável dependente\explicada** com o comportamento de uma **variável independente\explicativa** ou um conjunto **variáveis explicativas** com o intuito de:

- "**Explicar**" determinada realidade, nomeadamente explicar o valor esperado da variável dependente como função dos valores assumidos pela(s) variável(is) explicativa(s);

- "**Prever**" o comportamento da variável dependente, conhecido(s) o(s) valor(es) assumido(s) pela(s) variável(is) explicativa(s).

A distinção entre regressão simples e regressão múltipla assenta no número de variáveis explicativas no modelo.

Modelo de Regressão Linear

Distinguem-se vários tipos de modelos no que se refere à informação utilizada, nomeadamente:

Seccionais

Observações referentes a entidades independentes no mesmo período de tempo;

Exemplos: Var. dependente – média da licenciatura. Var. independente - média de acesso ao Ensino Superior de alunos que terminaram o Ensino Secundário. Ano 2016.

Var. dependente - Consumo das famílias Portuguesas. Var. independente - Rendimento disponível das famílias. Ano 2010

Nota: Supõe-se que os **dados seccionais** são obtidos por **amostragem casual** e por isso são $\boldsymbol{iid}$.

Modelo de Regressão Linear

Temporais

Observações referentes à mesma entidade ao longo de vários momentos no tempo;

Exemplos: Var. dependente – emprego,
Var. independente – salário mínimo. Observações anuais 2009-2016

Var. dependente – valor das acções de um clube desportivo,
Var. independente – montante gasto na contratação de jogadores. (2010-2016 – Observações anuais)

Nota: No caso dos **dados temporais** não é razoável supor que estes são obtidos por amostragem casual porque é difícil admitir que as observações de uma mesma variável ao longo do tempo sejam $\boldsymbol{iid}$.

Dados de painel Combinam dados sectoriais com dados temporais

Nota: O resto da exposição diz respeito, essencialmente, aos **modelos seccionais**.

Modelo de Regressão Linear

A amostra recolhida produziu 100 pontos:

Inês 12 anos de escolaridade, 348 €

Pedro
17 anos de escolaridade,
980 €

Joaquim

9 anos de escolaridade,
481 €

Teresa

11 anos de escolaridade,
502 €

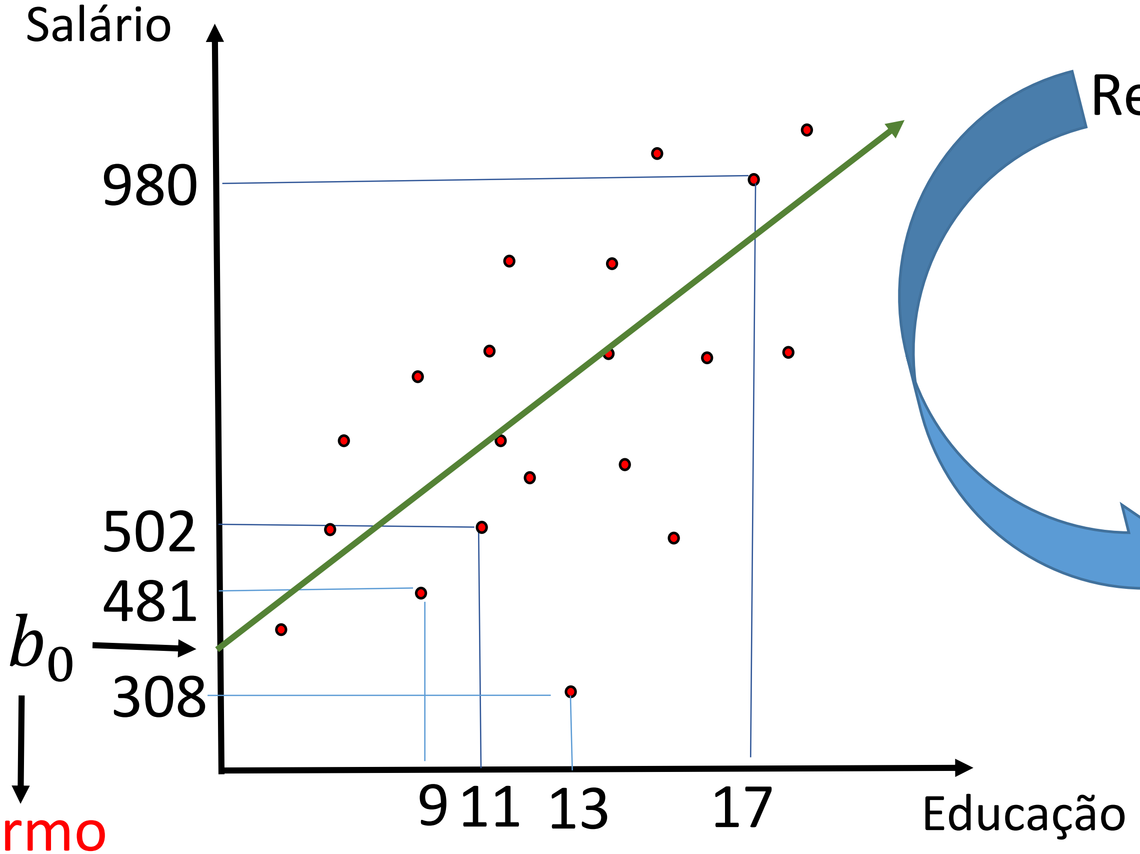
…

Modelo de Regressão Linear

Mas será possível saber mais sobre esta relação?

Sim, ajustando uma recta a esta nuvem de pontos

Qual o acréscimo de salário por cada ano de escolaridade adicional?



Termo constante

Recta $y = mx + b$ → Intersecção com eixo y

Inclinação da recta

Recta estimada

$E(salário|u) = b_0 + b_1$Educação

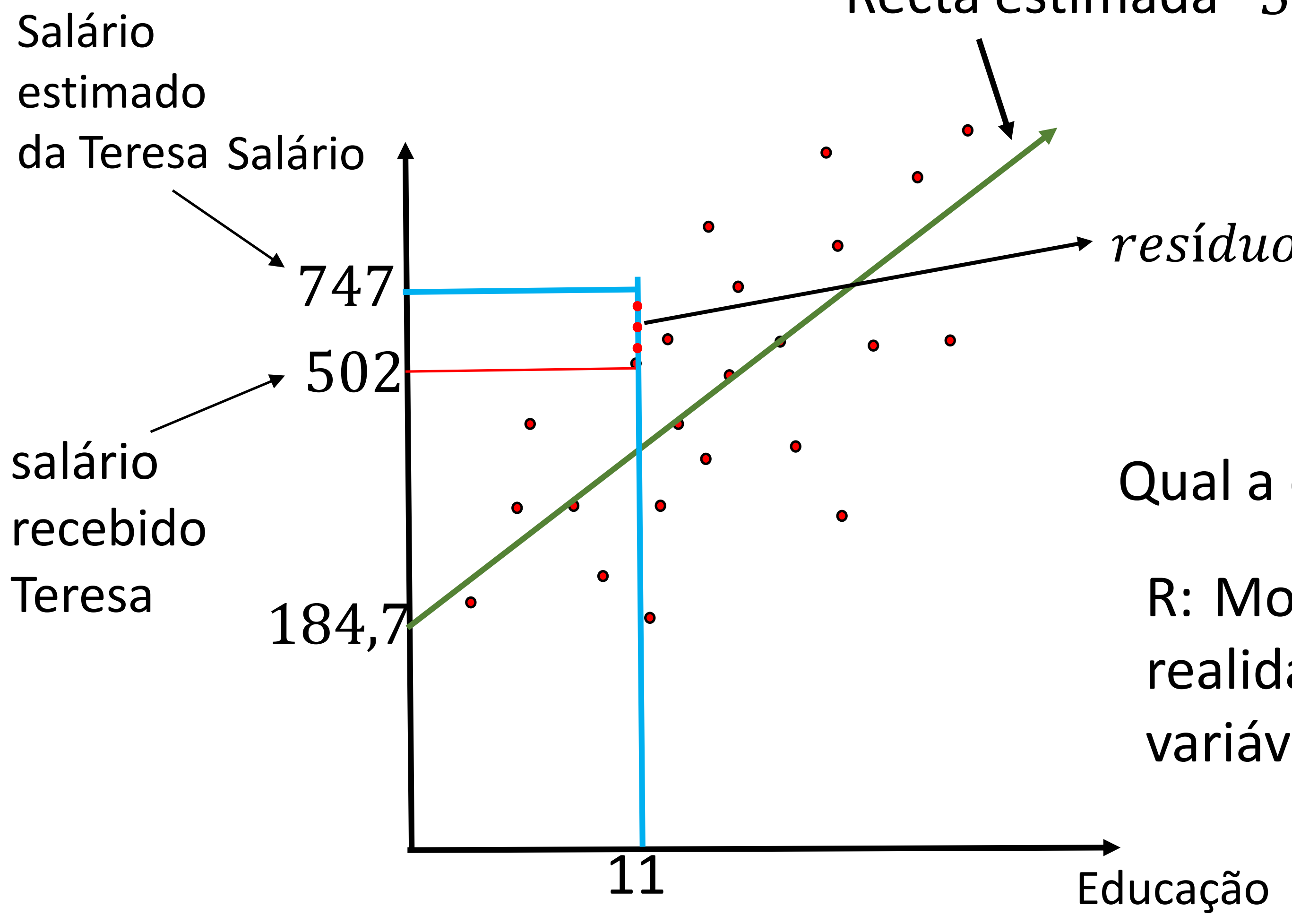
Variável explicativa (v.a.)

Variável explicada

Coeficientes Estimados

Modelo de Regressão Linear

Recta estimada $\widehat{Salário} = 184{,}7 + 51{,}12\ \text{Educação}$



$resíduo = \text{salário} - \widehat{salário}$

$= 502 - 747 = -245$

Qual a explicação para o resíduo?

R: Modelo não se ajusta perfeitamente à realidade porque há várias outras variáveis que afectam o salário

Modelo de Regressão Linear

Modelo da população $y_i = \beta_0 + \beta_1 x_i + u_i$ → Variável residual (não observável)

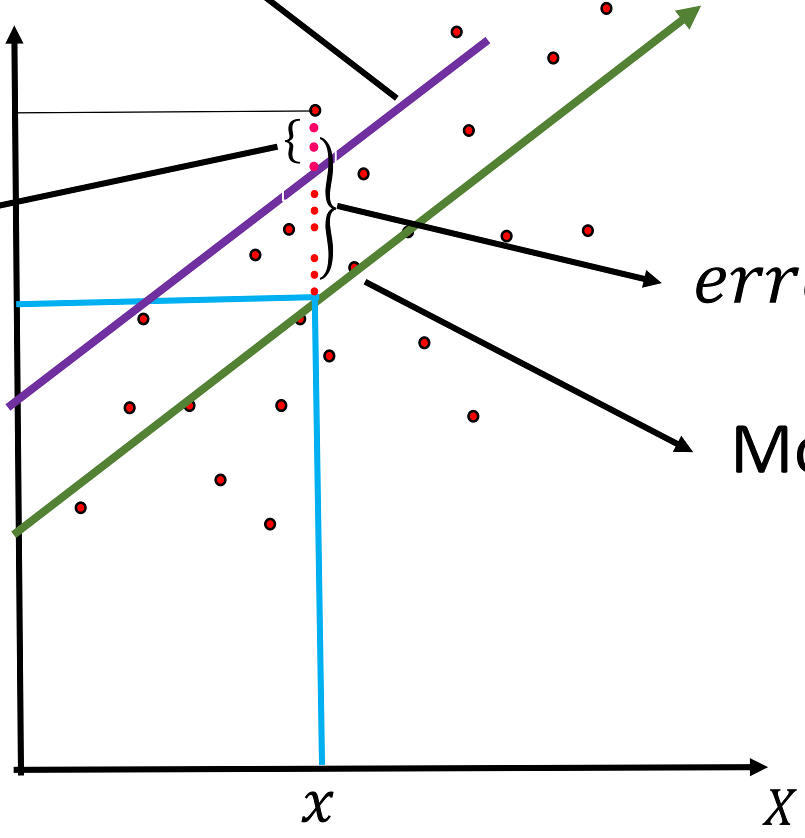
Representa o efeito de outras variáveis não consideradas no modelo

Exemplos: experiência, género, …

Modelo Regressão **Linear** $E(y|x) = \beta_0 + \beta_1 x$

Variável residual

$u = y - E(y|x)$



$erro \backslash resíduo = \hat{u} = y - \hat{y}$

Modelo estimado $\hat{y} = b_0 + b_1 x$

$\widehat{E(y|x)}$

Modelo de Regressão Linear

Terminologia do Modelo de Regressão Linear:

$\boldsymbol{y}$	$\boldsymbol{x}$
Variável dependente	Variável independente
Variável explicada	Variável explicativa
Regressando	Regressor
β_0, β_1	
Coeficientes ou parâmetros (fixos e desconh.)	
$\boldsymbol{u}$	
Variável residual (não observada)	

Modelo de Regressão Linear

Modelo da população: $y_i = \beta_0 + \beta_1 x_i + u_i \; (1)$

$\downarrow$

Modelo Regressão Linear: $E(y|x) = \beta_0 + \beta_1 x \;\; (2)$

Relação entre var. dependente e explicativa é estatística e não matemática

Hipóteses do modelo:

$H_1: E(u) = 0 \qquad (3)$

$H_2: E(u|x) = E(u) \qquad (4)$

$H_3: Cov(x, u) = 0 \;\Rightarrow$ u e x não são correlacionadas

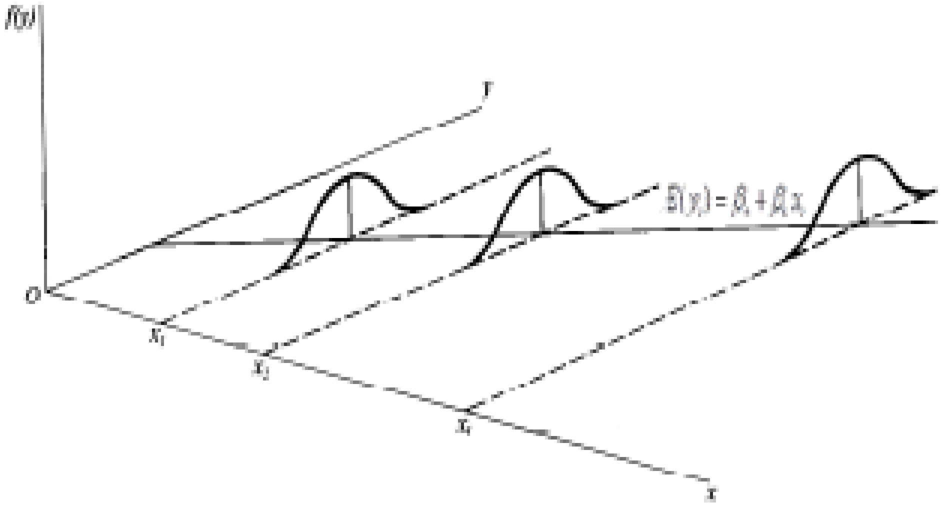
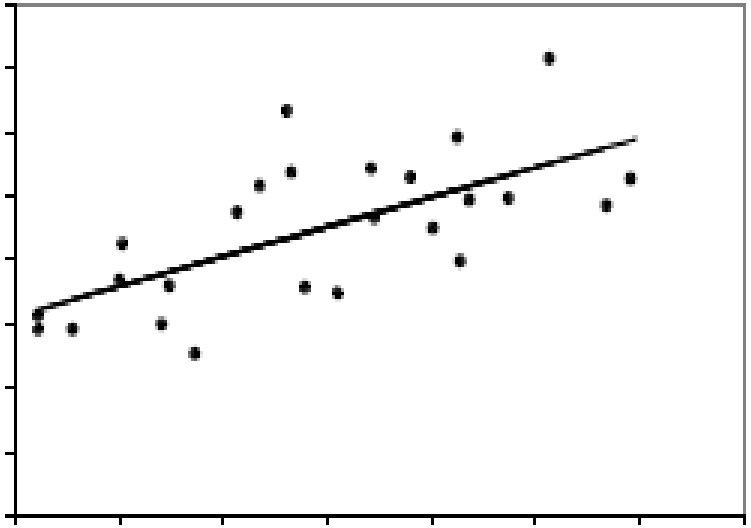
(4) O valor médio do efeito das variáveis não observadas (u) não depende do valor da var. explicativa (x) diz-se que a var. explicativa é exógena.

Modelo de Regressão Linear

Modelo da população $\longrightarrow$ $\text{Salário}_i = \beta_0 + \beta_1 \text{Educação}_i + u_i$ $\rightarrow$ Variável residual

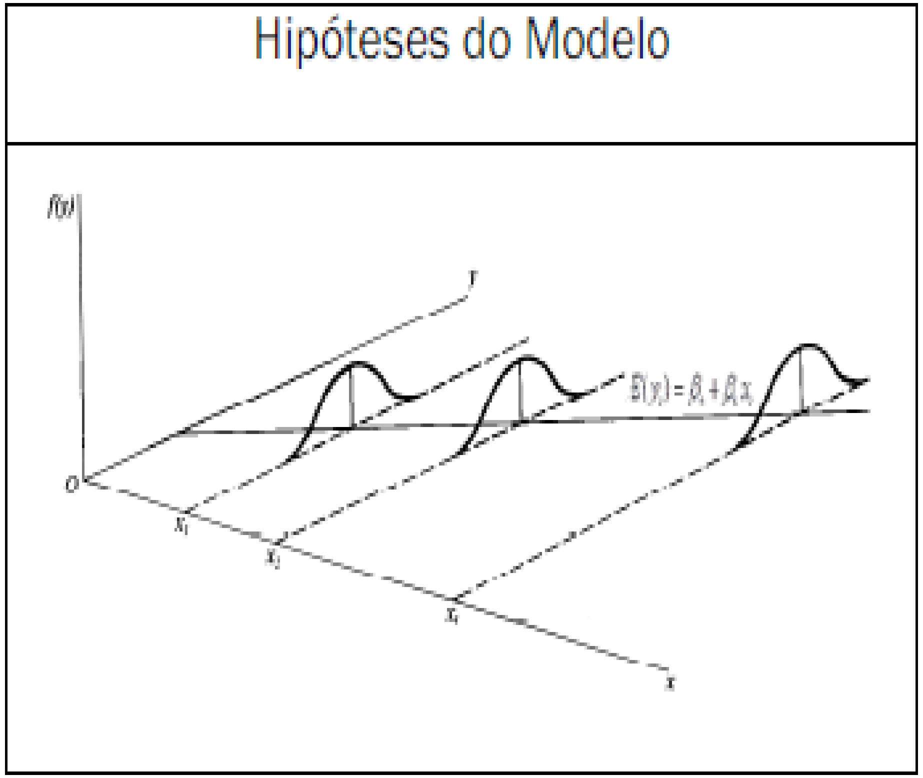
Flexibilidade

Modelo Regressão Linear $\longrightarrow$ $E(salário|X) = \beta_0 + \beta_1 \text{Educação}$

Hipóteses do Modelo	Nuvem de pontos e recta ajustada
	

Muito importante: Esta equação diz-nos como é que o valor médio do salário varia com a educação. Não diz que o salário é igual a $\beta_0 + \beta_1$Educação.

Modelo de Regressão Linear



$Var(u_i)$ é igual para todas as $i = 1, 2, \cdots, n$ observacões.

$$Var(u_i) = \sigma^2 \; i = 1, 2, \cdots, n$$

Homocedasticidade: variabilidade do salário à volta da média é constante .

Modelo de Regressão Linear

Apenas se vão estudar relações os modelos que envolvem uma **relação linear** ou **linearizável** em relação aos parâmetros, porque:

- abrangem uma variedade significativa de situações

- são de tratamento mais fácil

Muito importante: Não confundir **linearidade relativa aos parâmetros** com **linearidade relativa às variáveis**

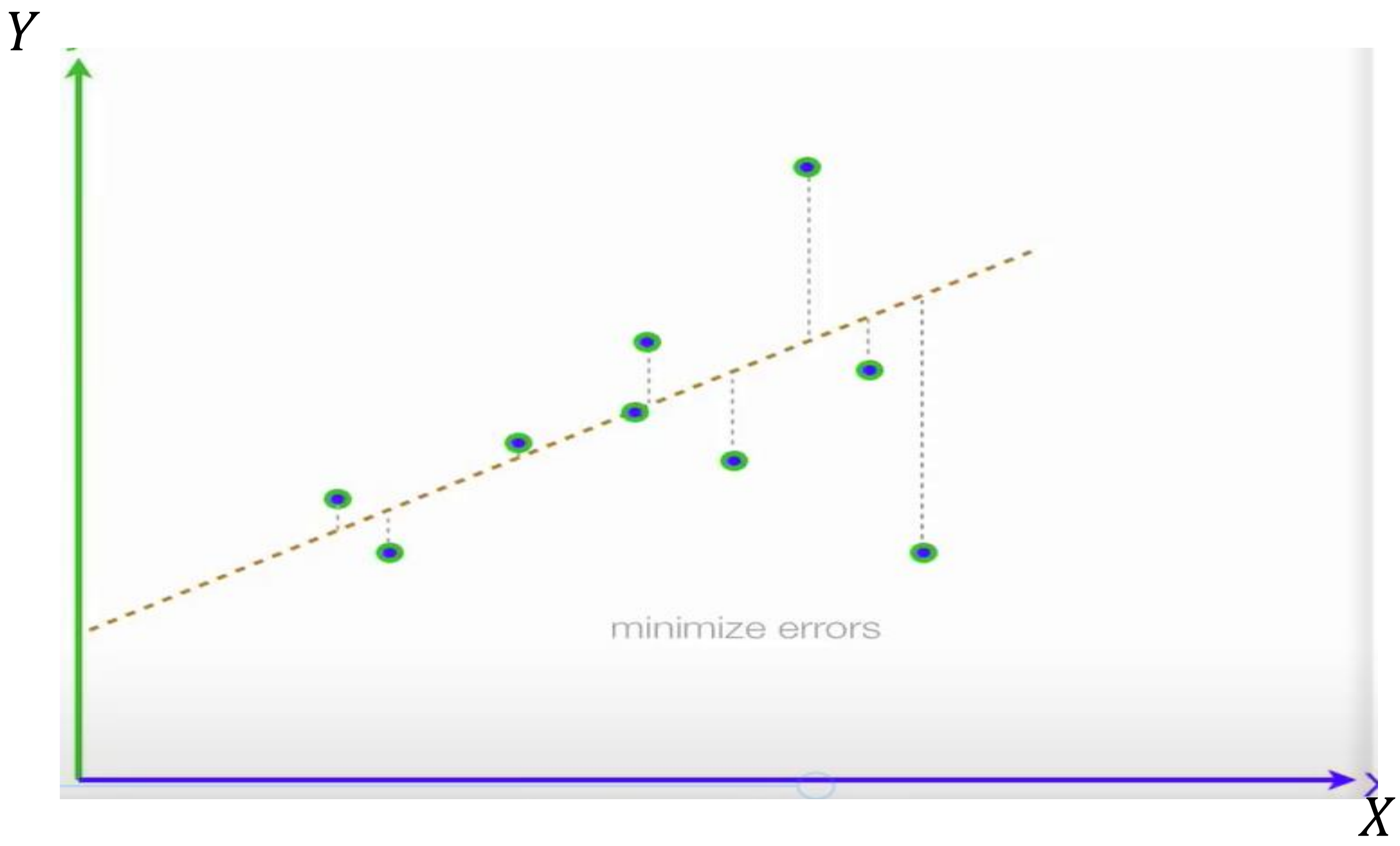
$Y = \beta_0 + \beta_1 X \longrightarrow$ É linear nos parâmetros

$Y = \beta_0 + \beta_1 X^2 \longrightarrow$ Não é linear mas é linearizável nos parâmetros

$Y = \beta_0 + {\beta_1}^2 X \longrightarrow$ Não é linear nos parâmetros

Modelo de Regressão linear

Estimação dos coeficientes de regressão pelo Método dos Mínimos Quadrados



IDEIA

Ajustamento da recta à nuvem de pontos será tanto melhor quanto menor for a distância dos pontos à recta

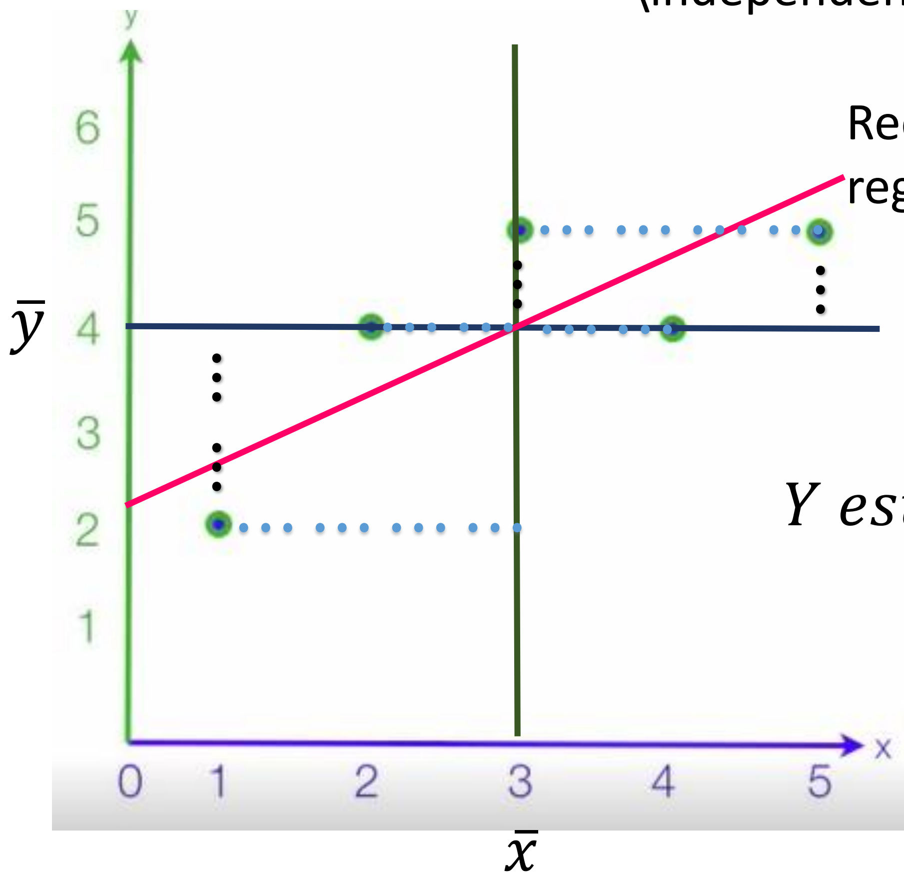
Minimizar a soma dos quadrados dos erros

Modelo de Regressão linear

Estimação dos coeficientes de regressão pelo Método dos Mínimos Quadrados

Variável dependente

Variável explicativa \independente



Recta regressão

Variável independente

Variável dependente

x	y	$x - \bar{x}$	$y - \bar{y}$	$(x - \bar{x})^2$	$(x - \bar{x})(y - \bar{y})$
1	2	-2	-2	4	4
2	4	-1	0	1	0
3	5	0	1	0	0
4	4	1	0	1	0
5	5	2	1	4	2
				10	6

Coef. estimados

$Y\ estimado$

$$\hat{y} = b_0 + b_1 x$$

Recta de regressão estimada

$$b_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} = \frac{6}{10}$$

$$E(B_1) = \beta_1\ (estimador\ não\ enviesado)$$

Qualquer recta de regressão linear passa pelo ponto $(\bar{x}, \bar{y}\ ) = (3{,}4)$

$$b_0 = \bar{y} + b_1\bar{x} \Leftrightarrow b_0 = 4 - 0.6 * 3 = 2.2$$

Modelo de Regressão Linear

Interpretação dos parâmetros

As interpretações são feitas em termos do valor esperado condicionado de y que é estimado por $\hat{y}_i$.

Para exemplificar as situações mais correntes considerem-se dois modelos:

No **1º Modelo** tem-se:

Modelo regressão linear → $E[y|x] = \beta_0 + \beta_1 x_i$

Modelo estimado → $\hat{y}_i = b_0 + b_1 x_i$

$\boldsymbol{\beta_1}$ - representa uma variação marginal, i.é, a **variação de $E[y|x]$** quando **x varia de uma unidade.** b_1 é estimativa de β_1

Modelo de Regressão Linear

Interpretação dos parâmetros

regressando

No **2º Modelo- A**: a variável explicada é z mas $\boldsymbol{y = lnz}$. Suponha-se ainda que a variável explicativa é $\boldsymbol{x}$.

Variável explicativa

Modelo de regressão → $E[\boldsymbol{y}|x\,] = \beta_0 + \beta_1 \boldsymbol{x}_i$

Modelo regressão transformado:

$$E[\boldsymbol{lnz}|x] = \beta_0 + \beta_1 \boldsymbol{x}_i$$

Modelo estimado → $\hat{y}_i = b_0 + b_1 x_i$

β_1 - diz-nos que quando **$\boldsymbol{x}$ varia de 1 unidade**, $E[z|x]$varia aproximadamente **$100 * \beta_1$%.** b_1 é estimativa de β_1.

Nota: A aproximação é tanto melhor quanto menores forem as variações de b_1

Modelo de Regressão Linear

Interpretação dos parâmetros

regressando

No **2º Modelo - B**: a variável explicada é z mas $\boldsymbol{y = lnz}$. Suponha-se ainda que a variável explicativa é $\boldsymbol{w}$, e $\boldsymbol{x} = \boldsymbol{ln}(\boldsymbol{w})$.

regressor

Modelo de regressão → $E[\boldsymbol{y}|x\,] = \beta_0 + \beta_1 \boldsymbol{x}$

Modelo regressão transformado:

$$E[\boldsymbol{lnz}|\boldsymbol{ln}(\boldsymbol{w})] = \beta_0 + \beta_1 \boldsymbol{ln}(\boldsymbol{w})$$

Modelo estimado → $\hat{y}_1 = b_0 + b_1 x_1$

β_1 - diz-nos que quando **w varia de 1%**, $\boldsymbol{E}[z|\boldsymbol{x}]$varia aproximadamente **$\beta_1$%.**

b_1 é estimativa de β_1.

Nota: A aproximação é tanto melhor quanto menores forem as variações percentuais de b_1 e $\boldsymbol{w}$

Modelo de Regressão Linear

Propriedades dos resíduos dos Mínimos Quadrados

1. $\sum_{i=1}^{n} \hat{u}_i = 0$

2. $\sum_{i=1}^{n} x_i \hat{u}_i = 0$ Erros não correlacionados com x

$$erro\backslash\text{resíduo} = \hat{u}_i = y_i - \hat{y}_i \Leftrightarrow y_i = \hat{y}_i + \hat{u}_i$$

MMQ decompõe cada y_i em duas partes

$$\frac{\sum_{i=1}^{n} y_i}{n} = \frac{\sum_{i=1}^{n} \hat{y}_i}{n} + \frac{\sum_{i=1}^{n} \hat{u}_i}{n} \Leftrightarrow \bar{y} = \bar{\hat{y}}$$

Variação **T**otal – medida da dispersão dos y_i na amostra $VT = \sum_{i=1}^{n} (y_i - \bar{y})^2$

Variação **E**xplicada – medida da dispersão dos $\hat{y}_i$ na amostra $VE = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$

Variação **R**esidual – medida da dispersão dos $\hat{u}_i$ na amostra $VR = \sum_{i=1}^{n} (\hat{u}_i)^2$

Modelo de Regressão Linear

Inferência estatística sobre o modelo

de $y_i = \hat{y}_i + \hat{u}_i$ e $\bar{y} = \bar{\hat{y}}$ Vem:

$$\underbrace{\sum_{i=1}^{n}(y_i - \bar{y})^2}_{Variação\ Total} = \underbrace{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}_{Variação\ Explicada} + \underbrace{+\sum_{i=1}^{n}\hat{u}_i{}^2}_{Variação\ Residual}$$

$VT = VE + VR$ assumindo $VT \neq 0$

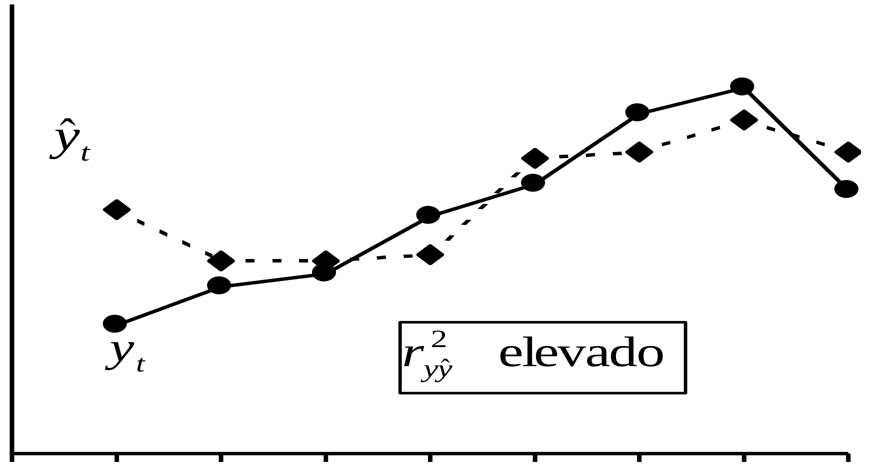
Dividindo a expressão anterior por VT vem: $1 = \frac{VE}{VT} + \frac{VR}{VT}$

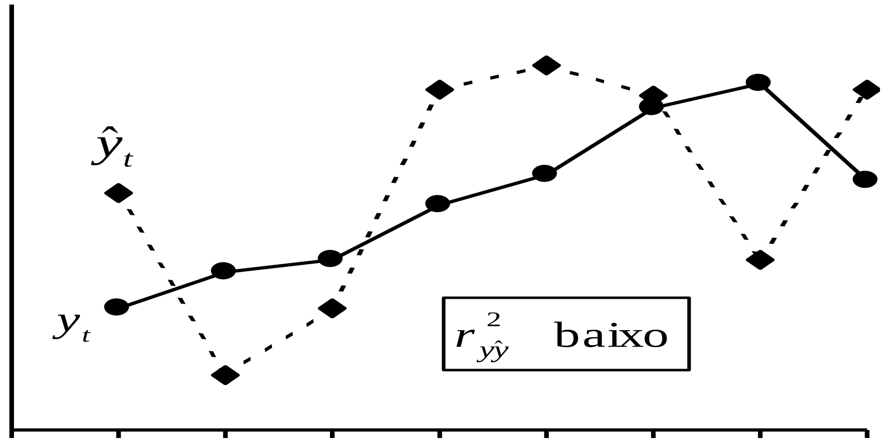
Um indicador avaliação da qualidade do ajustamento:

Coeficiente de determinação - R^2

$$R^2 = VE/VT = 1 - VR/VT$$

Modelo de Regressão Linear





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

Assim, quanto mais próximo de 1 estiver o coeficiente de determinação melhor é o "grau de ajustamento".

Notas:

1. Apenas se deve usar R^2 para comparar modelos que tenham a mesma variável dependente.

2. É uma medida de interpretação nem sempre fácil.

Modelo de Regressão Linear

Estimador não enviesado para σ^2 (variância da var. residual -u) é $\widehat{\sigma^2}$:

$$\widehat{\sigma^2} = S^2 = \frac{\sum_{i=1}^{n} \hat{u}_i{}^2}{n-2} = \frac{VR}{n-2}\ ; \qquad \sqrt{\widehat{\sigma^2}} = \hat{\sigma} = S$$ – erro padrão da regressão

$$S_{\beta_1} = \frac{\widehat{\sigma^2}}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}} \longrightarrow$$ erro padrão de $\beta 1$

Teste à nulidade do parâmetro β_1 $\qquad H_0: \beta_1 = 0 \quad contra\ \beta_1 \neq 0$

Estatística Teste $-\mathrm{T} = \frac{b_1 - \beta_1}{S_{\beta_1}} \sim t_{(n-k-1)} \qquad (\mathrm{k} = 1)$

Modelo de Regressão Linear

- Exemplo do "output" da regressão linear $\mathrm{E}(lnsal) = \beta_0 + \beta_1 Educ$
  $n = 30$

	Coeficientes não estandardizados		t	Sig.
	B	Erro Padrão		
Educ	,071	,022	3,169	,004
(Constant)	5,770	,294	19,627	,000

$\bar{y}$ =12.931; $\bar{x}$ =6.682

$\sum_{i=1}^{30}(x_i - \bar{x})^2$ =211.862;

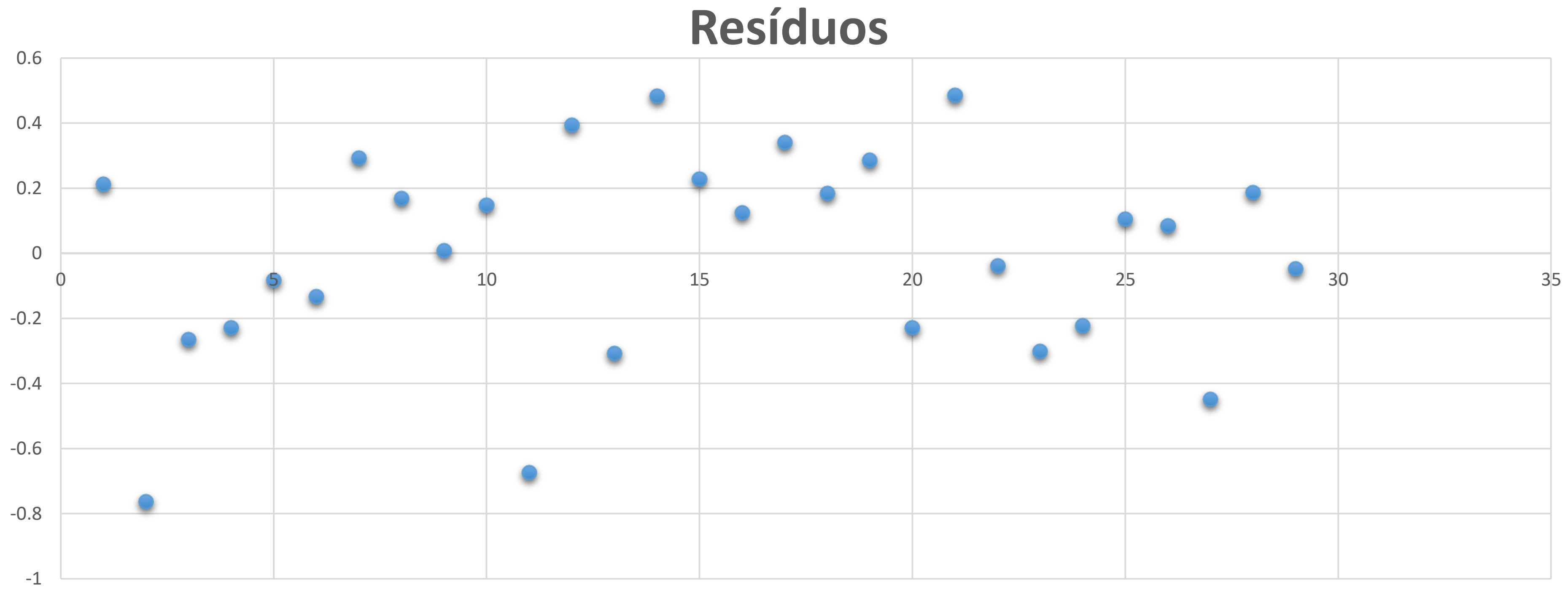
$\sum_{i=1}^{30}(x_i - \bar{x})(y_i - \bar{y})$ =14.94;

$b_1 = 14.94/211.862 = 0.071$

$b_0 = 12.931 - 0.071 * 6.682$
$= 5.77$

$$t = \frac{b_1 - \beta_1}{s_{\beta_1}} = \frac{0.071 - 0}{0.022}$$
=3.169

Modelo de Regressão Linear



ANOVA			
	Soma dos Quadrados	df	Média Quadrados
Regression	1,054	1	1,054
Residual	2,833	27	,105
Total	3,886	28	

$$R^2 = VE/VT$$
$$= 1.054/3.886$$
$$= 0.2711$$

Modelo de Regressão Linear

- Interpretação dos parâmetros estimados do modelo:

$b_1 = 0.071$

O acréscimo de um ano de escolaridade induz, **em média**, um acréscimo de salário de aproximadamente 7.1% no salário

O teste à nulidade de β_1 rejeita a nulidade o que significa que o parâmetro é significativo para um nível de significância de 5%.

Teste à nulidade do coeficiente avalia a qualidade do modelo

b_0- termo constante

Não tem qualquer interpretação de interesse neste modelo

Modelo de Regressão Linear Múltipla

- Modelo de Regressão Linear Múltipla (MRLM)

O “verdadeiro” modelo é linear e dado por:

Variáveis explicativas

Var. residual

Variável explicada

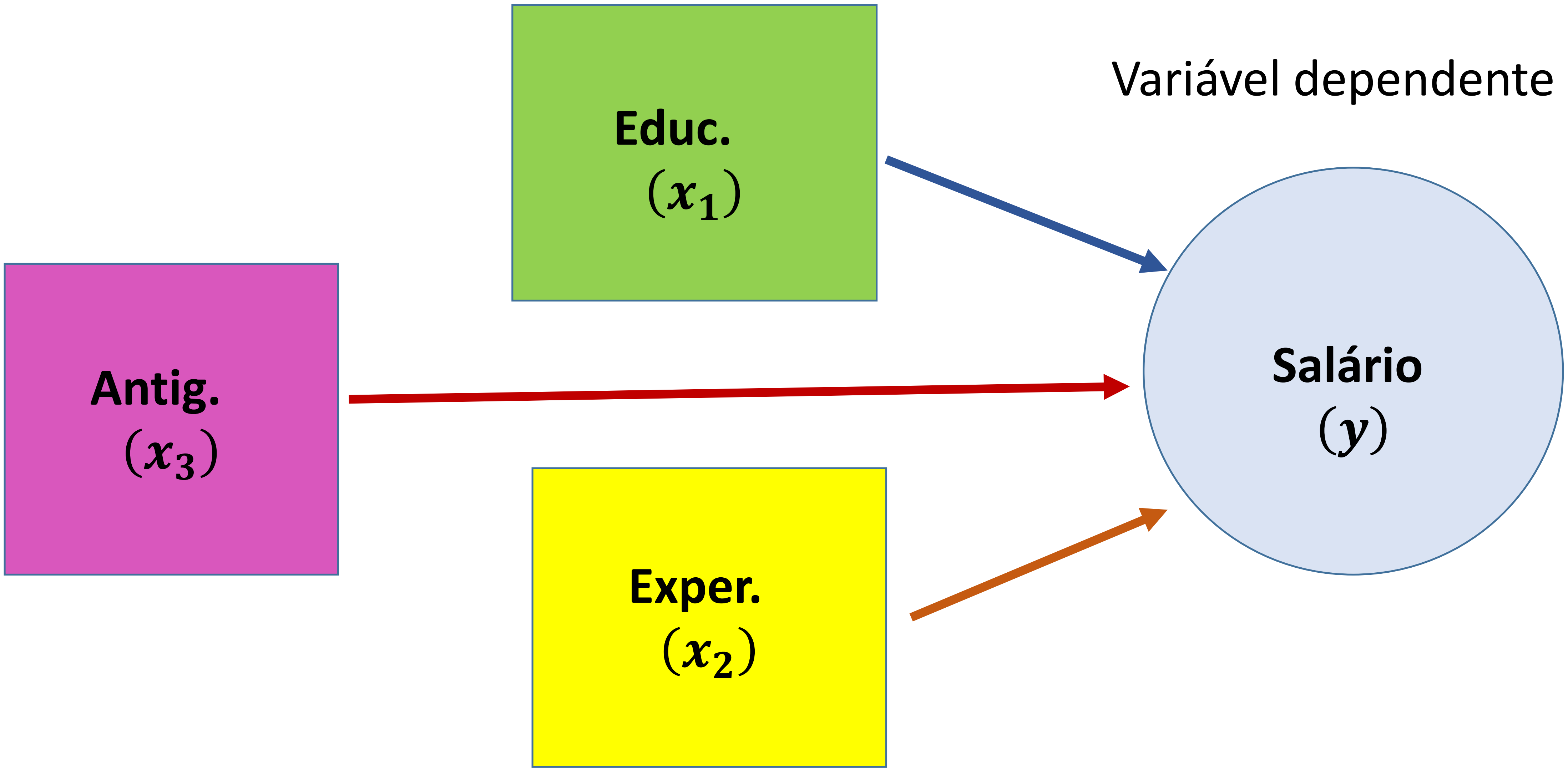
$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + u_i$$

Parâmetros\Coeficientes da regressão (fixos e desconhecidos)

Exemplo: $salário_i = \beta_0 + \beta_1 educ_i + \beta_2 exper_i + \beta_k antig_i + u_i$

Modelo teórico: $E(y_i|x_{i1}, x_{i2}, \cdots, x_{in}) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik}$

Modelo de Regressão Linear Múltipla



Modelo de Regressão Linear Múltipla

- A amostra:

Para estimar o modelo, é necessário dispor de uma amostra de dimensão n $(n > k)$ **muito maior**. (n deve ser, no mínimo, 5 a 10 vezes maior que k)

Para a amostra observada o modelo vai escrever-se:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + u_i \quad (i = 1,2,\cdots,n)$$

Nos modelos seccionais a amostra observada pode ser considerada uma amostra casual simples. O mesmo não acontece nos modelos cronológicos.

Modelo de Regressão Linear Múltipla

Hipóteses básicas do modelo:

H_1 – **Linearidade** $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + u$

H_2 – **Exogeneidade** $E(u_i|X) = 0 \quad (i = 1, 2, \cdots, n)$

Não há associação linear entre os regressores e a variável residual

H_3 – **Homocedasticidade condicionada** $Var(u_i|X) = \sigma^2 \quad (i = 1, 2, \cdots, n)$

A variância da variável residual é constante e não depende dos regressores

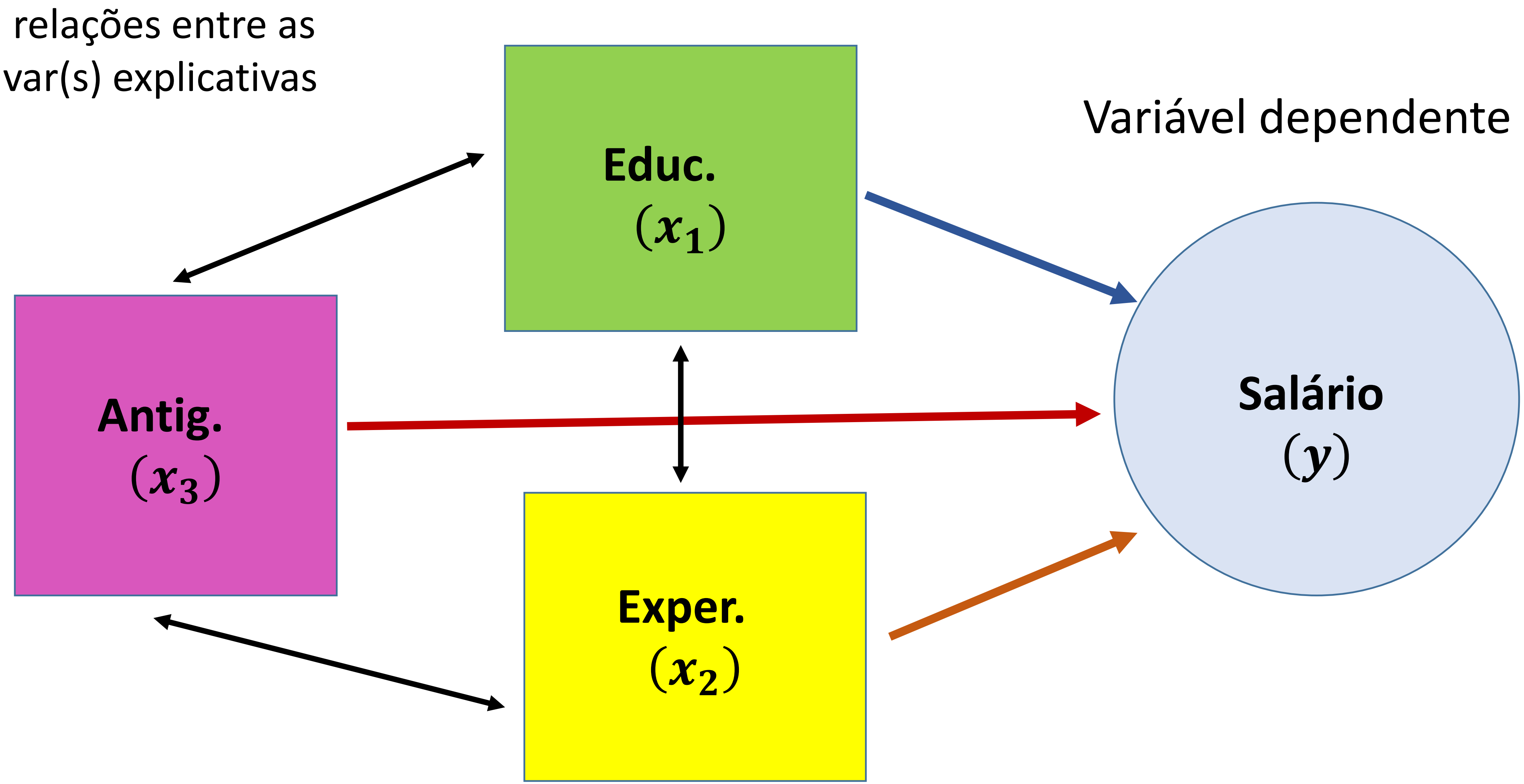
H_4 – **Ausência de autocorrelação** $Covar\left(u_i, u_j|X\right) = 0 \quad (i \neq j;\ i, j = 1, 2, \cdots, n)$

(nos modelos seccionais esta hipótese não tem grande importância)

H_5 – **Não existência de multicolinearidade exacta**

Nenhuma das variáveis explicativas é constante e não existe uma relação linear exacta entre elas.

Modelo de Regressão Linear Múltipla



Modelo de Regressão Linear Múltipla

Estimação do modelo pelo Método dos Mínimos Quadrados

1. Estimação dos coeficientes de regressão

Para estimar os $\beta_j \ (j = 1, 2, \cdots, k)$ recorre-se ao **Método dos Mínimos Quadrados**

A escolha de minimizar a soma dos quadrado dos resíduos tem por principal consequência a de dar maior peso aos grandes resíduos em detrimento dos pequenos

Função de regressão linear ajustada

$$\hat{y}_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + \cdots + b_k x_{ik}$$

Modelo de Regressão Linear Múltipla

Estimação do modelo pelo Método dos Mínimos Quadrados

2. Estimação da variância da var. residual - σ^2

- O método dos Mínimos Quadrados (MQ) permite obter as estimativas b_j
- Obtidas as estimativas b_j, estima-se σ^2 fazendo:

$$\widehat{\sigma^2} = S^2 = \frac{\sum_{i=1}^{n} \hat{u}_i^2}{n-k-1} = \frac{VR}{n-k-1} \qquad E\left(\widehat{\sigma^2}\right) = \sigma^2$$

- O erro padrão da regressão é dado por: $S = \sqrt{S^2}$
- Também se obtêm os erros-padrão dos estimadores b_j, isto é, as estimativas do desvio-padrão de cada um dos estimadores que se designam por s_{b_j}.

Modelo de Regressão Linear Múltipla

SUMMARY OUTPUT	
Regression Statistics	
Multiple R	0.4563
R Square	0.2082
Adjusted R Square	0.2050
Standard Error	0.3677
Observations	1000

ANOVA						
	df	SS	MS	F	Significance F	
Regression	4	35.3787	8.8447	65.404	3.8274E-49	
Residual	995	134.5547	0.1352			
Total	999	169.9335				
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	5.3134	0.1037	51.263	2E-281	5.1100	5.5168
educ	0.0552	0.0048	11.616	2E-29	0.0459	0.0646
exper	0.0240	0.0025	9.6429	4E-21	0.0192	0.0289
antig	0.0041	0.0024	1.7287	0.0842	-0.0006	0.0088
qi	0.0048	0.0007	6.6	7E-11	0.0034	0.0063

Notas:

Quando se acrescenta ao modelo mais um regressor, **qualquer que ele seja**, o R^2 cresce sempre.

Quando se comparam modelos com diferente nº var(s) explicativas deve usar-se o $\bar{R}^2$.

Modelo de Regressão Linear Múltipla

Exemplo – Obter a função de regressão ajustada. Recorrendo a um programa, no caso o EXCEL **que não é um software de estatística**, obtém-se directamente a função de regressão linear ajustada:

$$E(\widehat{lnsal_i}) = 5.3134 + 0.0552educ_i + 0.0240exper_i + 0.0041antig_i + 0.0048qi_i$$

Interpretação das estimativas (variável dependente de interesse está logaritmizada logo 2º modelo)

- A estimativa MQ da semi-elasticidade do salário (esperado) em relação ao número de anos de escolaridade (retorno da educação) é de 0.0552, isto é, se a escolaridade aumentar um ano, **em média**, o salário aumenta, **tudo o resto constante**, **aproximadamente** 5.52%.

- **Tudo o resto constante**, um ano mais de experiência leva, a um aumento relativo **esperado** do salário de **aproximadamente** 2.40%

-**Tudo o resto constante**, um ano mais de antiguidade leva a um aumento relativo **esperado** do salário de **aproximadamente** 0.41%

Modelo de Regressão Linear Múltipla

- Depois de estimar o modelo deve proceder-se à sua análise estatística. Só depois destas análises é que se deve utilizar o modelo.

Etapas da análise estatística:

- O ajustamento global do modelo parece adequado?

  Teste F à significância global da regressão

  Análise do R^2

- Análise individual dos coeficientes e princípio da parcimónia. Testes t

  Teste à significância estatística de um regressor

Muito importante: Caso se aceite a eliminação de algumas variáveis proceder a testes de nulidade conjunta deste sub-grupo. Regra geral, é preferível "guardar" uma variável irrelevante do que rejeitar uma variável de interesse.

Modelo de Regressão Linear Múltipla

Inferência estatística - O modelo de regressão linear

$\boldsymbol{H_6}$ – **Distribuição normal da variável residual** $\quad u_i|X \sim N(0, \sigma^2)$

<u>Inferência estatística sobre a variância das variáveis residuais</u>

Estatística Teste $\quad Q = \dfrac{\sum_{i=1}^{n} \hat{u}_i^2}{\sigma^2} = \dfrac{(n-k)s^2}{\sigma^2} \sim \chi^2_{(n-k-1)}$

<u>Inferência estatística sobre um coeficiente de regressão isolado</u>

Estatística Teste $\quad \mathrm{T} = \dfrac{b_j - \beta_j}{S_{\beta_j}} \sim t_{(n-k-1)} \text{ para } \quad (j = 2, 3, \cdots, k)$

Modelo de Regressão Linear Múltipla

Inferência estatística - O modelo de regressão linear

Casos mais frequentes: Estatística Teste $T_j = \dfrac{b_j - \beta_j}{S_{\beta_j}} \sim t_{(n-k-1)}$ para $(j = 2, 3, \cdots, k)$

Teste à significância estatística de um regressor: $H_0\!: \beta_j = 0 \quad contra\ \beta_j \neq 0$

Este teste é feito pela generalidade dos programas.

Teste ao sinal de um coeficiente: $H_0\!: \beta_j = 0 \quad contra\ \beta_j < 0$

Teste para um valor particular de um coeficiente: $H_0\!: \beta_j = c \quad contra\ \beta_j \neq c$

Quando a variável residual não tem distribuição normal mas a amostra é grande pode-se utilizar:

$$T_j = \frac{b_j - \beta_j}{S_{\beta_j}} \dot{\sim} N(0,1)$$

Modelo de Regressão Linear Múltipla

Inferência estatística sobre uma combinação linear dos coeficientes de regressão

Teste à nulidade de um subconjunto de coeficientes de regressão

$$H_0: \beta_{p+1} = 0, \beta_{p+2} = 0, \cdots, \beta_k \quad contra \quad H_1: \exists\, \beta_j \neq 0 \;\; (j = p+1, \cdots, k)$$

Para tal concebe-se um teste em 3 passos:

1 – estimar o modelo sem restricções , i.é, com todos os regressores e obter $VR_1 = \sum_{i=1}^{n} \widehat{u}_i^2$

2 – estimar o modelo com restricções, i.é, eliminando os regressores considerados nulos e obter $VR_0 = \sum_{i=1}^{n} \hat{u}_i^2$

3 – Comparar os modelos utilizando:

$$F = \frac{(VR_0 - VR_1)/\left(\overbrace{k-p}^{m}\right)}{VR_1/(n-k-1)} \sim F_{(m, n-k-1)}$$

Modelo de Regressão Linear Múltipla

Teste à significância global da regressão = nulidade de todos coeficientes de regressão

$$H_0: \beta_1 = \beta_2 = \cdots = \beta_k = 0 \quad contra \quad H_1: \exists\, \beta_j \neq 0 \;\; (j = 1, \cdots, k)$$

A não rejeição da hipótese H_0 leva a que o modelo deva ser posto de parte.

Não rejeitar a hipótese nula **corresponde a verificar que o modelo proposto não é adequado, na sua globalidade**, para descrever o comportamento do regressando.

Estatística teste: $$F = \frac{R^2/k}{(1-R^2)/(n-k-1)} = \frac{VE/k}{VR/(n-k-1)} \sim F_{(k,n-k-1)}$$

Estatística calculada por todos os software

A **região de rejeição** situa-se na **aba direita** da distribuição F.

Modelo de Regressão Linear Múltipla

Comentários:

- Se o teste individual de cada um dos coeficientes incluídos em H_0 não rejeita a nulidade e o teste conjunto a rejeita, **desconfiar de uma possível multicolinearidade**

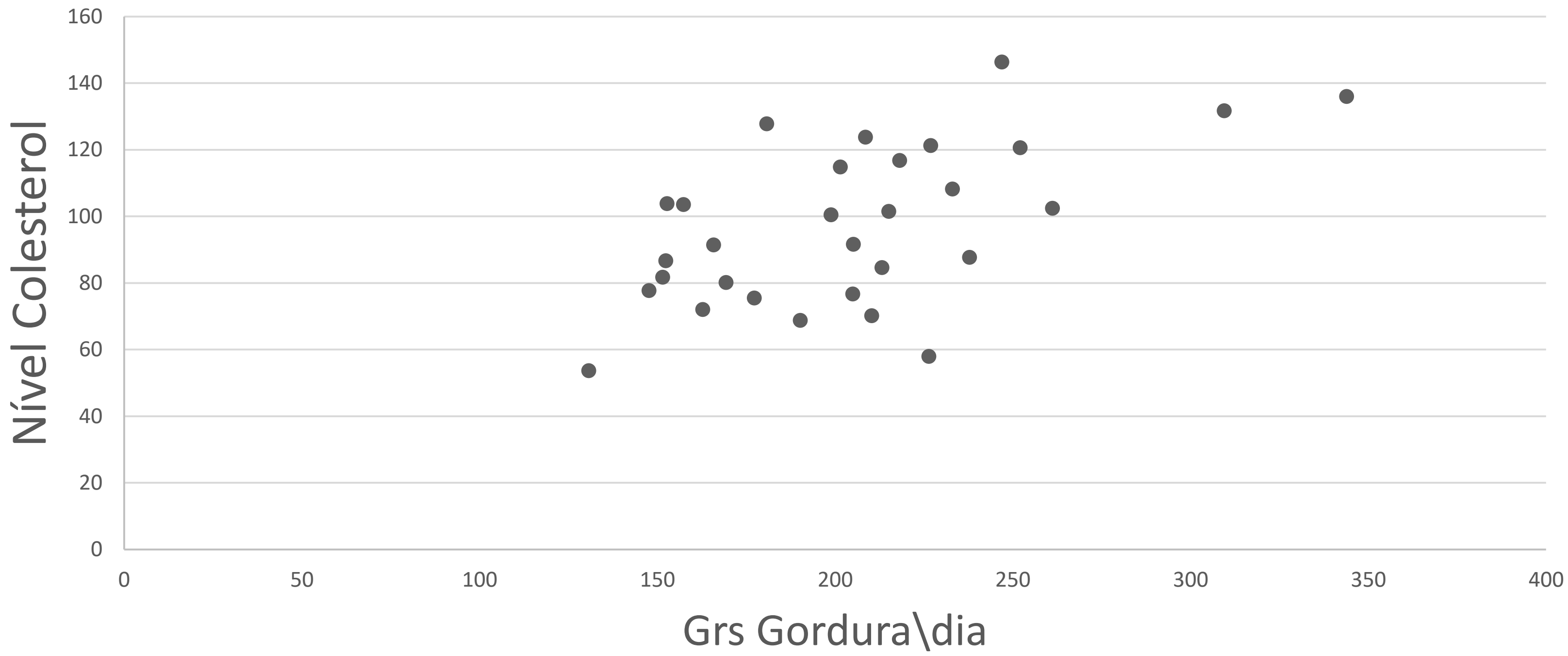
- A situação inversa

  (não se rejeita a nulidade conjunta de alguns regressores, com o teste F, mas rejeita-se para um particular coeficiente pelo teste ***t***)

  também é possível, mas neste caso é geralmente preferível confiar no teste ***t***.

Modelo de Regressão Linear Simples

- Exercício 4



	Qtde Colesterol	Grs Gordura\dia
Qtde Colesterol	1	
Grs Gordura\dia	0,586050325	1

Modelo de Regressão Linear Simples

- Exerc. 4

SUMMARY OUTPUT					
Regression Statistics					
Multiple R	0,586050325				
R Square	0,343454983				
Adjusted R Square	0,320006947				
Standard Error	39,39128311				
Observations	30				
ANOVA					
	df	SS	MS	F	Significance F
Regression	1	22728,12448	22728,12	14,64749	0,000667
Residual	28	43446,84918	1551,673		
Total	29	66174,97367			
	Coefficients	Standard Error	t Stat	P-value	
Intercept	91,5885625	30,5126691	3,001657	0,005594	
Grs Gordura\dia	1,167693221	0,305103428	3,827205	0,000667	

Modelo de Regressão Linear Simples

- Exercício 4

c) $H_0: \beta_1 = 1 \quad contra \quad H_0: \beta_1 > 1$

Estatística teste $\frac{b_1 - \beta_1}{s_{\beta_1}} \sim t_{(\underbrace{n-2}_{28})}$ $\qquad t_{obs} = \frac{1,1677 - 1}{0,3051} = 0,5496$

$$Valor - p = P\left(t_{(28)} > 0,5496\right) = 0,2935$$

d) $IC_{\beta_1}^{0.9} = ?$ $\quad$ variável fulcral: $\frac{b_1 - \beta_1}{s_{\beta_1}} \sim t_{(\underbrace{n-2}_{28})}$ $\qquad \alpha = 0,9 \Rightarrow t_{\alpha/2} = 1,701$

$$IC_{\beta_1}^{0.9} = (1.1677 - 1.701 * 0.3051,\ 1.1677 + 1.701 * 0.3051)$$
$$= (0.6487, 1.6867)$$

Modelo de Regressão Linear Múltipla

- Exercício 8 $$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + u$$

$$R^2 = 0.7 \qquad \sum_{i=1}^{30} \hat{u}_i^2 = 1.69$$

a) $$S^2 = \frac{\sum_{i=1}^{n} \hat{u}_i^2}{n-k-1} \Rightarrow s^2 = \frac{1.69}{30-4-1} = 0{,}0676$$

b) $H_0: \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0 \;\; contra \;\; H_1: \exists\, \beta_j \neq 0 \;\; (j = 1, \cdots, k)$

$$F = \frac{R^2/k}{(1-R^2)/(n-k-1)} \sim F_{(k,n-k-1)} \qquad f_{obs} = \frac{0.7/4}{(1-0.7)/(30-5)} = 14.5833$$

$$Valor - p = P\left(F_{(4,25)} > 14.5833\right) = 2{,}83805\text{E}{-}06 \approx 0$$

Modelo de Regressão Linear Múltipla

Exercício 12.

a) LEDUC – Por cada acréscimo de 1% no nº anos escolaridade o salário cresce 1.48%

IT – Por cada ida adicional ao teatro o salário decresce 2.9%

b) $H_0: \beta_{IT} = \beta_{HD} = 0 \quad contra\ H_1: \exists \beta_j \neq 0 \quad j = IT, HD$

Nº var(s) explicativas no modelo sem restrições

Estatística teste: $F = \dfrac{(VR_0 - VR_1)/\left(\overbrace{k-p}^{m}\right)}{VR_1/(n-k-1)} \sim F_{(m, n-k-1)}$

$k = 4, p = 2$

Nº β_j considerados nulos na H_0 = nº restrições

$$f = \frac{(47.2140 - 47.0867)/\left(\overbrace{4-2}^{2}\right)}{47.0867/(500-4-1)} = 0.669$$

$$Valor - p = P\left(F_{(2,495)} > 0.669\right) = 0.5127$$

Modelo de Regressão Linear Múltipla

Exercício 12.

c) $H_0: \beta_{LSAL} = 0.05 \quad contra \quad H_0: \beta_{LSAL} \neq 0.05$

$$t_{obs} = \frac{0.387138 - 0.5}{0,032692} = -3,45228$$ Estatística teste $\frac{b_j - \beta_j}{s_{\beta_j}} \sim t_{(\underbrace{n-4-1}_{495})}$

ou

$$\frac{b_j - \beta_j}{s_{\beta_j}} \dot{\sim} N(0,1)$$

$$Valor - p = P\left(t_{(495)} > |-3,45228\,|\right) = 0,000603$$

$IC_{\beta_1}^{0.9} = ?$

$$Valor - p = P(Z > |-3,45228|\,) \approx 0,000556$$

Modelo de Regressão Linear Múltipla

- Exercício 12

c) (continuação)

$IC_{LSAL}^{0.95} = ?$ Variável fulcral: $\frac{b_{LSAL}-\beta_{LSAL}}{s_{\beta_{LSAL}}} \sim t_{(\underbrace{n-k-1}_{495})}$

$$1-\alpha = 0{,}9 \Rightarrow \alpha = 0.1 \Rightarrow \frac{\alpha}{2} = 0.05 \Rightarrow t_{\alpha/2}: P\left(t_{(495)} > t_{\alpha/2}\right) = 0.05$$

$$t_{\alpha/2} = 1.647$$

$$IC_{\beta_1}^{0.9} = (0.3871 - 1.647 * 0{,}0327, 0.3871 + 1.647 * 0{,}0327)$$

$$= (0.3332,\ 0.441)$$