

---

# Estatística: Aplicação ao Sensoriamento Remoto

SER 204

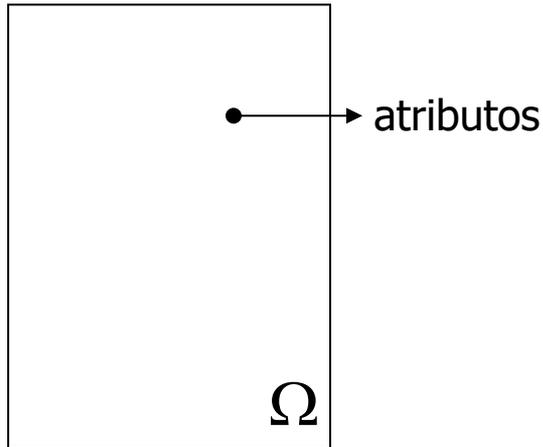
Análise de Regressão

Camilo Daleles Rennó

camilo.renno@inpe.br

acesso do conteúdo do curso em [Bibdigital do INPE](#) ou [GitHub](#)

# Relacionamento entre Variáveis



Em qualquer tipo de estudo, há sempre a necessidade de se focar em um ou mais atributos (características) dos elementos que compõem esta população ( $\Omega$ )

atributos quantitativos:

- . altura total
- . diâmetro da copa
- . diâmetro do tronco (DAP)
- . biomassa
- . etc

Estes atributos constituem as variáveis em estudo.

Quando adquiridas sobre o mesmo indivíduo, estas variáveis guardam alguma relação entre si?

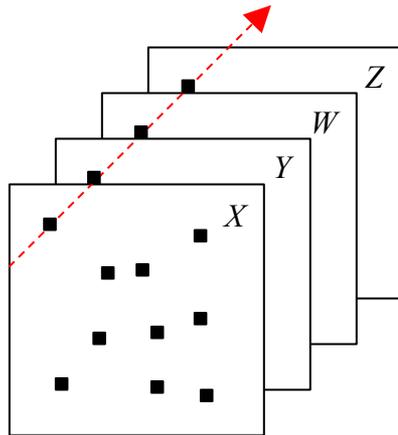
Há como aproveitar o conhecimento dessas relações em estudos dessa população?



gimm.at

# Relacionamento entre Variáveis

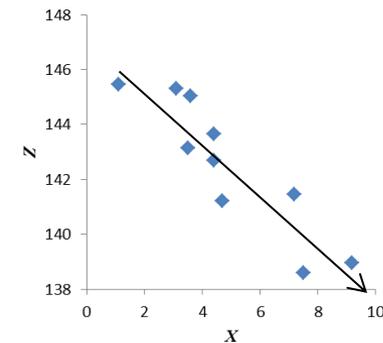
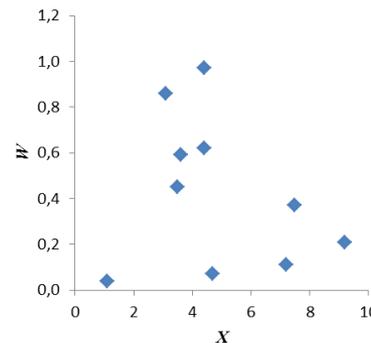
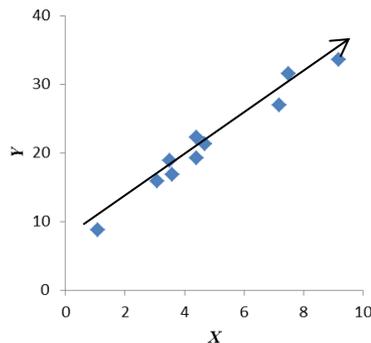
Em imagens ou mapas, o relacionamento aparece pela posição geográfica



Amostra	$X$	$Y$	$W$	$Z$
1	3,5	18,9	0,45	143,2
2	7,5	31,5	0,37	138,6
3	4,4	22,2	0,62	142,7
4	1,1	8,7	0,04	145,5
5	4,4	19,2	0,97	143,7
6	4,7	21,3	0,07	141,2
7	7,2	27,0	0,11	141,5
8	3,6	16,8	0,59	145,1
9	9,2	33,6	0,21	139,0
10	3,1	15,9	0,86	145,3

mesma posição geográfica

Diagrama de dispersão

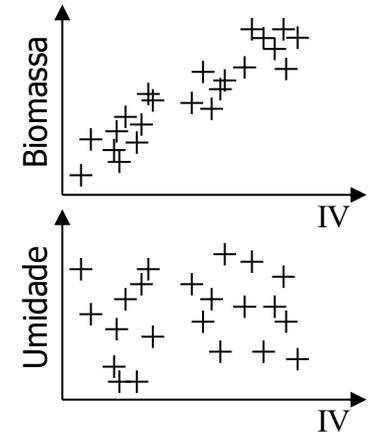


Muitos estudos buscam entender as relações de dependência entre variáveis de modo a construir modelos que permitam prever o comportamento de uma variável conhecendo-se os valores de outra ou outras variáveis

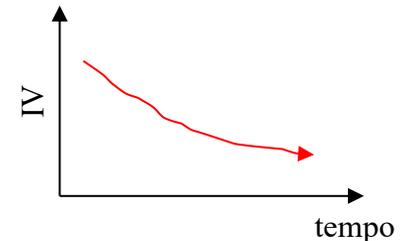
# Relacionamento entre Variáveis

Por exemplo:

Considere que um determinado índice de vegetação (IV) apresenta valores baixos para vegetações com pequena biomassa e apresenta valores altos para vegetações com grande biomassa. Por outro lado, este mesmo índice não tem qualquer relação com a umidade superficial do solo.



Se observarmos uma diminuição do valor deste índice de vegetação ao longo do tempo, o que podemos concluir quanto à mudança na biomassa da vegetação e na umidade superficial do solo deste lugar?



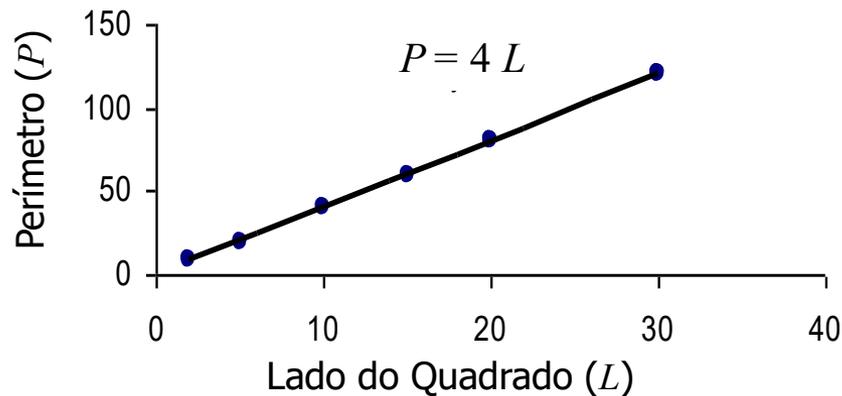
Quanto à biomassa, espera-se que tenha havido uma diminuição  
Quanto à umidade, nada podemos afirmar

# Relação funcional x Relação estatística

As variáveis podem possuir dois tipos de relações:

1) **Funcional:** a relação é expressa por uma fórmula matemática:  $Y = f(X)$

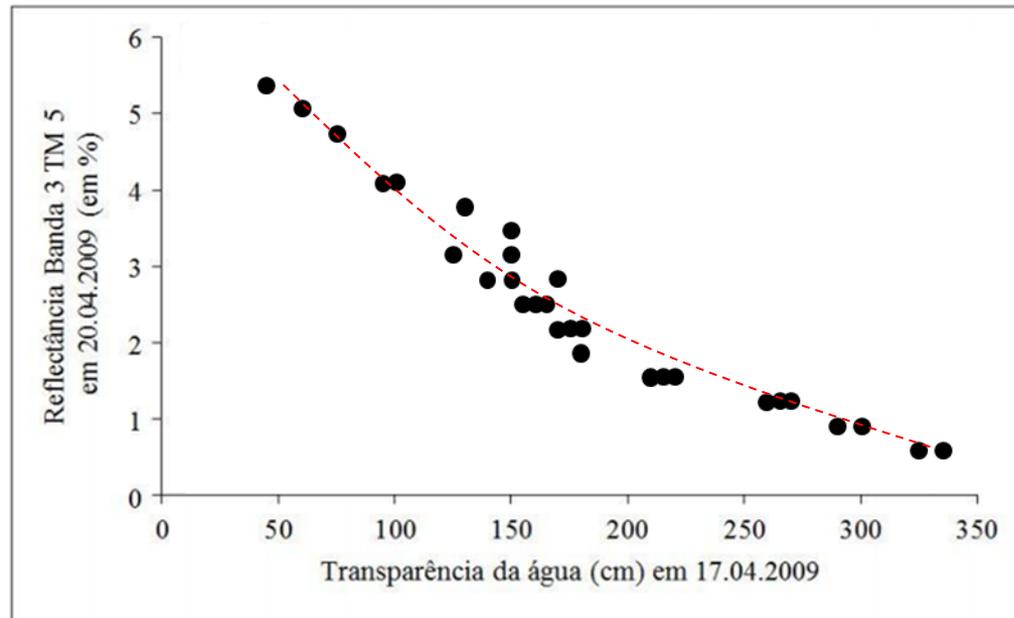
Ex: relação entre o perímetro ( $P$ ) e o lado ( $L$ ) de um quadrado



Todos os pontos caem perfeitamente sobre a linha que representa a relação funcional entre  $L$  e  $P$

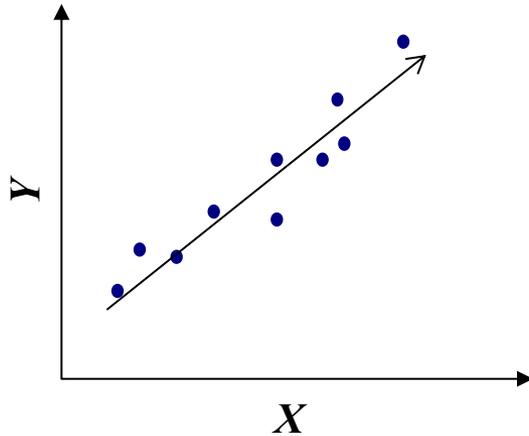
# Relação funcional x Relação estatística

- 2) **Estatística:** não há uma relação perfeita como no caso da relação funcional.  
As observações em geral não caem exatamente na linha que descreve a relação.  
Ex: relação entre transparência da água e a reflectância na banda 3 TM5

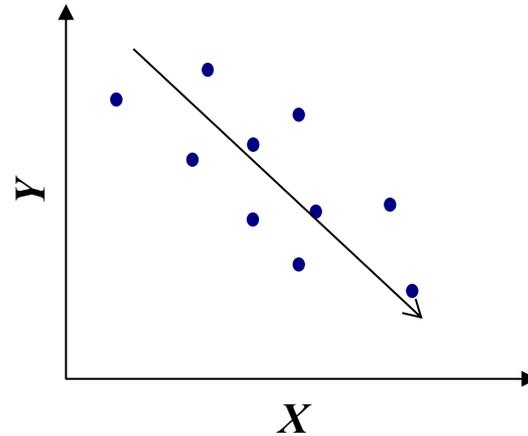


Fonte: Adaptado de Santos, F.C.; Pereira Filho, W.; Toniolo, G.R.. Transparência associada à reflectância da água do reservatório Passo Real. In: XVII SBSR, 2015. p. 6653-6659

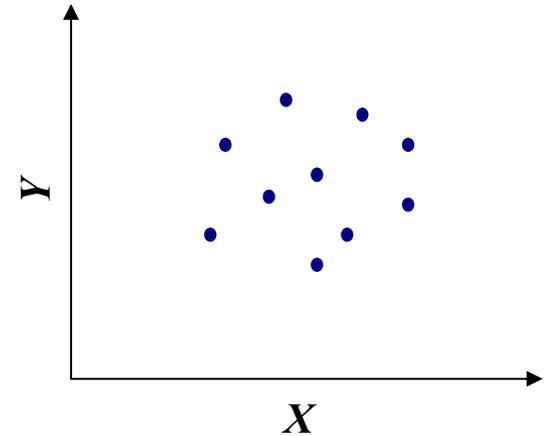
# Grau de Relacionamento



Relação direta ou positiva



Relação inversa ou negativa



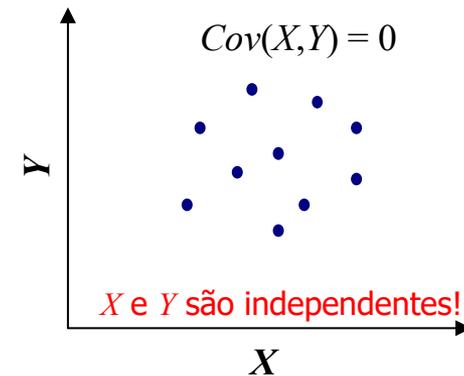
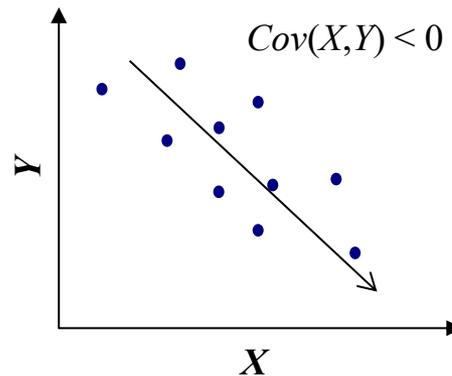
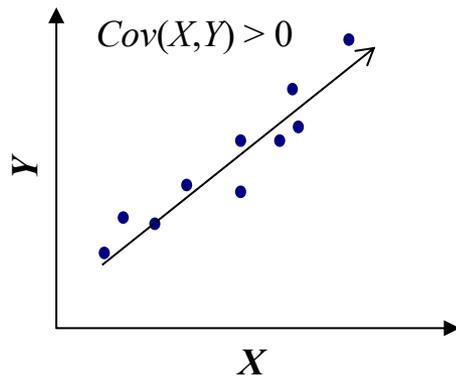
Ausência de relação

Como caracterizar o grau de relacionamento (ou associação) entre pares de variáveis?

Covariância  
Coeficiente de Correlação

# Covariância

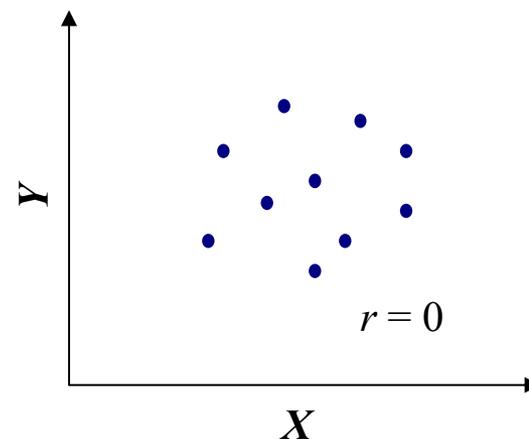
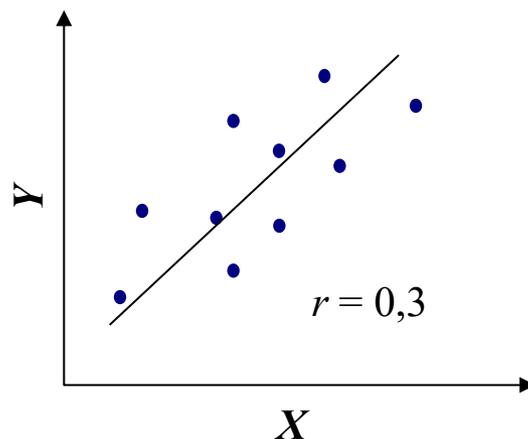
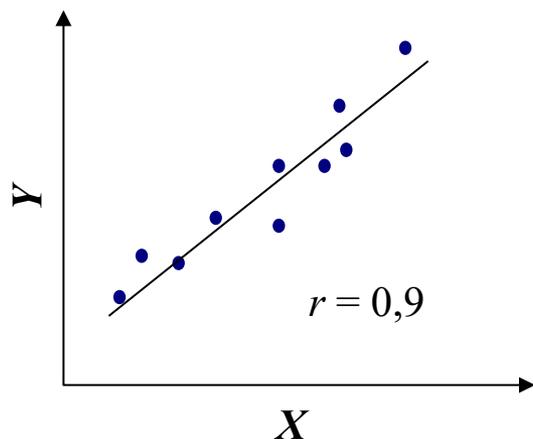
Covariância populacional	Covariância amostral
<b>v.a. discretas:</b> $\sigma_{XY} = \sum_{i=1}^N (x_i - \mu_x)(y_i - \mu_y)P(X = x_i; Y = y_i)$	$s_{XY} = \frac{\sum_{i=1}^n (x_i - \bar{X})(y_i - \bar{Y})}{n-1}$
<b>v.a. contínuas:</b> $\sigma_{XY} = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} (x - \mu_x)(y - \mu_y) f(x, y) dx dy$	



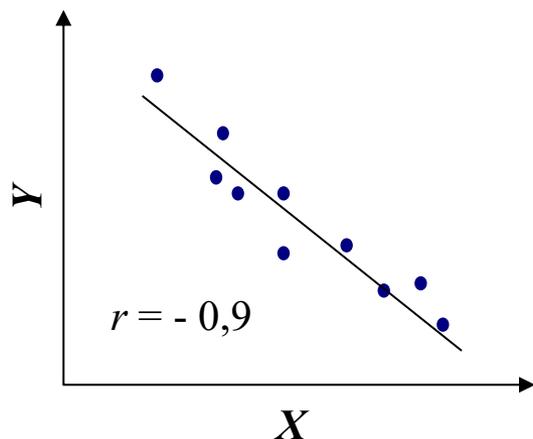
Quanto maior a covariância (em módulo), mais próximos estarão os pontos entorno da **reta** que representa a tendência principal da nuvem de pontos

Uma limitação da covariância é que seu valor calculado depende diretamente das unidades de medida, dificultando a comparação entre covariâncias.

# Coeficiente de Correlação



Coeficiente de Correlação (de Pearson)  
mede o grau de relação **linear** entre  $X$  e  $Y$



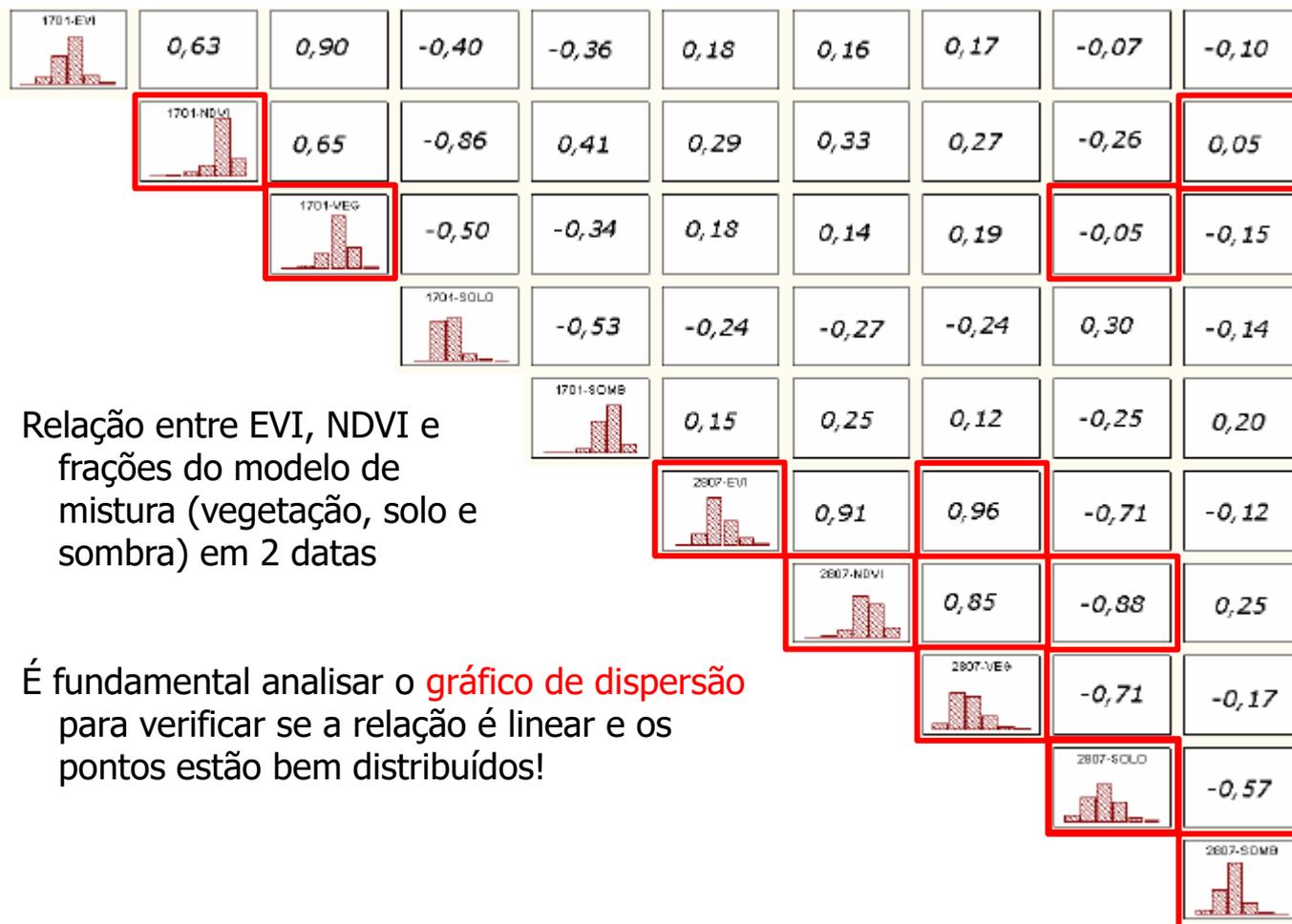
$$r = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}}$$

$-1 \leq r \leq 1$  (adimensional)

$$r = \frac{\sum_{i=1}^n (x_i - \bar{X})(y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{X})^2 \sum_{i=1}^n (y_i - \bar{Y})^2}}$$

$$= \frac{n \sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{\sqrt{\left[ n \sum_{i=1}^n x_i^2 - \left( \sum_{i=1}^n x_i \right)^2 \right] \left[ n \sum_{i=1}^n y_i^2 - \left( \sum_{i=1}^n y_i \right)^2 \right]}}$$

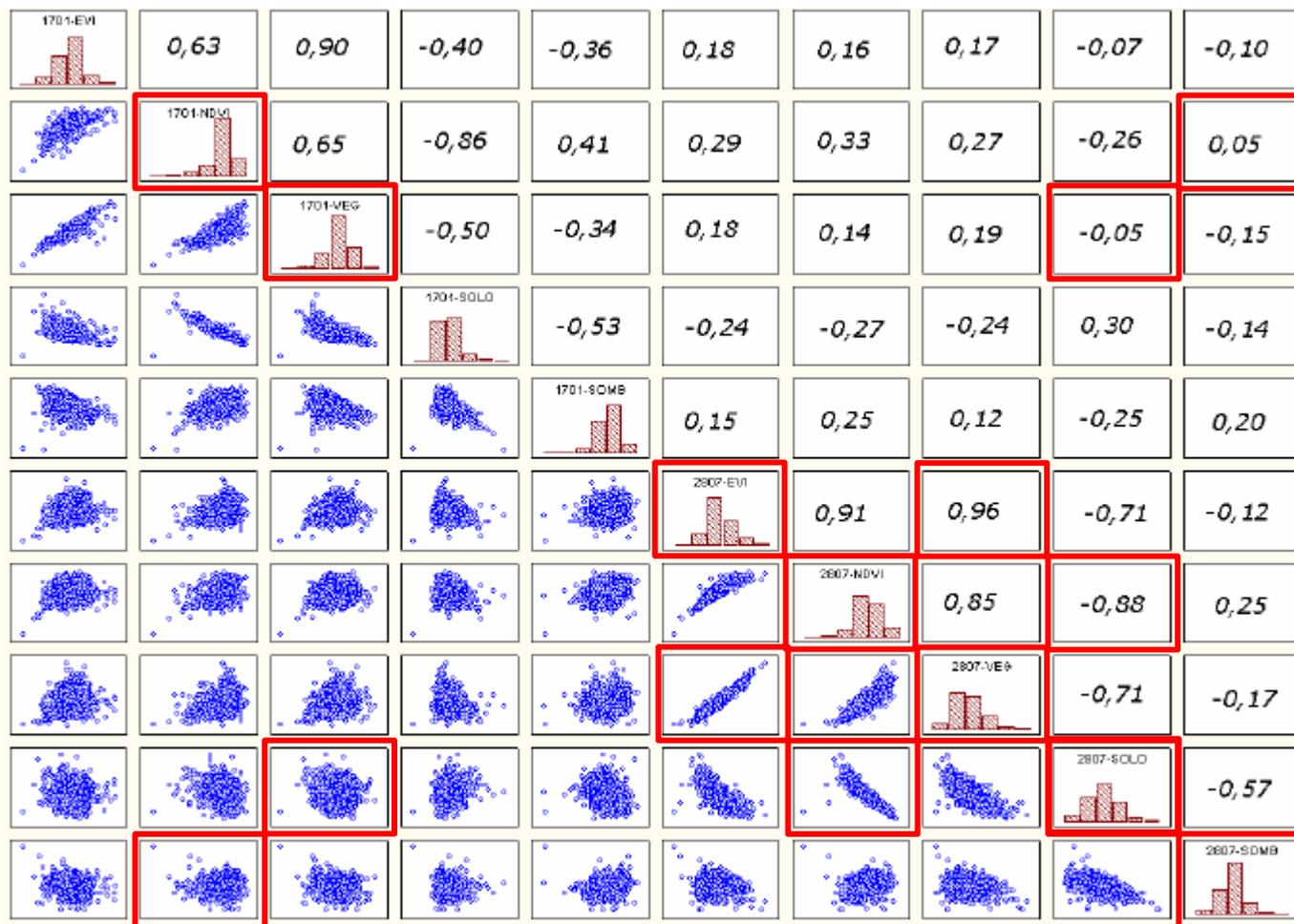
# Coeficiente de Correlação



Relação entre EVI, NDVI e frações do modelo de mistura (vegetação, solo e sombra) em 2 datas

É fundamental analisar o gráfico de dispersão para verificar se a relação é linear e os pontos estão bem distribuídos!

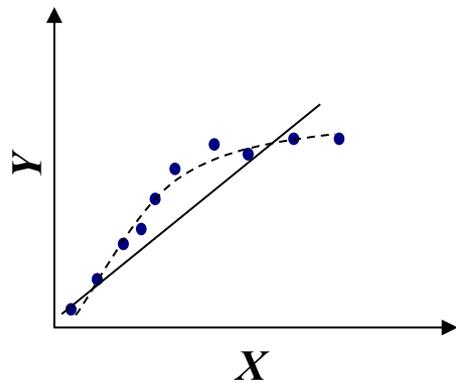
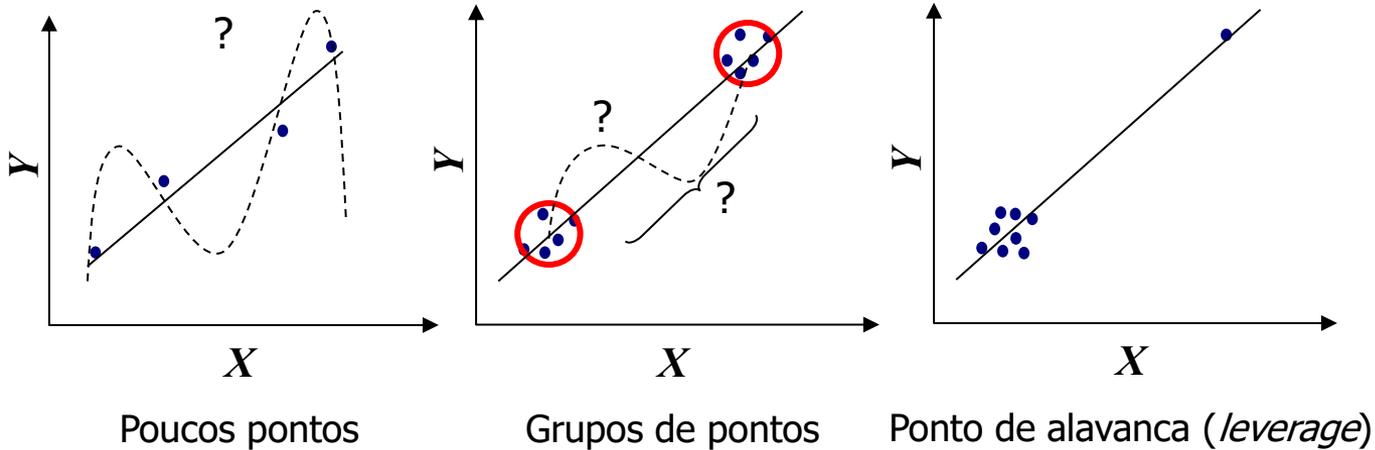
# Coeficiente de Correlação



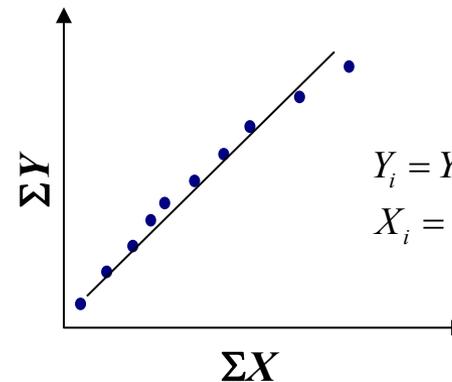
# Coeficiente de Correlação

## Interpretações errôneas do coeficiente de correlação

- Um alto coeficiente de correlação (em módulo) nem sempre indica que a equação de regressão estimada está bem ajustada aos dados.



Relação quase linear



Variáveis cumulativas

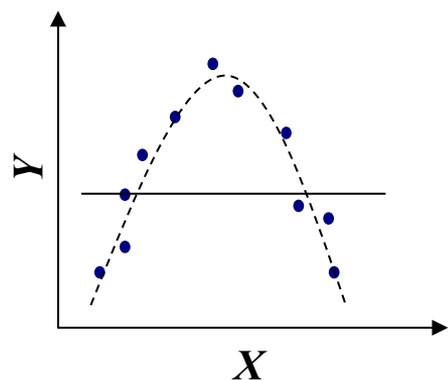
$$Y_i = Y_{i-1} + \Delta y_i \quad \Delta y_i \geq 0$$

$$X_i = X_{i-1} + \Delta x_i \quad \Delta x_i \geq 0$$

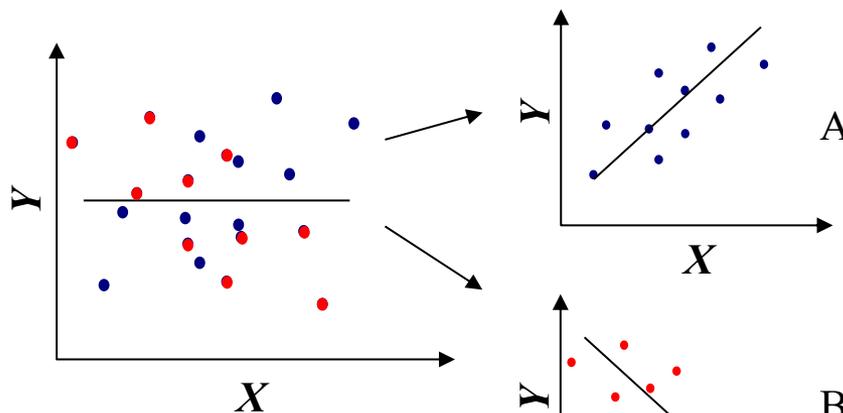
# Coeficiente de Correlação

## Interpretações errôneas do coeficiente de correlação

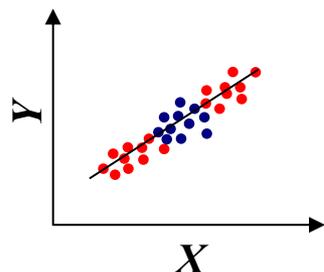
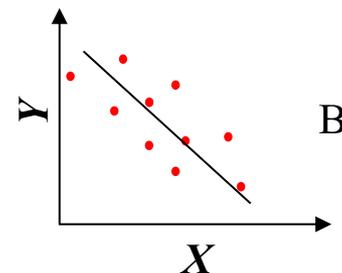
- Um coeficiente de correlação próximo de zero nem sempre indica que  $X$  e  $Y$  não são relacionadas.



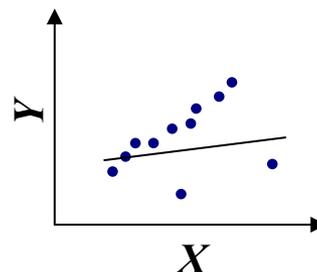
Relação não linear



Mistura de grupos com relações diferentes



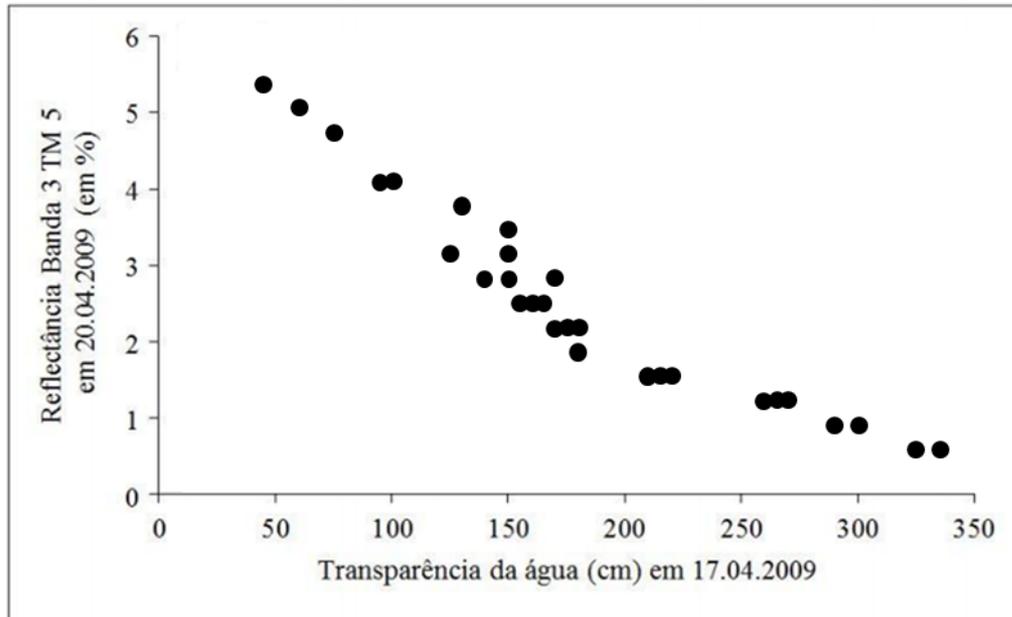
Amostragem não representativa



Presença de outliers ou pontos de alavanca

# Análise de Regressão

“Método estatístico que utiliza a relação entre duas ou mais variáveis para que uma variável possa ser estimada (ou predita) a partir da outra ou das outras”



A existência de uma relação estatística entre a variável **dependente**  $Y$  e a variável **independente**  $X$  não implica que  $Y$  realmente dependa de  $X$ , ou que exista uma relação de causa-efeito entre  $X$  e  $Y$ .

Fonte: Adaptado de Santos, F.C.; Pereira Filho, W.; Toniolo, G.R.. Transparência associada à reflectância da água do reservatório Passo Real. In: XVII SBSR, 2015. p. 6653-6659

# Análise de Regressão

---

Para que serve uma análise de regressão?

- Encontrar as variáveis mais relevantes que se relacionam com a variável dependente ( $Y$ )
- Encontrar a função que descreve como uma ou mais variáveis se relacionam com a variável dependente ( $Y$ ) e estimar os parâmetros que definem esta função (equação ajustada)
- Usar a equação ajustada para prever valores da variável dependente ( $Y$ )

## Regressão Linear Simples

# Modelo de Regressão Linear Simples

Pressuposições:

$$E(\varepsilon_i) = 0$$

$$Var(\varepsilon_i) = \sigma^2$$

$$Cov(\varepsilon_k, \varepsilon_j) = 0 \quad \forall k \neq j$$

→ erros independentes

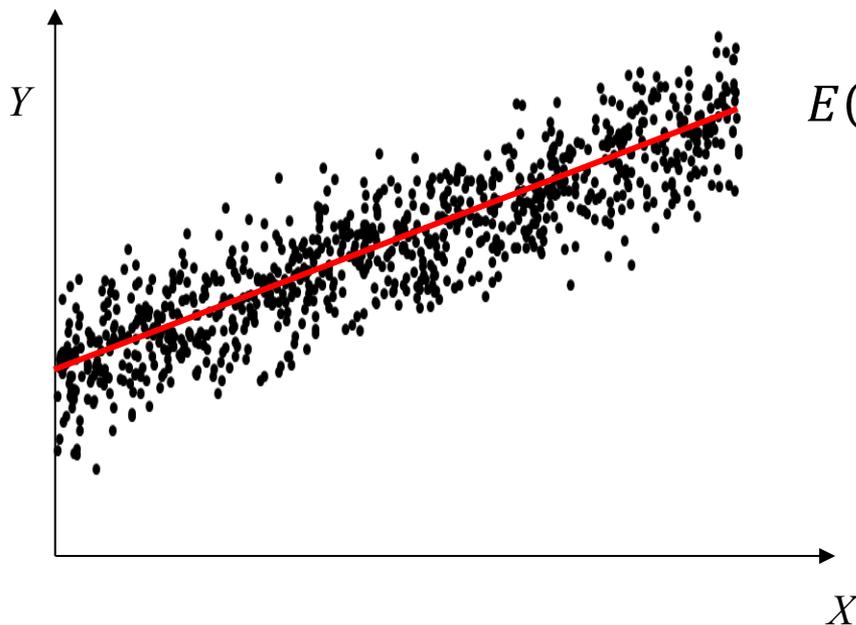
$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

→ componente aleatório  
(erro ou resíduo)

→ variável independente\*  
(variável explicativa)

→ variável dependente  
(variável resposta)

$\beta_0$  e  $\beta_1$  são parâmetros (fixos)

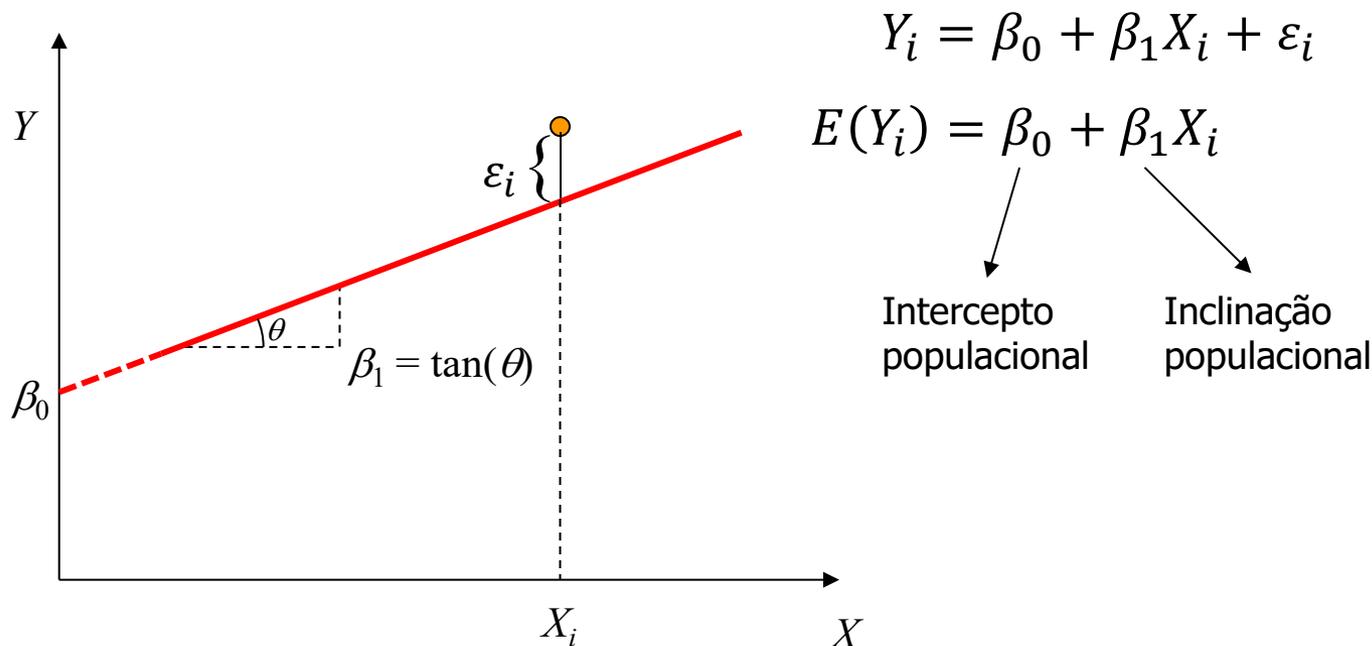


$$E(Y_i) = \beta_0 + \beta_1 X_i$$

A reta representa o valor médio da variável dependente ( $Y$ ) para todos os níveis da variável independente ( $X$ )

\* na regressão clássica, a **variável independente** não é considerada uma variável aleatória, ou seja, supõe-se que seus **valores são fixos conhecidos**

# Modelo de Regressão Linear Simples



$\beta_0$  representa o valor de  $E(Y_i)$  quando  $X_i = 0$

$\beta_1$  é o coeficiente angular da reta e representa o alteração em  $E(Y_i)$  quando  $X_i$  é incrementado em uma unidade

# Estimação dos parâmetros $\beta_0$ e $\beta_1$

Em geral não se conhece os valores de  $\beta_0$ ,  $\beta_1$  e  $\sigma^2$

Eles podem ser estimados através de dados obtidos por amostras

O método comumente utilizado na estimação dos parâmetros é o **método dos mínimos quadrados**, o qual considera os desvios quadráticos dos  $Y_i$  em relação a seu valor esperado:

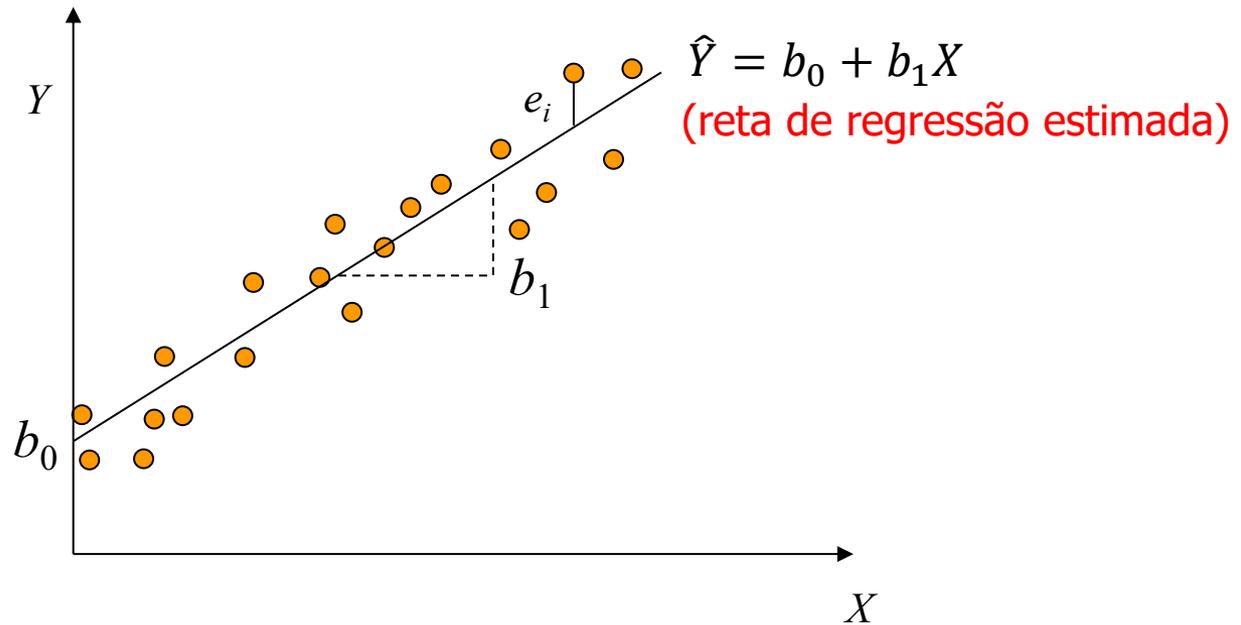
$$\varepsilon_i^2 = [Y_i - E(Y_i)]^2 \qquad \varepsilon_i^2 = [Y_i - (\beta_0 + \beta_1 X_i)]^2$$

Em particular, o método dos mínimos quadrados requer que consideremos a soma de  $n$  desvios quadrados, denotado por  $Q$ :

$$Q = \sum_{i=1}^n \varepsilon_i^2 = \sum_{i=1}^n [Y_i - (\beta_0 + \beta_1 X_i)]^2$$

De acordo com o método dos mínimos quadrados, os estimadores de  $\beta_0$  e  $\beta_1$  são aqueles, denotados por  $b_0$  e  $b_1$ , que tornam mínimo o valor de  $Q$ . Isso é feito derivando-se  $Q$  em relação a  $\beta_0$  e  $\beta_1$  e igualando-se as expressões encontradas a zero.

# Estimação dos parâmetros $\beta_0$ e $\beta_1$



$$b_1 = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2}$$

$$b_0 = \bar{Y} - b_1 \bar{X}$$

$b_0$  e  $b_1$  são v.a. (não independentes!) e portanto variam de amostra para amostra

$$E(b_0) = \beta_0$$

$$E(b_1) = \beta_1$$

$$e_i = Y_i - \hat{Y}_i \quad (\text{resíduo amostral})$$

# Estimação da Variância do Erro ( $\sigma^2$ )

A variância dos erros  $\varepsilon_i$ , denotada por  $\sigma^2$ , é um parâmetro do modelo de regressão, e necessita ser estimada através dos desvios quadráticos de  $Y_i$  em torno de sua própria média estimada  $\hat{Y}_i$ .

Soma dos quadrados dos erros ou resíduos (*SQE*):

$$SQE = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^n (Y_i - b_0 - b_1 X_i)^2 = \sum_{i=1}^n e_i^2$$

A soma dos quadrados dos erros tem  $n - 2$  graus de liberdade, pois 2 graus de liberdade foram perdidos por estimar  $\beta_0$  e  $\beta_1$ .

Portanto, o estimador de  $\sigma^2$ , denominado de **Quadrado Médio do Erro ou Resíduo (*QME*)**, é dado pela razão entre a *SQE* e  $n - 2$ :

$$QME = \frac{SQE}{n - 2}$$

Pode ser demonstrado que  $E[QME] = \sigma^2$

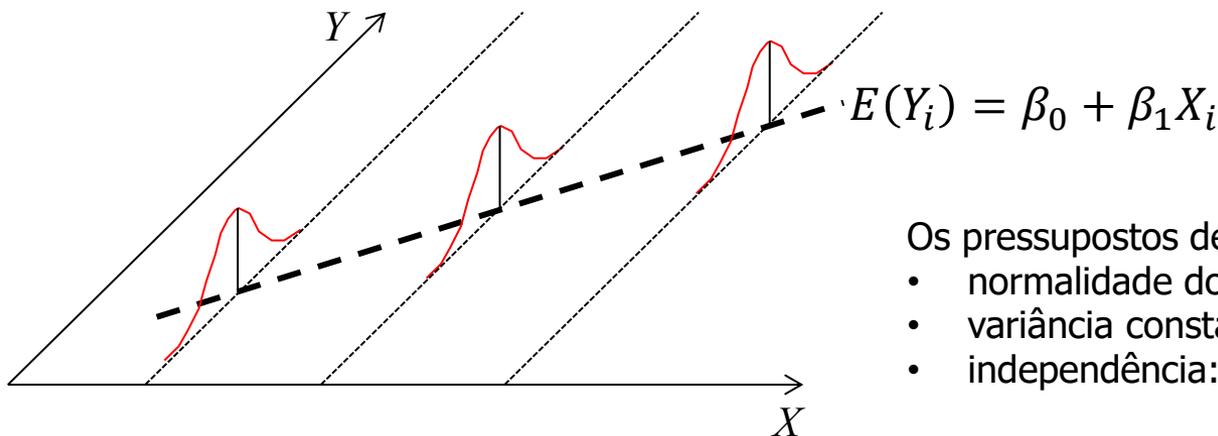
# Inferência em Análise de Regressão

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

Para a estimação dos parâmetros  $\beta_0$ ,  $\beta_1$  e  $\sigma^2$ , foi adotado o método dos mínimos quadrados e para tanto foi considerado que  $E(\varepsilon_i) = 0$ ,  $Var(\varepsilon_i) = \sigma^2$  e  $Cov(\varepsilon_k, \varepsilon_j) = 0 \quad \forall k \neq j$ . Note que nenhuma consideração foi feita a respeito da distribuição de  $\varepsilon_i$ .

Qualquer inferência que se faça a partir da equação ajustada utilizando-se uma amostra, deve pressupor a existência de uma distribuição associada a  $\varepsilon_i$ . A distribuição mais comum de ser adotada nesse caso é a **distribuição normal**, ou seja

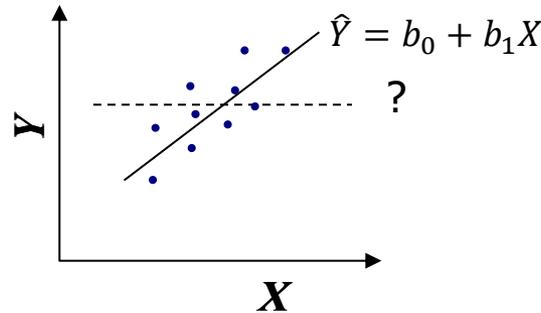
$$\varepsilon_i \sim N(0, \sigma^2) \text{ e } Cov(\varepsilon_k, \varepsilon_j) = 0 \quad \forall k \neq j$$



Os pressupostos devem ser verificados:

- normalidade dos erros: teste de Shapiro-Wilk
- variância constante: teste Breusch-Pagan
- independência: garantida pela amostragem

# Teste de Hipótese para $\beta_1$



$$E(Y_i) = \beta_0 ?$$

existe uma relação entre  $Y$  e  $X$ ?

$$t = \frac{b_1 - \beta_1}{s(b_1)} \sim t_{n-2}$$

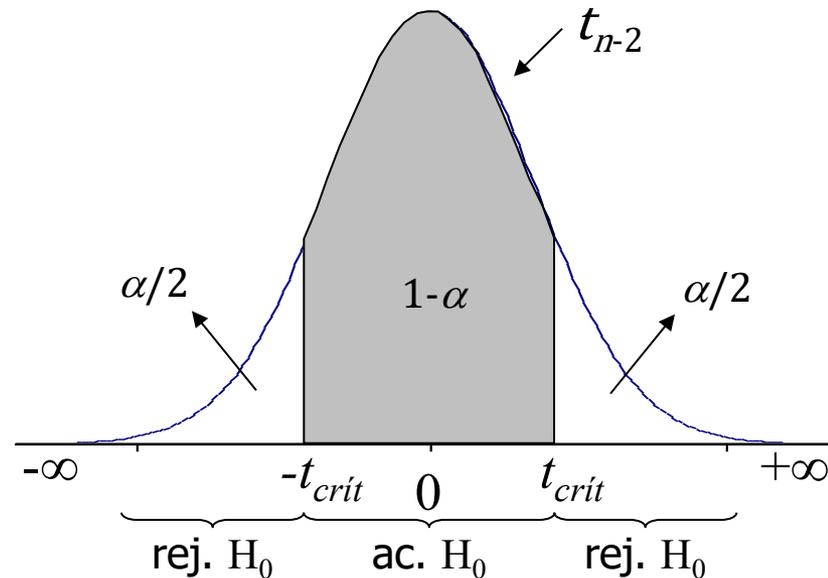
$$H_0: \beta_1 = 0$$

$$H_1: \beta_1 \neq 0$$

$$s^2(b_1) = \frac{QME}{\sum_{i=1}^n (X_i - \bar{X})^2}$$

se  $H_0$  verdadeira:

$$t = \frac{b_1}{s(b_1)} \sim t_{n-2}$$

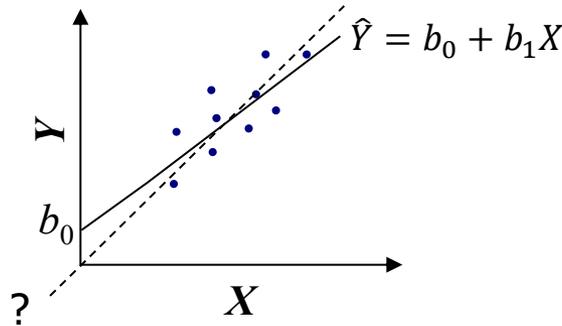


Região Crítica:

- aceito  $H_0$  se  $-t_{crit} < t < t_{crit} \rightarrow P(-t_{crit} < t < t_{crit}) = 1 - \alpha$
- rejeito  $H_0$  caso contrário  $\rightarrow P(|t| > t_{crit}) = \alpha$

OBS: se  $H_0$  for aceita, então a regressão não é significativa e, portanto, não há relação entre as variáveis  $X$  e  $Y$  ( $X$  e  $Y$  podem ser consideradas independentes).

# Teste de Hipótese para $\beta_0$



$$E(Y_i) = \beta_1 X_i ?$$

a regressão passa pela origem ?

$$t = \frac{b_0 - \beta_0}{s(b_0)} \sim t_{n-2}$$

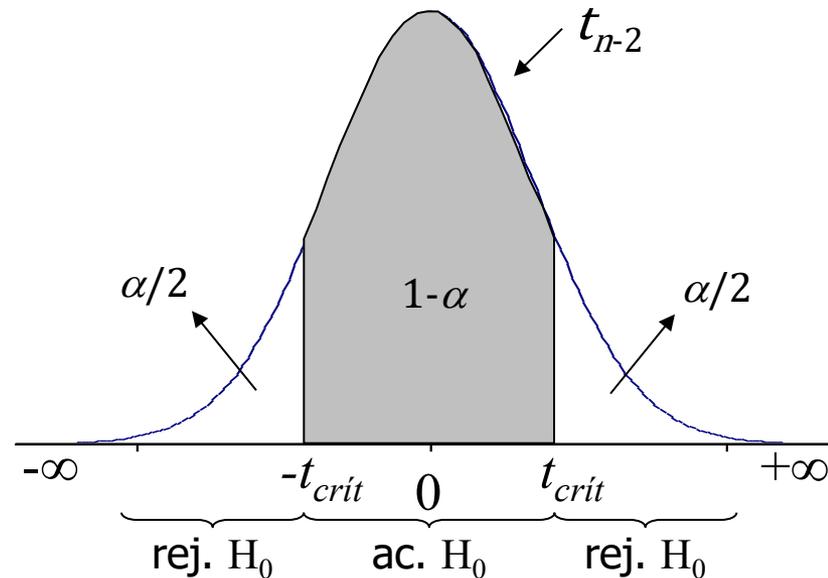
$$H_0: \beta_0 = 0$$

$$H_1: \beta_0 \neq 0$$

$$s^2(b_0) = QME \left[ \frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^n (X_i - \bar{X})^2} \right]$$

se  $H_0$  verdadeira:

$$t = \frac{b_0}{s(b_0)} \sim t_{n-2}$$



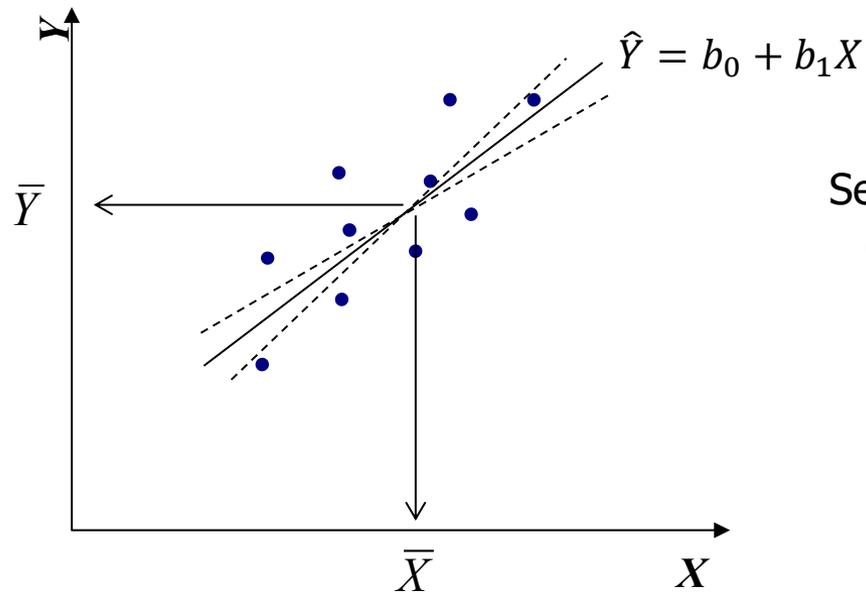
Região Crítica:

- aceito  $H_0$  se  $-t_{crit} < t < t_{crit}$   $\rightarrow P(-t_{crit} < t < t_{crit}) = 1 - \alpha$
- rejeito  $H_0$  caso contrário  $\rightarrow P(|t| > t_{crit}) = \alpha$

OBS: se  $H_0$  for aceita, então a reta de regressão passa pela origem. Isso não tem qualquer relação com a existência ou não de relação entre  $X$  e  $Y$ . Muitas vezes este teste é irrelevante (especialmente quando  $X=0$  não tem significado prático)

# Inferências para $E(Y_h)$

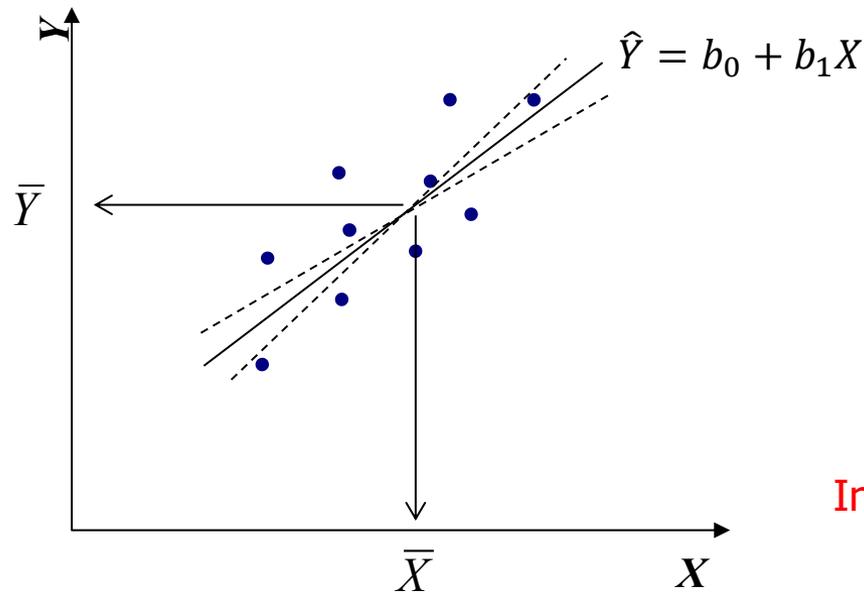
Considerando um determinado valor de  $X_h$ , quais as incertezas relacionadas às estimativas de  $E(Y_h)$ ?



Se  $b_0$  e  $b_1$  são variáveis aleatórias, então eles podem variar de amostra para amostra...

# Inferências para $E(Y_h)$

Considerando um determinado valor de  $X_h$ , quais as incertezas relacionadas às estimativas de  $E(Y_h)$ ?



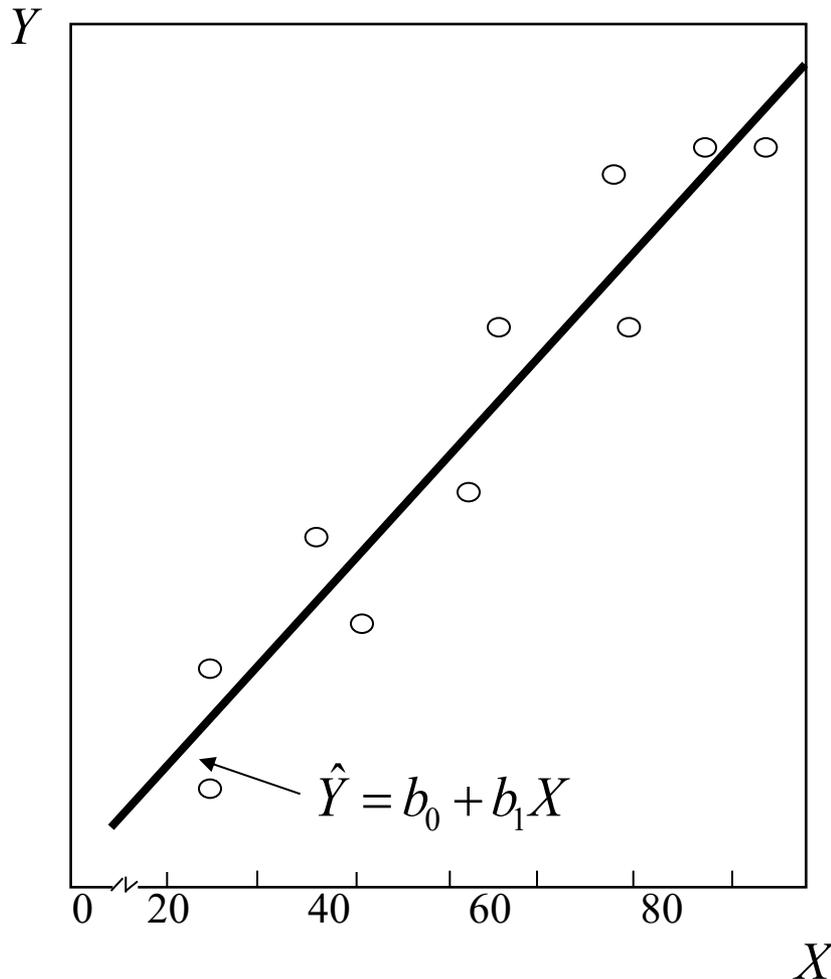
$$\frac{\hat{Y}_h - E(Y_h)}{s(\hat{Y}_h)} \sim t_{n-2}$$

$$s^2(\hat{Y}_h) = QME \left[ \frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum_{i=1}^n (X_i - \bar{X})^2} \right]$$

A red arrow points to the term  $(X_h - \bar{X})^2$  in the numerator of the fraction inside the brackets.

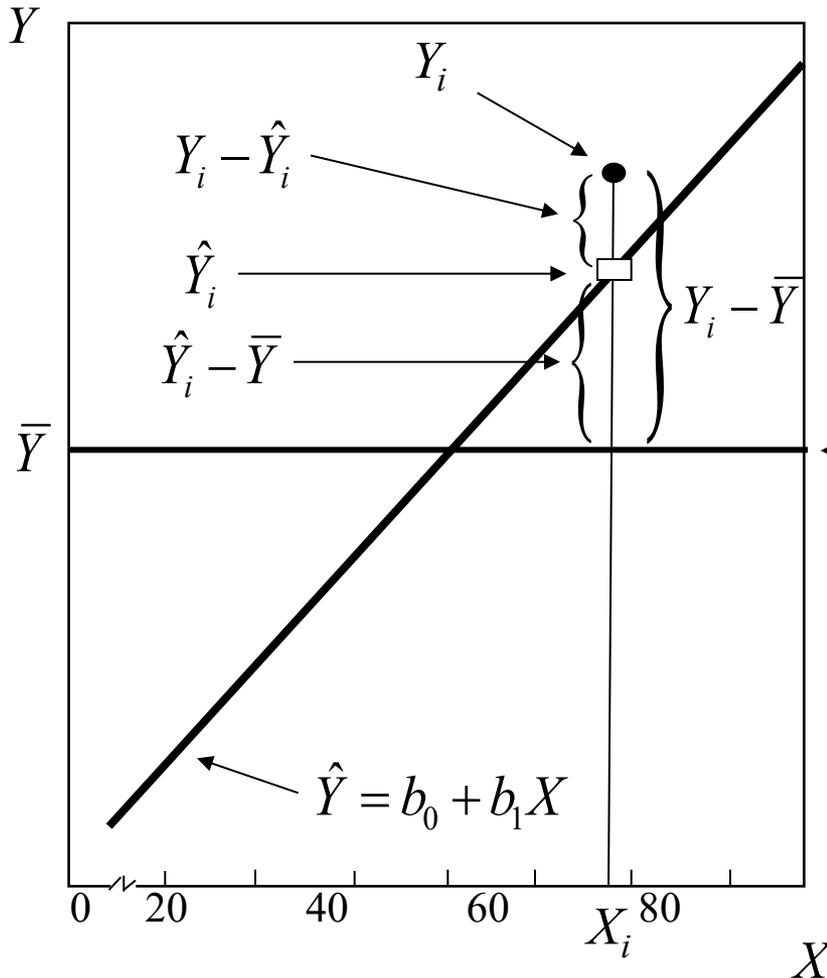
**Interpretação:** quanto mais distante  $X_h$  estiver de  $\bar{X}$ , maiores serão as incertezas nas estimativas de  $E(Y_h)$ . Por isso, extrapolações para faixa de valores de  $X$  extremos ou não observados devem ser evitadas!

# Particionamento do Erro



Assim como na Análise de Variância, na Análise de Regressão podemos analisar o erro (ou resíduo) sob diferentes aspectos...

# Particionamento do Erro



Assim como na Análise de Variância, na Análise de Regressão podemos analisar o erro (ou resíduo) sob diferentes aspectos...

← Caso não existisse relação entre  $X$  e  $Y$

$$Y_i - \bar{Y} = (\hat{Y}_i - \bar{Y}) + (Y_i - \hat{Y}_i)$$

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

$$SQTO = SQReg + SQE$$

# ANOVA x Análise de Regressão

Causas da Variação	Soma de Quadrados	Graus de Liberdade	Quadrados Médios
Regressão	$\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2$	1	$\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2$
Resíduo	$\sum_{i=1}^n (Y_i - \hat{Y}_i)^2$	$n - 2$	$\frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{n - 2}$
Total	$\sum_{i=1}^n (Y_i - \bar{Y})^2$	$n - 1$	

$$E(QMReg) = \sigma^2 + \beta_1^2 \sum_{i=1}^n (X_i - \bar{X})^2$$

$$E(QME) = \sigma^2$$



estimador tendencioso, exceto se  $\beta_1 = 0$

$$H_0: \beta_1 = 0$$

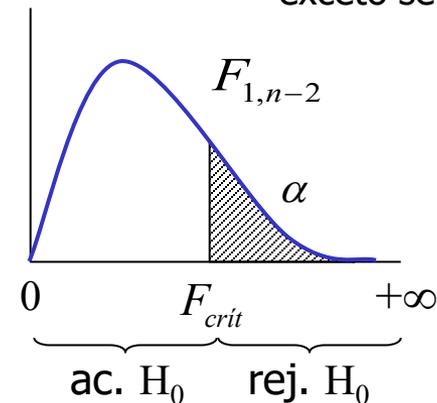
se  $H_0$  verdadeira:

$$H_1: \beta_1 \neq 0$$

$$F = \frac{QMReg}{QME} \sim F_{1, n-2}$$

Região Crítica:

- aceito  $H_0$  se  $F < F_{crit} \rightarrow P(F < F_{crit}) = 1 - \alpha$
- rejeito  $H_0$  caso contrário  $\rightarrow P(F > F_{crit}) = \alpha$



Se  $H_0$  for aceita, então a regressão não é significativa e, portanto, não há relação entre as variáveis  $X$  e  $Y$  ( $X$  e  $Y$  podem ser consideradas independentes).

**OBS:** Para regressão **linear simples**: teste  $F$  é equivalente ao teste  $t$  bilateral para  $\beta_1$

# Coeficiente de Determinação

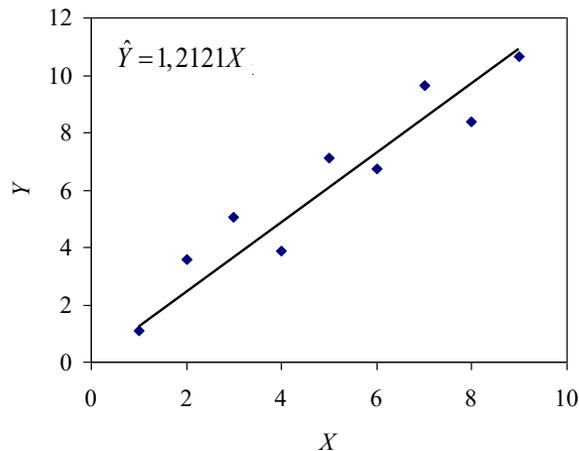
$$r^2 = \frac{SQReg}{SQTO} = \frac{SQTO - SQE}{SQTO} = 1 - \frac{SQE}{SQTO} \quad 0 \leq r^2 \leq 1$$

**Interpretação:**  $r^2$  mede a fração da variação total de  $Y$  explicada pela regressão e por isso pode ser representada em porcentagem

**OBS:** o coeficiente de determinação equivale ao quadrado do coeficiente de correlação para regressões **lineares simples**

**Atenção:**

Regressão passando pela origem ( $\beta_0 = 0$ )



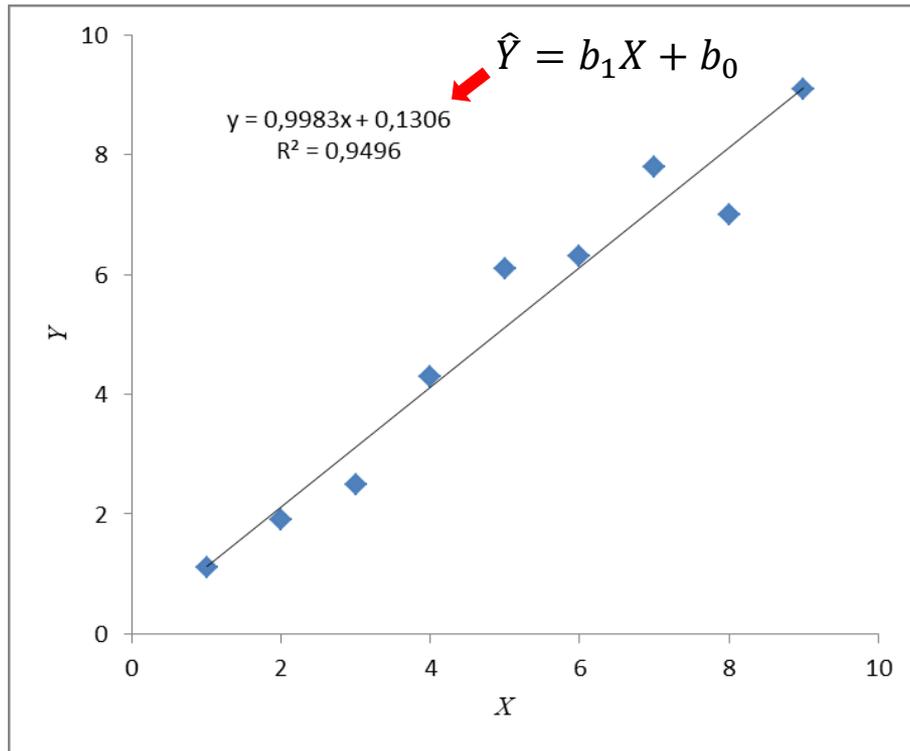
$$Y_i = \beta_1 X_i + \xi_i \quad b_1 = \frac{\sum_{i=1}^n X_i Y_i}{\sum_{i=1}^n X_i^2} \quad s^2(b_1) = \frac{QME}{\sum_{i=1}^n X_i^2}$$

~~$r^2 = 1 - SQE/SQTO$~~  ( $r^2$  pode ser negativo!)

$$r^2 = 1 - SQE/SQTO * \begin{cases} SQE = \sum_{i=1}^n (Y_i - b_1 X_i)^2 \\ SQTO^* = \sum_{i=1}^n Y_i^2 \end{cases}$$

# Análise de Regressão no EXCEL

$X$	$Y$
1	1,1
2	1,9
3	2,5
4	4,3
5	6,1
6	6,3
7	7,8
8	7,0
9	9,1



# Análise de Regressão no EXCEL

X	Y
1	1,1
2	1,9
3	2,5
4	4,3
5	6,1
6	6,3
7	7,8
8	7,0
9	9,1

The screenshot displays the Microsoft Excel interface with the 'Análise de dados' (Data Analysis) task pane open. The 'Regressão' (Regression) tool is selected. The 'Intervalo Y de entrada' (Input Y Range) is set to '\$C\$1:\$C\$10', the 'Intervalo X de entrada' (Input X Range) is '\$B\$1:\$B\$10', and the 'Intervalo de saída' (Output Range) is '\$E\$1'. The 'Rótulos' (Labels) checkbox is checked. The background shows a spreadsheet with columns X and Y highlighted in red.

**Ferramentas de análise**

Análise de Fourier  
Histograma  
Média móv.  
Geração de  
Ordem e p  
Regressão  
Amostragem  
Teste-T: d  
Teste-T: d

**Entrada**

Intervalo Y de entrada:

Intervalo X de entrada:

Rótulos  Constante é zero

Nível de confiança: 95 %

**Opções de saída**

Intervalo de saída:

Nova planilha:

Nova pasta de trabalho

**Resíduos**

Resíduos  Plotar resíduos

Resíduos padronizados  Plotar ajuste de linha

**Probabilidade normal**

Plotagem de probabilidade normal

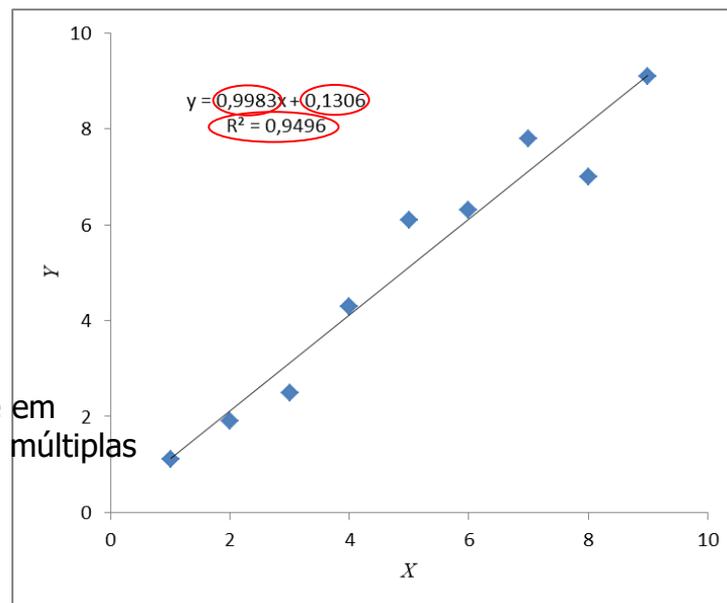
# Análise de Regressão no EXCEL

X	Y
1	1,1
2	1,9
3	2,5
4	4,3
5	6,1
6	6,3
7	7,8
8	7,0
9	9,1

## RESUMO DOS RESULTADOS

Estatística de regressão	
R múltiplo	0,9745
R-Quadrado	0,9496
R-quadrado ajustado	0,9424
Erro padrão	0,6735
Observações	9

importante em regressões múltiplas  $\sqrt{QME}$



## ANOVA

	gl	SQ	MQ	F	F de significação
Regressão	1	59,8002	59,8002	131,8267	8,55E-06
Resíduo	7	3,1754	0,4536		
Total	8	62,9756			

bilateral

	Coefficientes	Erro padrão	Stat t	valor-P	95% inferiores	95% superiores
Interseção	0,1306	0,4893	0,2668	0,7973	-1,0265	1,2876
X	0,9983	0,0870	11,4816	8,55E-06	0,7927	1,2039

OBS: Para regressão linear simples: teste  $F$  é equivalente ao teste  $t$  bilateral para  $\beta_1$

# Análise de Regressão no R

X	Y
1	1,1
2	1,9
3	2,5
4	4,3
5	6,1
6	6,3
7	7,8
8	7,0
9	9,1

```

>x <- c(1,2,3,4,5,6,7,8,9)
>y <- c(1.1,1.9,2.5,4.3,6.1,6.3,7.8,7.9,9.1)
>plot(x, y, xlim = c(1,9), ylim = c(1,10))
>reg <- lm(y ~ x)
>abline(reg)
>ypred <- predict(reg)
>summary(reg)
>anova(reg)
    
```

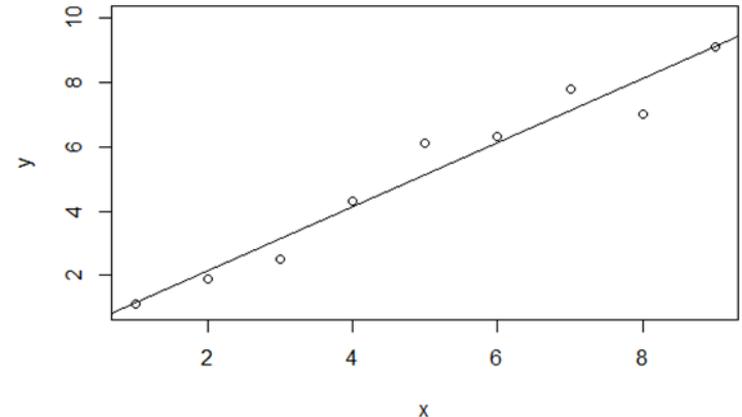
Call:  
lm(formula = y ~ x)

Residuals:  
Min 1Q Median 3Q Max  
-1.11722 -0.22722 -0.01556 0.17944 0.97778

Coefficients:  
Estimate Std. Error t value Pr(>|t|)  
(Intercept) 0.13056 0.48930 0.267 0.797  
x 0.99833 0.08695 11.482 8.55e-06 \*\*\*

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

Residual standard error: 0.61  $\sqrt{QME}$  of freedom  
Multiple R-squared: 0.9496,  $r^2$  Adjusted R-squared: 0.9424  
F-statistic: 131.8 on 1 and 7 DF, p-value: 8.547e-06 valor-P



Analysis of Variance Table  
Response: y

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
x	1	59.8	59.8	131.8	8.55e-06***
Residuals	7	3.175	0.454		

---  
*QME*

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

# Modelos Linearizáveis

Modelo Padrão:  $Y_i = \beta_0 + \beta_1 X_i + \xi_i$

exponencial

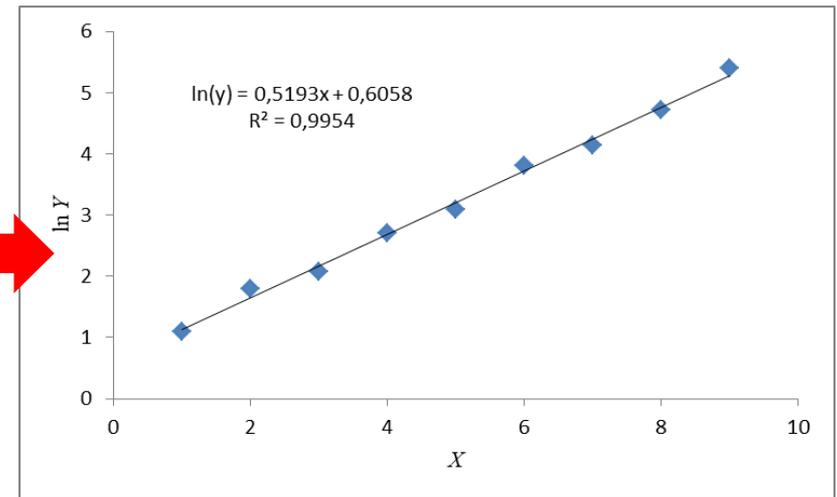
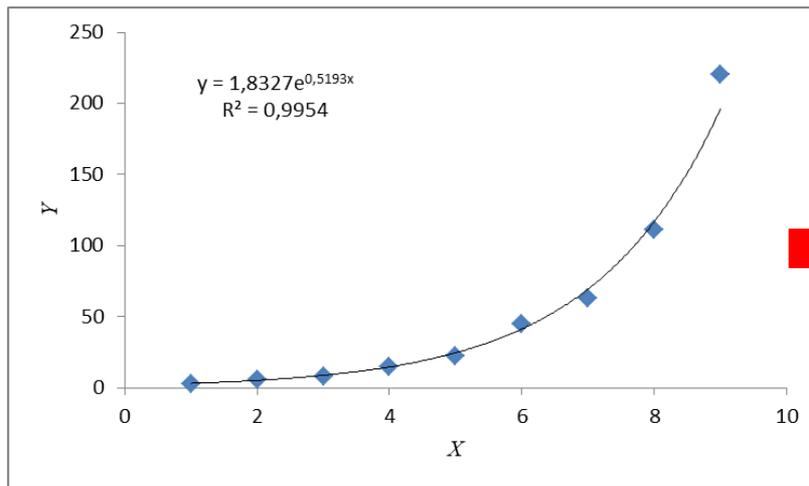
$$Y_i = \beta_0 e^{\beta_1 X_i} \xi_i$$

$$\ln Y_i = \ln \beta_0 + \beta_1 X_i + \ln \xi_i$$

$$Y'_i = \beta'_0 + \beta_1 X_i + \xi'_i$$

$$\xi'_i \sim N(0, \sigma^2)$$

exponencial



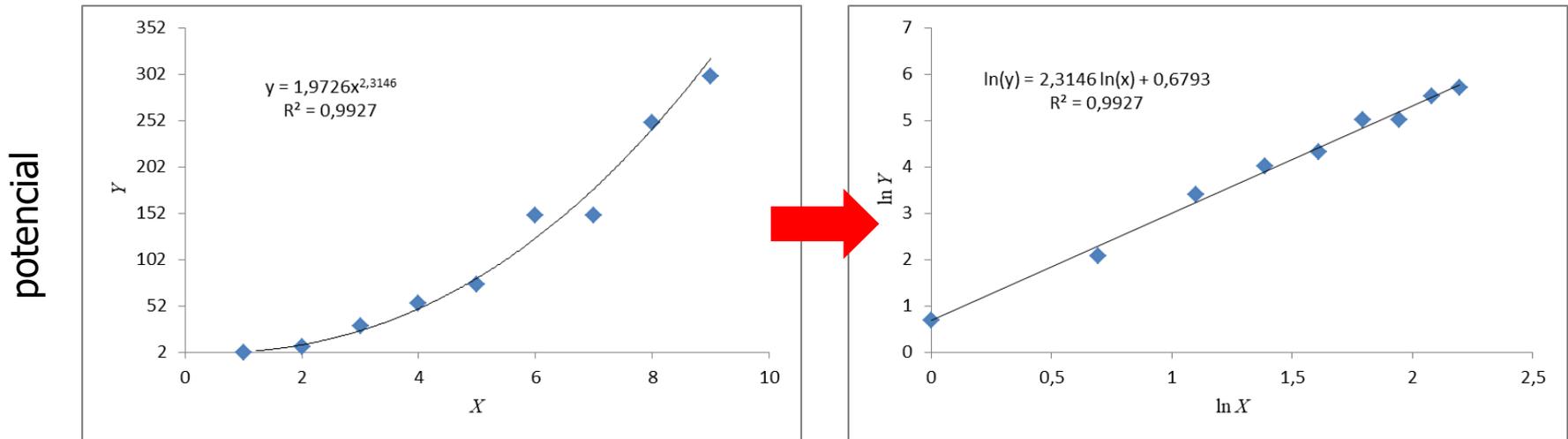
# Modelos Linearizáveis

Modelo Padrão:  $Y_i = \beta_0 + \beta_1 X_i + \xi_i$

potencial

$$Y_i = \beta_0 X_i^{\beta_1} \xi_i \quad \ln Y_i = \ln \beta_0 + \beta_1 \ln X_i + \ln \xi_i \quad Y'_i = \beta'_0 + \beta_1 X'_i + \xi'_i$$

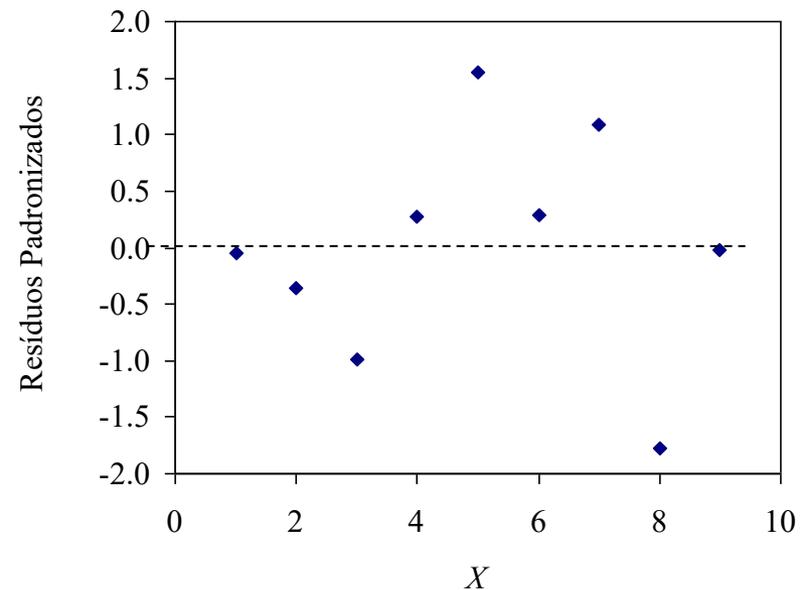
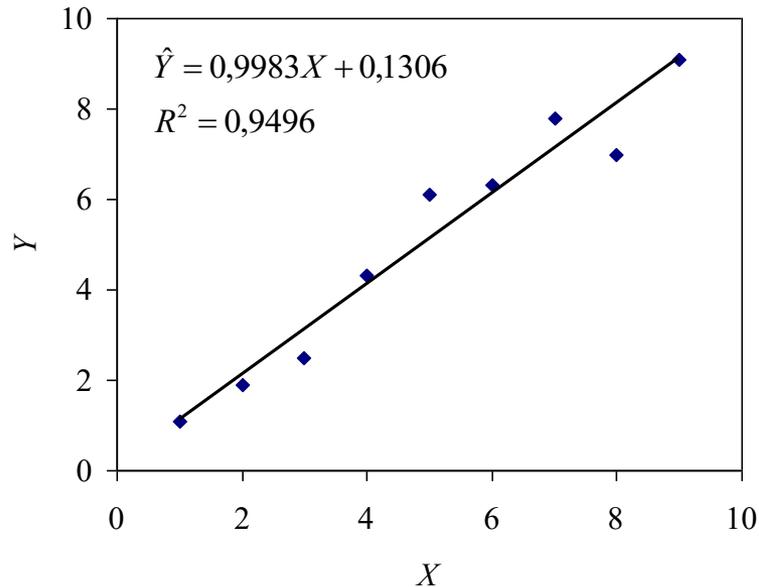
$$\xi'_i \sim N(0, \sigma^2)$$



A adoção de uma série de transformações (logaritmo, potência, inverso, etc) em  $X$ ,  $Y$  ou ambos permite que o modelo linear simples possa ser utilizado para representar relações mais complexas  $\Rightarrow$  **Modelos Lineares Generalizados**

# Análise de Resíduos

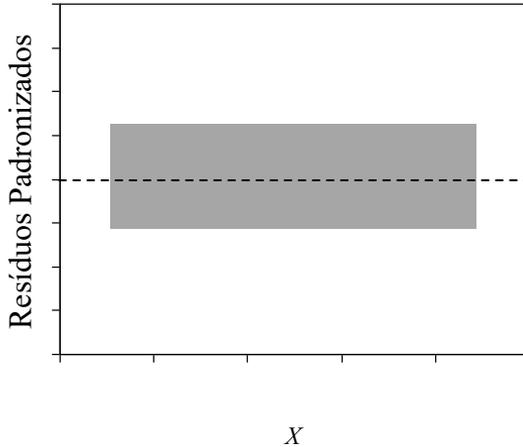
É uma etapa fundamental na Análise de Regressão pois auxilia na identificação de problemas que afetam diretamente a interpretação dos resultados



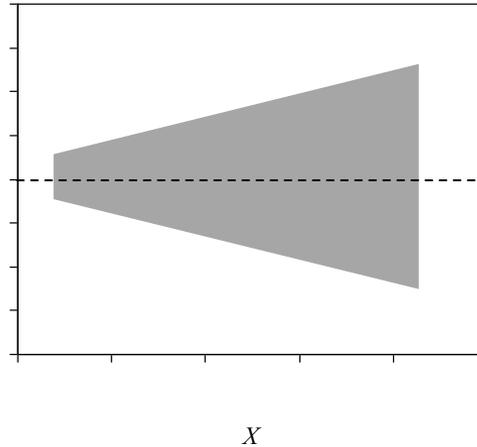
$$\text{Resíduo Padronizado} = e_i / \sqrt{QME}$$

# Análise de Resíduos

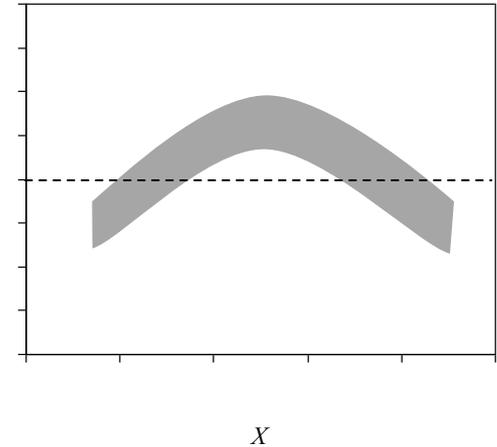
"ideal"



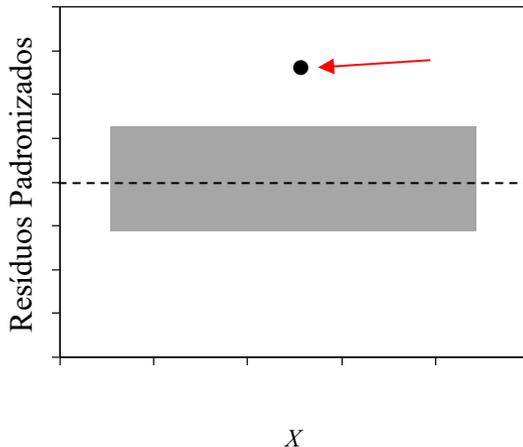
$\sigma^2$  não constante



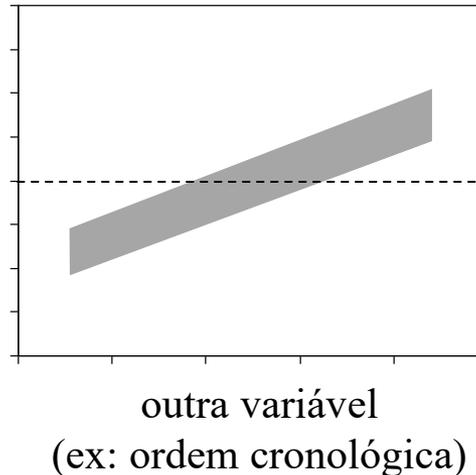
não linearidade



"outlier"



não independência



Essa análise qualitativa pode antecipar muitos problemas e indicar a necessidade de adequação do modelo a ser utilizado e/ou a necessidade de retirada ou adição de amostras

Não exclui a utilização de testes e análises específicas!

# Modelo de Regressão Linear Múltipla

## Modelo Geral

$$Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \dots + \beta_{p-1} X_{p-1,i} + \xi_i$$

$\beta_0, \beta_1, \beta_2, \dots, \beta_{p-1}$  são parâmetros do modelo ( $p$  parâmetros no total)

$X_{1,i}, X_{2,i}, \dots, X_{p-1,i}$  são valores fixos conhecidos

$\xi_i$  são erros independentes  $\xi_i \sim N(0, \sigma^2)$

$i = 1, 2, \dots, n$  (cada amostra deve conter valores para todos  $X_k$ )

Fazendo  $X_{0,i} = 1$ , podemos reescrever o modelo como

$$\begin{aligned} Y_i &= \beta_0 X_{0,i} + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \dots + \beta_{p-1} X_{p-1,i} + \xi_i \\ &= \sum_{k=0}^{p-1} \beta_k X_{k,i} + \xi_i \end{aligned}$$

# Casos Especiais

## Regressão Polinomial

Considere um modelo de regressão de 3º grau com uma variável independente:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \beta_3 X_i^3 + \xi_i$$

Se considerarmos  $X_{1,i} = X_i$ ,  $X_{2,i} = X_i^2$  e  $X_{3,i} = X_i^3$  então

$$Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \beta_3 X_{3,i} + \xi_i$$

## Efeito de Interação

Considere um modelo de regressão com duas variáveis independentes:

$$Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \beta_3 X_{1,i} X_{2,i} + \xi_i$$

Se considerarmos  $X_{3,i} = X_{1,i} X_{2,i}$  então

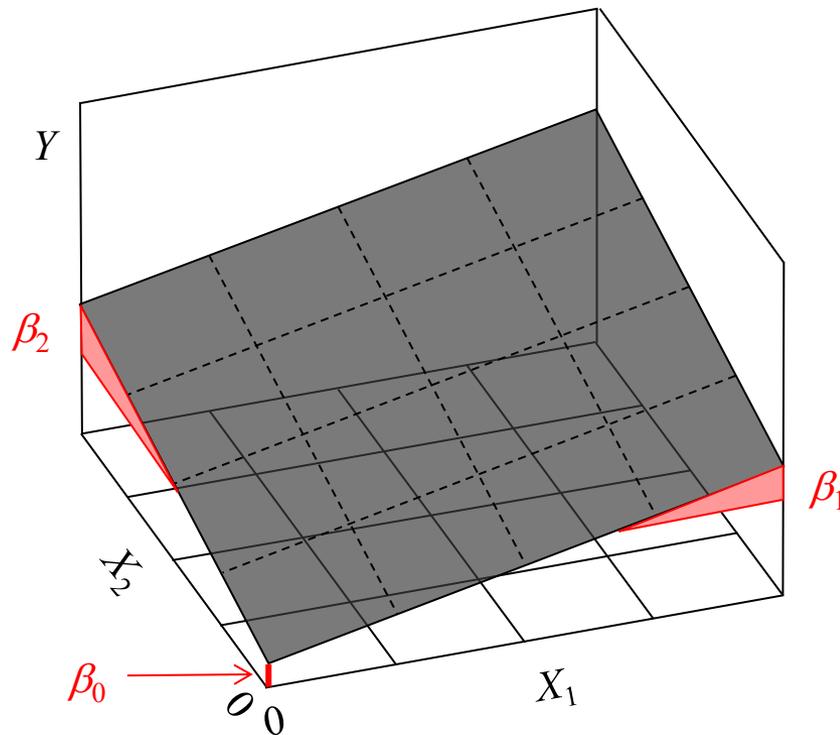
$$Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \beta_3 X_{3,i} + \xi_i$$

**Importante:** o modelo geral de regressão linear não é restrito às superfícies planas. O termo **linear** refere-se ao fato de que ele é **linear nos parâmetros**, não na forma da superfície.

# Modelo de Regressão Linear Múltipla

Exemplo: duas variáveis independentes

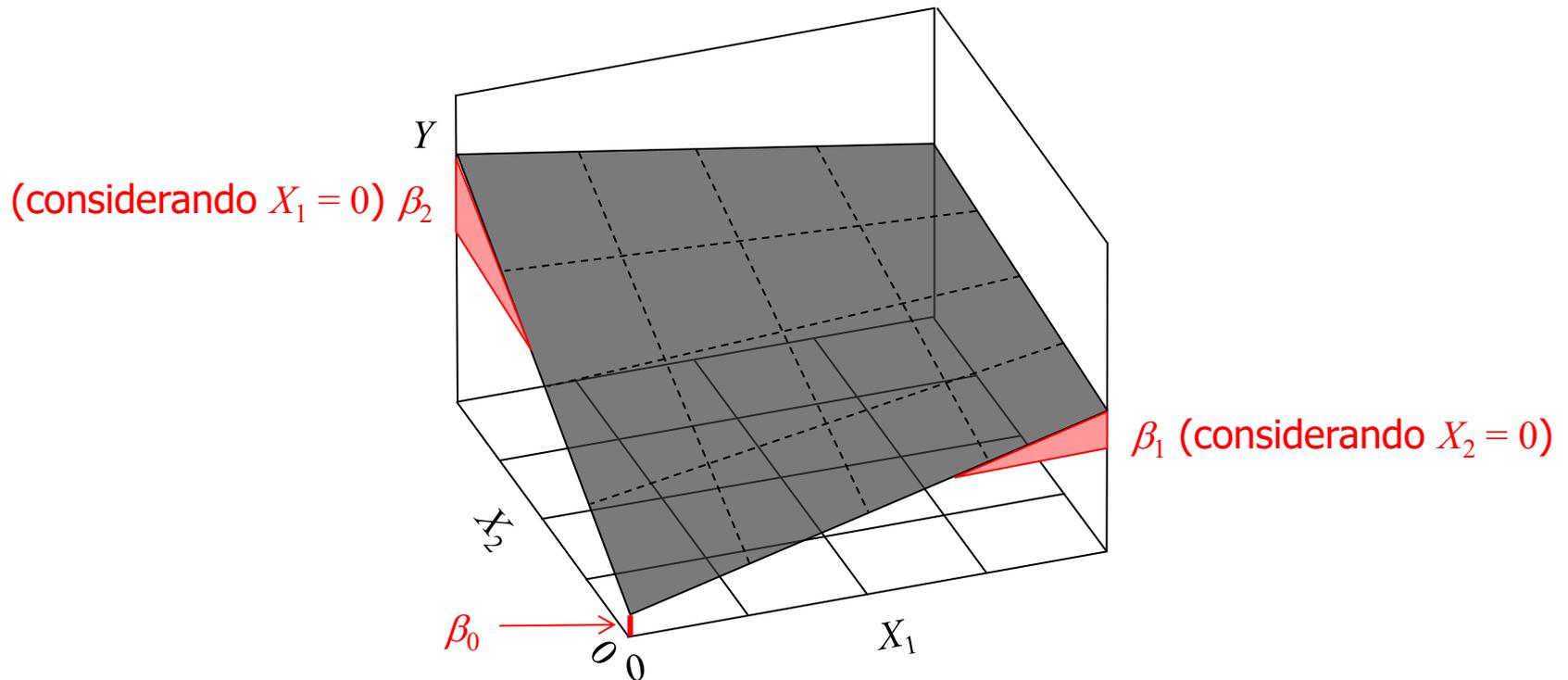
$$Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \xi_i$$



# Modelo de Regressão Linear Múltipla

Exemplo: duas variáveis independentes com interação

$$Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \beta_3 X_{1,i} X_{2,i} + \xi_i$$



# Notação Matricial

Modelo Geral

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\xi}$$

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} \quad \mathbf{X} = \begin{bmatrix} 1 & X_{1,1} & X_{2,1} & \cdots & X_{p-1,1} \\ 1 & X_{1,2} & X_{2,2} & \cdots & X_{p-1,2} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & X_{1,n} & X_{2,n} & \cdots & X_{p-1,n} \end{bmatrix} \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{bmatrix} \quad \boldsymbol{\xi} = \begin{bmatrix} \xi_1 \\ \xi_2 \\ \vdots \\ \xi_n \end{bmatrix}$$

$n \times 1$                        $n \times p$                        $p \times 1$                        $n \times 1$

→ primeira coluna de 1's para representar o  $\beta_0$

$$\mathbf{b} = \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_{p-1} \end{bmatrix} = (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{Y}) \quad \hat{\mathbf{Y}} = \mathbf{X}\mathbf{b} \quad \mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}}$$

# ANOVA x Análise de Regressão

Causas da Variação	Soma de Quadrados	Graus de Liberdade	Quadrados Médios
Regressão	$SQ_{TO} - SQE$	$p - 1$	$\frac{SQ_{Reg}}{p - 1}$
Resíduo	$Y'Y - b'X'Y$	$n - p$	$\frac{SQE}{n - p}$
Total	$\sum_{i=1}^n (Y_i - \bar{Y})^2$	$n - 1$	

$$Y_i = \beta_0 + \beta_1 X_{1,i} + \dots + \beta_{p-1} X_{p-1,i} + \xi_i$$

$$E(QM_{Reg}) \geq \sigma^2$$

$$E(QME) = \sigma^2$$

$H_0: \beta_k = 0 \ (k = 1, \dots, p - 1)$

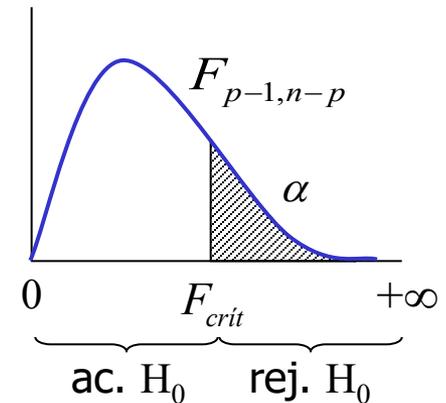
$H_1: \text{pelo menos um dos } \beta_k \neq 0$

se  $H_0$  verdadeira:

$$F = \frac{QM_{Reg}}{QME} \sim F_{p-1, n-p}$$

Região Crítica:

- aceito  $H_0$  se  $F < F_{crit} \rightarrow P(F < F_{crit}) = 1 - \alpha$
- rejeito  $H_0$  caso contrário  $\rightarrow P(F > F_{crit}) = \alpha$



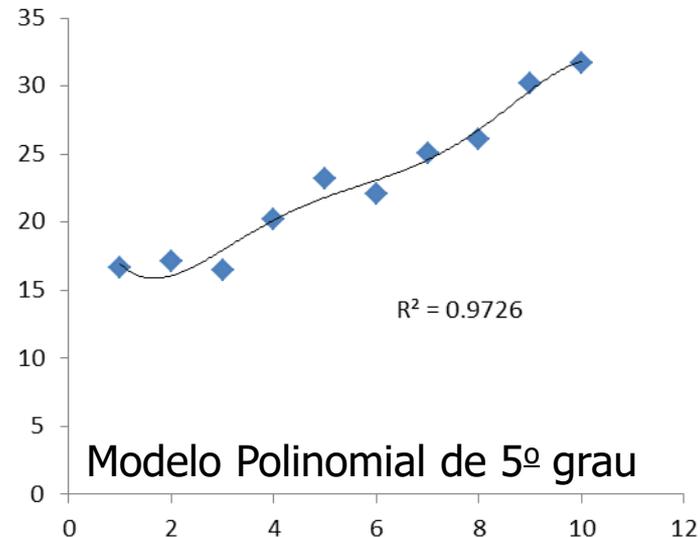
**OBS:** o coeficiente  $\beta_0$  não tem nenhuma influência sobre este teste  
o teste F não é capaz de identificar qual ou quais  $\beta_k$  são diferentes de zero, nem quais  $\beta_k$  são diferentes entre si

# Coeficiente de Determinação Múltiplo

$$SQTO = SQReg + SQE$$

$$r^2 = \frac{SQReg}{SQTO}$$
$$= 1 - \frac{SQE}{SQTO}$$

**Atenção:**  $r^2$  é fortemente influenciado pelo número de parâmetros considerados no modelo.  
Quanto maior o número de parâmetros ( $p \rightarrow n$ ), melhor o ajuste e portanto maior o  $r^2$ .  
Quando  $p = n$ , o ajuste é perfeito!!!



# Coeficiente de Determinação Múltiplo

$$SQTO = SQReg + SQE$$

$$r^2 = \frac{SQReg}{SQTO}$$
$$= 1 - \frac{SQE}{SQTO}$$

**Atenção:**  $r^2$  é fortemente influenciado pelo número de parâmetros considerados no modelo.

Quanto maior o número de parâmetros ( $p \rightarrow n$ ), melhor o ajuste e portanto maior o  $r^2$ .

Quando  $p = n$ , o ajuste é perfeito!!!

$$r_a^2 = 1 - \left( \frac{n-1}{n-p} \right) \frac{SQE}{SQTO}$$

## Coeficiente de Determinação Múltiplo Ajustado

Este coeficiente pode até diminuir se as variáveis acrescentadas ao modelo não representarem contribuições importantes.

# Teste de Hipótese para $\beta_k$

$$t = \frac{b_k - \beta_k}{s(b_k)} \sim t_{n-p}$$

$$s^2(\mathbf{b}) = \begin{bmatrix} s^2(b_0) & s(b_0, b_1) & \dots & s(b_0, b_{p-1}) \\ s(b_0, b_1) & s^2(b_1) & \dots & s(b_1, b_{p-1}) \\ \vdots & \vdots & \ddots & \vdots \\ s(b_0, b_{p-1}) & s(b_1, b_{p-1}) & \dots & s^2(b_{p-1}) \end{bmatrix} = QME(\mathbf{X}'\mathbf{X})^{-1}$$

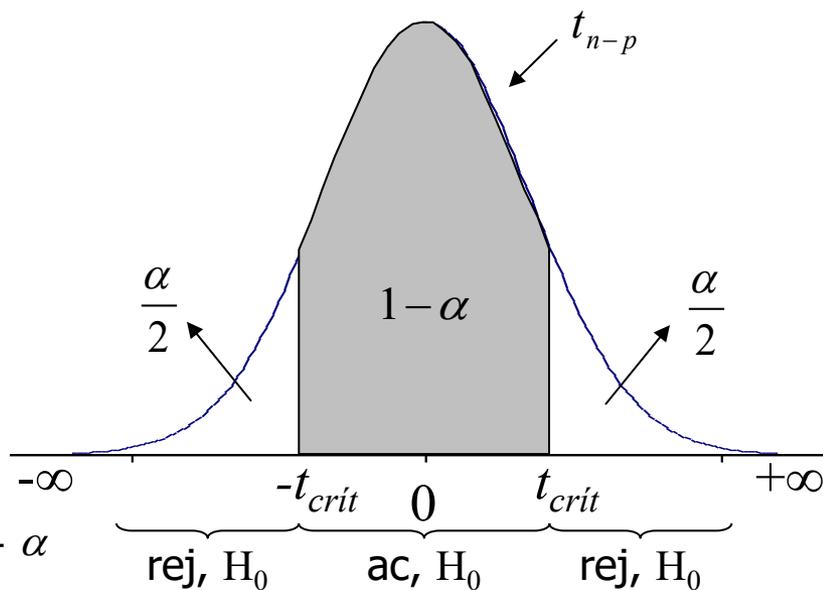
$$H_0 : \beta_k = 0 \quad E(Y_i) = \sum_{\substack{j=0 \\ j \neq k}}^{p-1} \beta_j X_{j,i}$$

$$H_1 : \beta_k \neq 0$$

se  $H_0$  verdadeira:

$$t = \frac{b_k}{s(b_k)} \sim t_{n-p}$$

Todos os outros  $\beta_j$   
estão no modelo,  
menos o  $\beta_k$



Região Crítica:

- aceito  $H_0$  se  $-t_{crit} < t < t_{crit} \rightarrow P(-t_{crit} < t < t_{crit}) = 1 - \alpha$
- rejeito  $H_0$  caso contrário  $\rightarrow P(|t| > t_{crit}) = \alpha$

OBS: se  $H_0$  for aceita, então  $\beta_k = 0$  e, portanto, a variável  $X_k$  não contribui significativamente para explicar  $Y$  (considerando que **todas** as demais variáveis independentes estejam presentes no modelo).

# Eliminando-se variáveis independentes

Y	X <sub>1</sub>	X <sub>2</sub>	X <sub>3</sub>	X <sub>4</sub>
11,70	126,92	174,56	226,69	364,26
16,34	75,02	129,40	117,43	329,68
16,76	51,00	106,17	75,41	592,57
16,83	47,75	110,50	66,58	471,11
22,02	145,83	148,78	258,84	1151,11
23,43	62,91	113,04	99,85	327,56
24,75	73,34	97,81	117,23	850,26
29,96	79,87	92,83	126,21	695,32
30,31	131,55	139,24	235,10	820,23
33,51	163,68	141,01	294,77	884,83
38,12	93,25	98,44	152,29	291,09
38,42	110,57	99,38	195,38	1162,36
40,63	93,28	88,63	159,74	338,08
46,15	196,54	140,37	363,28	508,84
47,98	184,33	128,83	334,06	764,28
54,58	119,84	71,83	204,97	709,91
58,22	163,02	102,36	295,87	626,23
66,27	155,43	84,14	284,87	50,34
86,27	273,91	109,00	514,30	620,11
89,29	212,29	53,56	392,89	1186,30

$$Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \beta_3 X_{3,i} + \beta_4 X_{4,i} + \xi_i$$

ANOVA

	gl	SQ	MQ	F	valor-P
Regressão	4	9354,57	2338,64	587,45	2,78E-16
Resíduo	15	59,71	3,98		
Total	19	9414,28			

altamente significativo

	Erro			
	Coefficientes	padrão	Stat t	valor-P
Interseção	64,4359	4,8424	13,3067	1,04E-09
X <sub>1</sub>	-0,2129	0,3081	-0,6908	0,5002
X <sub>2</sub>	-0,4741	0,0160	-29,5575	1,04E-14
X <sub>3</sub>	0,2659	0,1553	1,7123	0,1074
X <sub>4</sub>	-0,0075	0,0015	-4,8827	0,0002

não significativos a 5%

~~Conclusão:  $\beta_1 = 0$  e  $\beta_3 = 0$  ?~~

**Atenção:** não se pode considerar que todos os  $\beta_{k,i}$  cujas estatísticas t são não significativas, sejam simultaneamente iguais a zero!

Este problema pode ocorrer quando as variáveis independentes são correlacionadas (problema de colinearidade)

# Teste de Hipótese para múltiplos $\beta_k$

Considere um modelo **completo** dado por:

$$Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \beta_3 X_{3,i} + \beta_4 X_{4,i} + \xi_i$$

Suponha que se queira testar as hipóteses

$$H_0 : \beta_1 = \beta_3 = 0$$

$$H_1 : \beta_1 \neq 0 \text{ e/ou } \beta_3 \neq 0$$

Se  $H_0$  for verdadeiro então, o modelo é **reduzido** para:

$$Y_i = \beta_0 + \beta_2 X_{2,i} + \beta_4 X_{4,i} + \xi_i$$

Neste caso:

$$F = \frac{SQE_R - SQE_C}{p_{C-R}} \div \frac{SQE_C}{n-p} \sim F_{p_{C-R}, n-p} \qquad F = \frac{r_C^2 - r_R^2}{p_{C-R}} \div \frac{1 - r_C^2}{n-p}$$

onde  $p_{C-R}$  é o número de parâmetros testados em  $H_0$ , ou seja, o número de parâmetros ausentes no modelo reduzido

# Teste de Hipótese para múltiplos $\beta_k$

Y	X <sub>1</sub>	X <sub>2</sub>	X <sub>3</sub>	X <sub>4</sub>
11,70	126,92	174,56	226,69	364,26
16,34	75,02	129,40	117,43	329,68
16,76	51,00	106,17	75,41	592,57
16,83	47,75	110,50	66,58	471,11
22,02	145,83	148,78	258,84	1151,11
23,43	62,91	113,04	99,85	327,56
24,75	73,34	97,81	117,23	850,26
29,96	79,87	92,83	126,21	695,32
30,31	131,55	139,24	235,10	820,23
33,51	163,68	141,01	294,77	884,83
38,12	93,25	98,44	152,29	291,09
38,42	110,57	99,38	195,38	1162,36
40,63	93,28	88,63	159,74	338,08
46,15	196,54	140,37	363,28	508,84
47,98	184,33	128,83	334,06	764,28
54,58	119,84	71,83	204,97	709,91
58,22	163,02	102,36	295,87	626,23
66,27	155,43	84,14	284,87	50,34
86,27	273,91	109,00	514,30	620,11
89,29	212,29	53,56	392,89	1186,30

$$Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \beta_3 X_{3,i} + \beta_4 X_{4,i} + \xi_i$$

ANOVA

	gl	SQ	MQ	F	valor-P
Regressão	4	9354,57	2338,64	587,45	2,78E-16
Resíduo	15	59,71	3,98		
Total	19	9414,28			

$$Y_i = \beta_0 + \beta_2 X_{2,i} + \beta_4 X_{4,i} + \xi_i$$

ANOVA

	gl	SQ	MQ	F	valor-P
Regressão	2	3168,92	1584,46	4,31	0,0306
Resíduo	17	6245,37	367,37		
Total	19	9414,28			

$$F = \frac{SQE_R - SQE_C}{p_{C-R}} \div \frac{SQE_C}{n-p} \sim F_{2,15}$$

$$F = \frac{6245,37 - 59,71}{2} \div \frac{59,71}{15} = 776,8983 \quad \text{Valor-P} \cong 0$$

Conclusão: os modelos completo e reduzido são diferentes e portanto **não** se deve retirar as duas variáveis de uma só vez!

# Teste de Hipótese para múltiplos $\beta_k$

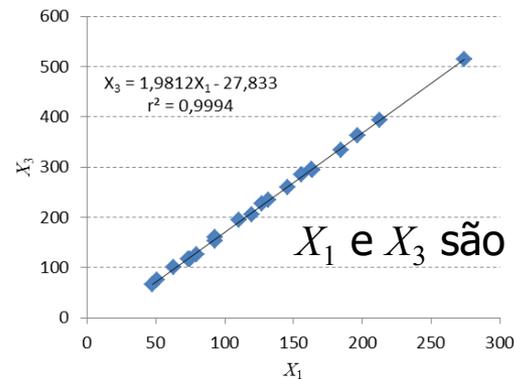
Y	X <sub>1</sub>	X <sub>2</sub>	X <sub>3</sub>	X <sub>4</sub>
11,70	126,92	174,56	226,69	364,26
16,34	75,02	129,40	117,43	329,68
16,76	51,00	106,17	75,41	592,57
16,83	47,75	110,50	66,58	471,11
22,02	145,83	148,78	258,84	1151,11
23,43	62,91	113,04	99,85	327,56
24,75	73,34	97,81	117,23	850,26
29,96	79,87	92,83	126,21	695,32
30,31	131,55	139,24	235,10	820,23
33,51	163,68	141,01	294,77	884,83
38,12	93,25	98,44	152,29	291,09
38,42	110,57	99,38	195,38	1162,36
40,63	93,28	88,63	159,74	338,08
46,15	196,54	140,37	363,28	508,84
47,98	184,33	128,83	334,06	764,28
54,58	119,84	71,83	204,97	709,91
58,22	163,02	102,36	295,87	626,23
66,27	155,43	84,14	284,87	50,34
86,27	273,91	109,00	514,30	620,11
89,29	212,29	53,56	392,89	1186,30

$$Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \beta_3 X_{3,i} + \beta_4 X_{4,i} + \xi_i$$

ANOVA

	gl	SQ	MQ	F	valor-P
Regressão	4	9354,57	2338,64	587,45	2,78E-16
Resíduo	15	59,71	3,98		
Total	19	9414,28			

De fato:



Mas afinal, quem eu retiro,  $X_1$  ou  $X_3$ ?

# Multicolinearidade

Y	X <sub>1</sub>	X <sub>2</sub>	X <sub>3</sub>	X <sub>4</sub>
11,70	126,92	174,56	226,69	364,26
16,34	75,02	129,40	117,43	329,68
16,76	51,00	106,17	75,41	592,57
16,83	47,75	110,50	66,58	471,11
22,02	145,83	148,78	258,84	1151,11
23,43	62,91	113,04	99,85	327,56
24,75	73,34	97,81	117,23	850,26
29,96	79,87	92,83	126,21	695,32
30,31	131,55	139,24	235,10	820,23
33,51	163,68	141,01	294,77	884,83
38,12	93,25	98,44	152,29	291,09
38,42	110,57	99,38	195,38	1162,36
40,63	93,28	88,63	159,74	338,08
46,15	196,54	140,37	363,28	508,84
47,98	184,33	128,83	334,06	764,28
54,58	119,84	71,83	204,97	709,91
58,22	163,02	102,36	295,87	626,23
66,27	155,43	84,14	284,87	50,34
86,27	273,91	109,00	514,30	620,11
89,29	212,29	53,56	392,89	1186,30

$$Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \beta_3 X_{3,i} + \beta_4 X_{4,i} + \xi_i$$

Em geral, a multicolinearidade provoca a perda da significância do coeficiente  $\beta_k$  associado à variável independente  $k$  que é relacionada a outra ou outras variáveis independentes. Isso acontece pois a variância do estimador  $s^2(b_k)$  é superestimada.

A detecção da multicolinearidade nem sempre é fácil e em geral recorre-se à análise do fator de inflação da variância (*VIF* – *Variance Inflation Factor*):

$$VIF_k = \frac{1}{1 - r_k^2}$$

onde  $r_k^2$  é o coeficiente de determinação obtido da regressão entre  $X_k$  e as demais variáveis independentes.

Em geral, se  $VIF_k > 10$  então  $X_k$  têm forte multicolinearidade

# Multicolinearidade

Y	X <sub>1</sub>	X <sub>2</sub>	X <sub>3</sub>	X <sub>4</sub>
11,70	126,92	174,56	226,69	364,26
16,34	75,02	129,40	117,43	329,68
16,76	51,00	106,17	75,41	592,57
16,83	47,75	110,50	66,58	471,11
22,02	145,83	148,78	258,84	1151,11
23,43	62,91	113,04	99,85	327,56
24,75	73,34	97,81	117,23	850,26
29,96	79,87	92,83	126,21	695,32
30,31	131,55	139,24	235,10	820,23
33,51	163,68	141,01	294,77	884,83
38,12	93,25	98,44	152,29	291,09
38,42	110,57	99,38	195,38	1162,36
40,63	93,28	88,63	159,74	338,08
46,15	196,54	140,37	363,28	508,84
47,98	184,33	128,83	334,06	764,28
54,58	119,84	71,83	204,97	709,91
58,22	163,02	102,36	295,87	626,23
66,27	155,43	84,14	284,87	50,34
86,27	273,91	109,00	514,30	620,11
89,29	212,29	53,56	392,89	1186,30

$$Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \beta_3 X_{3,i} + \beta_4 X_{4,i} + \xi_i$$

Exemplo: calculando-se o *VIF* para a variável  $X_1$

- Estima-se a regressão entre  $X_1$  e as demais variáveis independentes;
- Calcula-se o  $r^2$  e
- Obtém-se o *VIF*

$$X_{1,i} = a_0 + a_2 X_{2,i} + a_3 X_{3,i} + a_4 X_{4,i}$$

$$X_{1,i} = 14,04 - 0,003X_{2,i} + 0,50X_{3,i} + 0,0009X_{4,i}$$

$$r_1^2 = 0,9994$$

$$VIF_1 = \frac{1}{1 - 0,9994} = 1606,4$$

# Multicolinearidade no R

Y	X <sub>1</sub>	X <sub>2</sub>	X <sub>3</sub>	X <sub>4</sub>
11,70	126,92	174,56	226,69	364,26
16,34	75,02	129,40	117,43	329,68
16,76	51,00	106,17	75,41	592,57
16,83	47,75	110,50	66,58	471,11
22,02	145,83	148,78	258,84	1151,11
23,43	62,91	113,04	99,85	327,56
24,75	73,34	97,81	117,23	850,26
29,96	79,87	92,83	126,21	695,32
30,31	131,55	139,24	235,10	820,23
33,51	163,68	141,01	294,77	884,83
38,12	93,25	98,44	152,29	291,09
38,42	110,57	99,38	195,38	1162,36
40,63	93,28	88,63	159,74	338,08
46,15	196,54	140,37	363,28	508,84
47,98	184,33	128,83	334,06	764,28
54,58	119,84	71,83	204,97	709,91
58,22	163,02	102,36	295,87	626,23
66,27	155,43	84,14	284,87	50,34
86,27	273,91	109,00	514,30	620,11
89,29	212,29	53,56	392,89	1186,30

$$Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \beta_3 X_{3,i} + \beta_4 X_{4,i} + \xi_i$$

```
> Y <- c(11.7, ... ,89.29)
> X1 <- c(126.92, ... ,212.29)
> X2 <- c(174.56, ... ,53.56)
> X3 <- c(226.69, ... ,392.89)
> X4 <- c(364.26, ... ,1186.3)
> reg <- lm(Y~X1+X2+X3+X4)
> library(car)
> vif(reg)
```

X1	X2	X3	X4
1606.4	1.01	1602.7	1.12

Neste caso, explica-se a baixa significância de  $X_1$  e  $X_3$  pela multicolinearidade

Pode-se então eliminar a variável que apresente o maior *VIF* **repetindo-se a análise**

É importante que a eliminação das variáveis seja feita uma a uma até que não haja *VIF* muito altos

# Eliminando-se variáveis independentes

Y	X <sub>1</sub>	X <sub>2</sub>	X <sub>3</sub>	X <sub>4</sub>
11,70	126,92	174,56	226,69	364,26
16,34	75,02	129,40	117,43	329,68
16,76	51,00	106,17	75,41	592,57
16,83	47,75	110,50	66,58	471,11
22,02	145,83	148,78	258,84	1151,11
23,43	62,91	113,04	99,85	327,56
24,75	73,34	97,81	117,23	850,26
29,96	79,87	92,83	126,21	695,32
30,31	131,55	139,24	235,10	820,23
33,51	163,68	141,01	294,77	884,83
38,12	93,25	98,44	152,29	291,09
38,42	110,57	99,38	195,38	1162,36
40,63	93,28	88,63	159,74	338,08
46,15	196,54	140,37	363,28	508,84
47,98	184,33	128,83	334,06	764,28
54,58	119,84	71,83	204,97	709,91
58,22	163,02	102,36	295,87	626,23
66,27	155,43	84,14	284,87	50,34
86,27	273,91	109,00	514,30	620,11
89,29	212,29	53,56	392,89	1186,30

$$Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \beta_3 X_{3,i} + \beta_4 X_{4,i} + \xi_i$$

ANOVA

	gl	SQ	MQ	F	valor-P
Regressão	4	9354,57	2338,64	587,45	2,78E-16
Resíduo	15	59,71	3,98		
Total	19	9414,28			

	Erro			
	Coefficientes	padrão	Stat t	valor-P
Interseção	64,4359	4,8424	13,3067	1,04E-09
X <sub>1</sub>	-0,2129	0,3081	-0,6908	0,5002
X <sub>2</sub>	-0,4741	0,0160	-29,5575	1,04E-14
X <sub>3</sub>	0,2659	0,1553	1,7123	0,1074
X <sub>4</sub>	-0,0075	0,0015	-4,8827	0,0002

Outra maneira é eliminar-se primeiramente a variável que apresenta o maior *valor-P* não significativo

Em seguida, refaz-se a análise

# Eliminando-se variáveis independentes

Y	X <sub>2</sub>	X <sub>3</sub>	X <sub>4</sub>	Ŷ
11,70	174,56	226,69	364,26	11,97
16,34	129,40	117,43	329,68	16,28
16,76	106,17	75,41	592,57	18,58
16,83	110,50	66,58	471,11	16,06
22,02	148,78	258,84	1151,11	23,21
23,43	113,04	99,85	327,56	21,25
24,75	97,81	117,23	850,26	27,19
29,96	92,83	126,21	695,32	32,16
30,31	139,24	235,10	820,23	26,51
33,51	141,01	294,77	884,83	34,64
38,12	98,44	152,29	291,09	36,76
38,42	99,38	195,38	1162,36	36,44
40,63	88,63	159,74	338,08	42,23
46,15	140,37	363,28	508,84	48,71
47,98	128,83	334,06	764,28	47,57
54,58	71,83	204,97	709,91	54,49
58,22	102,36	295,87	626,23	55,11
66,27	84,14	284,87	50,34	66,43
86,27	109,00	514,30	620,11	86,67
89,29	53,56	392,89	1186,30	89,29

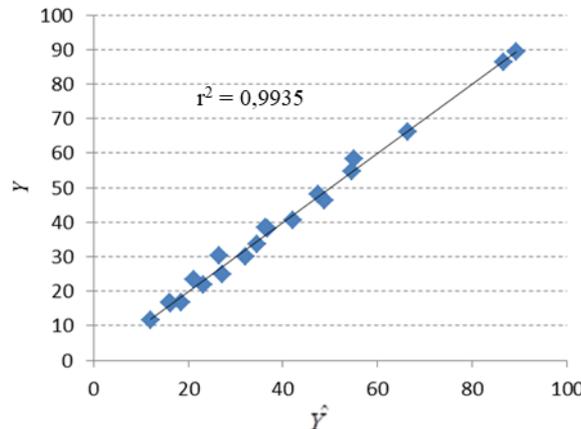
$$Y_i = \beta_0 + \beta_2 X_{2,i} + \beta_3 X_{3,i} + \beta_4 X_{4,i} + \xi_i$$

ANOVA

	gl	SQ	MQ	F	valor-P
Regressão	3	9352,67	3117,56	809,56	1,12E-17
Resíduo	16	61,61	3,85		
Total	19	9414,28			

	Coefficientes	Erro padrão	Stat t	valor-P
Interseção	61,4478	2,1408	28,7030	3,44E-15
X <sub>2</sub>	-0,4734	0,0157	-30,0637	1,66E-15
X <sub>3</sub>	0,1587	0,0040	40,0206	1,81E-17
X <sub>4</sub>	-0,0077	0,0015	-5,1698	9,31E-05

todos significativos a 5%



Coefficiente de correlação múltiplo

$$r = \sqrt{r^2}$$

$$r = 0,9964$$

(evite usar este índice!)

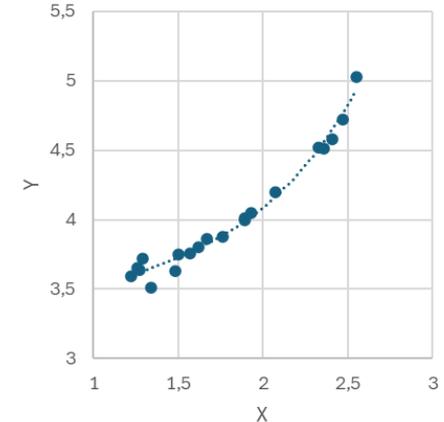
# Regressões Polinomiais

Y	X	X <sup>2</sup>	X <sup>3</sup>
4,2	2,07	4,28	8,87
3,8	1,62	2,62	4,25
3,88	1,76	3,1	5,45
3,63	1,48	2,19	3,24
4,05	1,93	3,72	7,19
3,86	1,67	2,79	4,66
3,75	1,5	2,25	3,38
4,58	2,41	5,81	14
3,59	1,22	1,49	1,82
3,76	1,57	2,46	3,87
4	1,89	3,57	6,75
3,51	1,34	1,8	2,41
4,72	2,47	6,1	15,07
3,65	1,26	1,59	2
3,72	1,29	1,66	2,15
5,03	2,55	6,5	16,58
4,01	1,89	3,57	6,75
4,51	2,36	5,57	13,14
3,64	1,27	1,61	2,05
4,52	2,33	5,43	12,65

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \beta_3 X_i^3 + \xi_i$$

ANOVA

	gl	SQ	MQ	F	valor-P
Regressão	3	3,4926	1,1642	323,7193	1,58E-14
Resíduo	16	0,0575	0,0036		
Total	19	3,5501			



	Erro			
	Coefficientes	padrão	Stat t	valor-P
Interseção	1,2692	1,5608	0,8132	0,4281
X	3,9589	2,6748	1,4800	0,1583
X <sup>2</sup>	-2,3363	1,4873	-1,5709	0,1358
X <sup>3</sup>	0,5291	0,2681	1,9739	0,0659

	VIF
X	7480,64
X <sup>2</sup>	32554,03
X <sup>3</sup>	9028,64

Mas como pode todos os coeficientes serem não significativos com o valor-P da ANOVA tão pequeno?

**Atenção:** Termos polinomiais e interações podem apresentar *VIF* alto pois podem apresentar relações quase lineares entre os termos do polinômio.

# Regressões Polinomiais

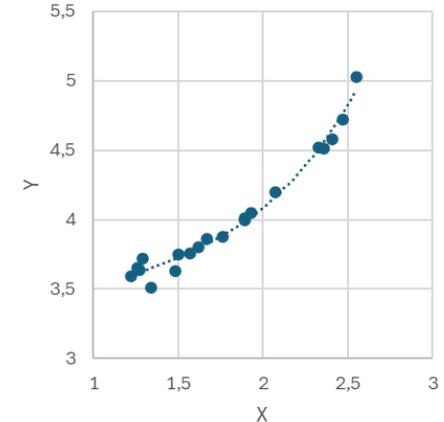
Y	X	X <sup>2</sup>	X <sup>3</sup>
4,2	2,07	4,28	8,87
3,8	1,62	2,62	4,25
3,88	1,76	3,1	5,45
3,63	1,48	2,19	3,24
4,05	1,93	3,72	7,19
3,86	1,67	2,79	4,66
3,75	1,5	2,25	3,38
4,58	2,41	5,81	14
3,59	1,22	1,49	1,82
3,76	1,57	2,46	3,87
4	1,89	3,57	6,75
3,51	1,34	1,8	2,41
4,72	2,47	6,1	15,07
3,65	1,26	1,59	2
3,72	1,29	1,66	2,15
5,03	2,55	6,5	16,58
4,01	1,89	3,57	6,75
4,51	2,36	5,57	13,14
3,64	1,27	1,61	2,05
4,52	2,33	5,43	12,65

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \beta_3 X_i^3 + \xi_i$$

No R, é possível gerar uma regressão polinomial:

```
summary(lm(Y ~ poly(X,3)))
```

	Erro			
	Coeficientes	padrão	Stat t	valor-P
Interseção	4,0205	0,0141	285,710	< 2e-16
X	1,8190	0,0629	28,904	3,08e-15
X <sup>2</sup>	0,4147	0,0629	6,590	6,22e-6
X <sup>3</sup>	0,0778	0,0629	1,237	0,234



Observe que agora os termos X e X<sup>2</sup> são significativos!

**Cuidado:** os valores estimados dos parâmetros não podem ser diretamente interpretados (estimação usa transformações ortogonais!)

Observe que X<sup>3</sup> não é significativo e pode ser descartado

Para se estimar os parâmetros do polinômio, deve-se realizar a regressão clássica:

```
summary(lm(Y ~ poly(X,2, raw=TRUE))) ou summary(lm(Y ~ X + I(X^2)))
```

	Erro			
	Coeficientes	padrão	Stat t	valor-P
Interseção	4,3188	0,3113	13,875	1,06e-10
X	-1,3081	0,3477	-3,763	0,0016
X <sup>2</sup>	0,6013	0,0927	0,0927	5,55e-6

ignore esses valores!

# Variáveis Indicadoras

Variáveis indicadoras (ou *dummies*) são variáveis que só pode assumir dois valores: 0 ou 1

Elas são amplamente utilizadas na regressão podendo aparecer isoladas ou interagindo com outras variáveis independentes

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 W_i + \beta_3 X_i W_i + \xi_i \quad W_i = \{0,1\}$$

Com esse modelo, é possível ajustar 2 regressões simultaneamente:

$$\begin{aligned} \text{se } W_i = 0 & \quad Y_i = \beta_0 + \beta_1 X_i + \xi_i \\ \text{se } W_i = 1 & \quad Y_i = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) X_i + \xi_i = \beta'_0 + \beta'_1 X_i + \xi_i \end{aligned}$$

Aplicações:

- a) Representar variáveis categóricas
- b) Verificar se certos parâmetros diferem entre grupos (comparação de modelos)
- c) Modelar mudanças abruptas ou estruturais

# Variáveis Indicadoras – Variável Categórica

Se uma variável categórica for representada por apenas 2 classes (classes A e B) então apenas uma variável indicadora precisa ser definida

$$W_i = \begin{cases} 0 & \text{se } i \text{ pertencer a Classe A} \\ 1 & \text{se } i \text{ pertencer a Classe B} \end{cases}$$

Agora, se houverem 3 classes (classes A, B e C), são necessárias 2 variáveis indicadoras:

$W_{1i}$	$W_{2i}$	Classe
0	0	A
0	1	B
1	1	C

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 W_{1i} + \beta_3 X_i W_{1i} + \beta_4 W_{2i} + \beta_5 X_i W_{2i} + \xi_i$$

$$\text{Classe A: } Y_i = \beta_0 + \beta_1 X_i + \xi_i$$

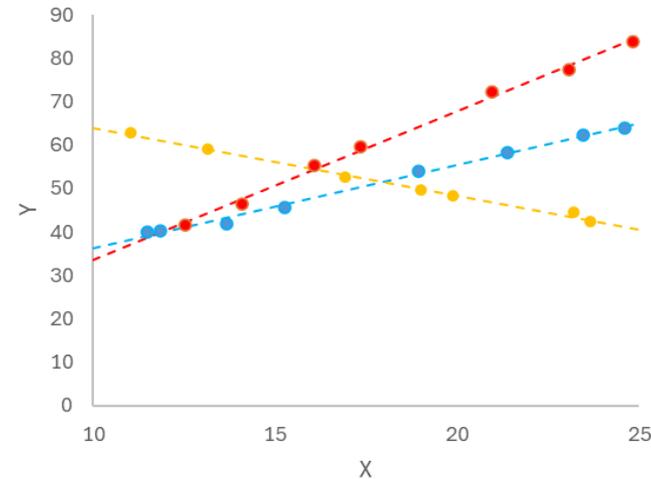
$$\text{Classe B: } Y_i = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) X_i + \xi_i$$

$$\text{Classe C: } Y_i = (\beta_0 + \beta_2 + \beta_4) + (\beta_1 + \beta_3 + \beta_5) X_i + \xi_i$$

# Variáveis Indicadoras – Variável Categórica

Y	X <sub>1</sub>	X <sub>2</sub>	W <sub>1</sub>	W <sub>2</sub>	X <sub>1</sub> W <sub>1</sub>	X <sub>1</sub> W <sub>2</sub>
40,12	11,51	A	0	0	0	0
41,98	13,67	A	0	0	0	0
53,89	18,95	A	0	0	0	0
64,07	24,60	A	0	0	0	0
40,24	11,85	A	0	0	0	0
45,58	15,28	A	0	0	0	0
58,38	21,38	A	0	0	0	0
62,3	23,47	A	0	0	0	0
59,74	17,35	B	0	1	0	17,35
72,26	20,97	B	0	1	0	20,97
46,38	14,09	B	0	1	0	14,09
55,43	16,08	B	0	1	0	16,08
41,69	12,53	B	0	1	0	12,53
77,42	23,08	B	0	1	0	23,08
83,79	24,83	B	0	1	0	24,83
59,24	13,17	C	1	1	13,17	13,17
52,63	16,92	C	1	1	16,92	16,92
49,82	19,02	C	1	1	19,02	19,02
44,5	23,22	C	1	1	23,22	23,22
62,79	11,04	C	1	1	11,04	11,04
48,27	19,88	C	1	1	19,88	19,88
42,39	23,67	C	1	1	23,67	23,67

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 W_{1i} + \beta_3 X_{1i}W_{1i} + \beta_4 W_{2i} + \beta_5 X_{1i}W_{2i} + \xi_i$$



ANOVA

	gl	SQ	MQ	F	valor-P
Regressão	5	3160,54	632,11	834,31	9,56E-19
Resíduo	16	12,12	0,7576		
Total	21	3172,66			

altamente significativo

	Erro			
	Coefficientes	padrão	Stat t	valor-P
Interseção	17,06	1,15	14,82	9,15E-11
X <sub>1</sub>	1,92	0,06	30,44	1,36E-15
W <sub>1</sub>	80,27	2,01	40,01	1,81E-17
W <sub>2</sub>	-17,71	1,85	-9,58	5,01E-08
X <sub>1</sub> W <sub>1</sub>	-4,98	0,11	-46,65	1,59E-18
X <sub>1</sub> W <sub>2</sub>	1,50	0,10	15,16	6,54E-11

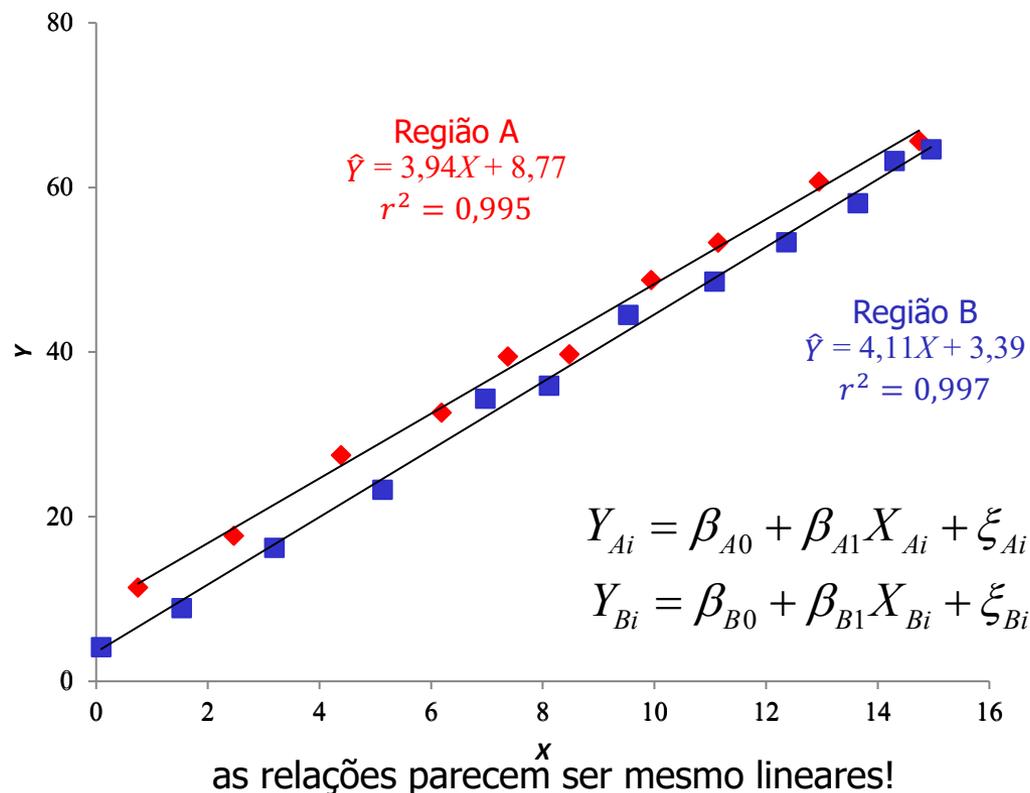
# Variáveis Indicadoras – Comparação de modelos

Suponha que temos 2 modelos ajustados para 2 regiões distintas

Como saber se os parâmetros de ambos os modelos são iguais, ou seja, se ambas as regiões possuem o mesmo modelo?

Região A	
Y	X
11,40	0,75
17,69	2,47
27,48	4,39
32,65	6,19
39,46	7,38
39,73	8,47
48,76	9,94
53,30	11,14
60,71	12,95
65,65	14,75

Região B	
Y	X
4,17	0,09
8,92	1,53
16,23	3,19
23,28	5,13
34,33	6,97
35,93	8,11
44,53	9,53
48,57	11,08
53,37	12,36
58,11	13,64
63,21	14,30
64,67	14,96



Para que ambas regressões sejam a mesma:  $\beta_{A0} = \beta_{B0}$  e  $\beta_{A1} = \beta_{B1}$

# Variáveis Indicadoras – Comparação de modelos

Y	X
11,40	0,75
17,69	2,47
27,48	4,39
32,65	6,19
39,46	7,38
39,73	8,47
48,76	9,94
53,30	11,14
60,71	12,95
65,65	14,75
4,17	0,09
8,92	1,53
16,23	3,19
23,28	5,13
34,33	6,97
35,93	8,11
44,53	9,53
48,57	11,08
53,37	12,36
58,11	13,64
63,21	14,30
64,67	14,96

Define-se uma nova variável  $W$ :

$$W_i = \begin{cases} 0 & \text{se } i \text{ pertencer a Região A} \\ 1 & \text{se } i \text{ pertencer a Região B} \end{cases}$$

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 W_i + \beta_3 X_i W_i + \xi_i$$

Para Região A ( $W_i = 0$ ):

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

Para Região B ( $W_i = 1$ ):

$$Y_i = \underbrace{(\beta_0 + \beta_2)}_{\beta'_0} + \underbrace{(\beta_1 + \beta_3)}_{\beta'_1} X_i + \xi_i$$

# Variáveis Indicadoras – Comparação de modelos

$Y$	$X$	$W$	$XW$
11,40	0,75	0	0
17,69	2,47	0	0
27,48	4,39	0	0
32,65	6,19	0	0
39,46	7,38	0	0
39,73	8,47	0	0
48,76	9,94	0	0
53,30	11,14	0	0
60,71	12,95	0	0
65,65	14,75	0	0
4,17	0,09	1	0,09
8,92	1,53	1	1,53
16,23	3,19	1	3,19
23,28	5,13	1	5,13
34,33	6,97	1	6,97
35,93	8,11	1	8,11
44,53	9,53	1	9,53
48,57	11,08	1	11,08
53,37	12,36	1	12,36
58,11	13,64	1	13,64
63,21	14,30	1	14,30
64,67	14,96	1	14,96

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 W_i + \beta_3 X_i W_i + \xi_i$$

Conclusões possíveis:

Se  $\beta_2 = \beta_3 = 0$ , então ambas regiões possuem a mesma regressão

Se  $\beta_2 \neq 0$ , então as regressões diferem-se entre si pelo intercepto

Se  $\beta_3 \neq 0$ , então as regressões diferem-se entre si pelo coeficiente angular

# Variáveis Indicatoras – Comparação de modelos

Y	X	W	XW
11,40	0,75	0	0
17,69	2,47	0	0
27,48	4,39	0	0
32,65	6,19	0	0
39,46	7,38	0	0
39,73	8,47	0	0
48,76	9,94	0	0
53,30	11,14	0	0
60,71	12,95	0	0
65,65	14,75	0	0
4,17	0,09	1	0,09
8,92	1,53	1	1,53
16,23	3,19	1	3,19
23,28	5,13	1	5,13
34,33	6,97	1	6,97
35,93	8,11	1	8,11
44,53	9,53	1	9,53
48,57	11,08	1	11,08
53,37	12,36	1	12,36
58,11	13,64	1	13,64
63,21	14,30	1	14,30
64,67	14,96	1	14,96

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 W_i + \beta_3 X_i W_i + \xi_i$$

ANOVA

	gl	SQ	MQ	F	valor-P
Regressão	3	7692,29	2564,10	1499,97	8,9E-22
Resíduo	18	30,77	1,71		
Total	21	7723,06			

altamente significativo

	Erro			
	Coefficientes	padrão	t	valor-P
Interseção	8,77	0,86	10,17	6,86E-09
X	3,94	0,10	40,84	3,36E-19
W	-5,38	1,14	-4,70	0,0002
XW	0,17	0,12	1,36	0,1915

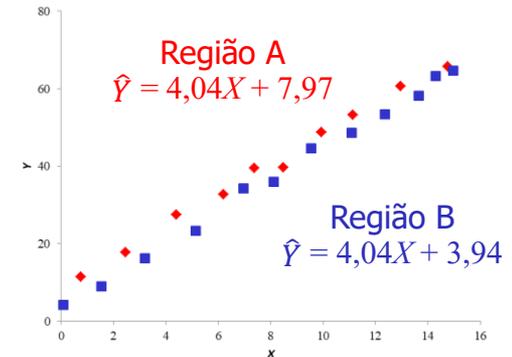
não significativo ( $\beta_3 = 0$ )

Elimina-se o termo  $\beta_3 X_i W_i$  e refaz-se a análise...

# Variáveis Indicatoras – Comparação de modelos

Y	X	W
11,40	0,75	0
17,69	2,47	0
27,48	4,39	0
32,65	6,19	0
39,46	7,38	0
39,73	8,47	0
48,76	9,94	0
53,30	11,14	0
60,71	12,95	0
65,65	14,75	0
4,17	0,09	1
8,92	1,53	1
16,23	3,19	1
23,28	5,13	1
34,33	6,97	1
35,93	8,11	1
44,53	9,53	1
48,57	11,08	1
53,37	12,36	1
58,11	13,64	1
63,21	14,30	1
64,67	14,96	1

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 W_i + \xi_i$$



ANOVA

	gl	SQ	MQ	F	valor-P
Regressão	2	7689,14	3844,57	2153,57	4,03E-23
Resíduo	19	33,92	1,79		
Total	21	7723,06			

	Erro			
	Coefficientes	padrão	t	valor-P
Interseção	7,97	0,64	12,41	1,46E-10
X	4,04	0,06	65,56	7,44E-24
W	-4,03	0,57	-7,03	1,09E-06

ambos significativos a 5%

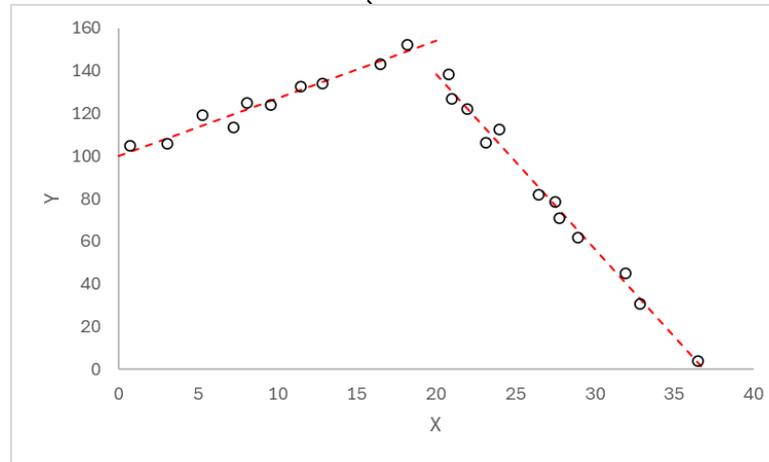
**Conclusão:** a 5% de significância, as regressões de ambas regiões possuem o mesmo coeficiente angular. Elas diferem-se apenas pelo intercepto.

Em média, a região B produz estimativas para Y menores que a região A em 4,03 unidades.

# Variáveis Indicadoras – Mudanças estruturais

Y	X	W	XW	Trecho
104,93	0,75	0	0	X < 20
105,75	3,11	0	0	X < 20
119,19	5,33	0	0	X < 20
113,34	7,27	0	0	X < 20
125,02	8,12	0	0	X < 20
124,23	9,59	0	0	X < 20
132,58	11,49	0	0	X < 20
133,82	12,84	0	0	X < 20
143,09	16,49	0	0	X < 20
152,37	18,21	0	0	X < 20
138,34	20,81	1	20,81	X >= 20
126,91	21,03	1	21,03	X >= 20
122,30	21,98	1	21,98	X >= 20
106,38	23,15	1	23,15	X >= 20
112,40	24,00	1	24,00	X >= 20
81,85	26,45	1	26,45	X >= 20
78,82	27,52	1	27,52	X >= 20
70,81	27,80	1	27,80	X >= 20
61,85	28,97	1	28,97	X >= 20
45,15	31,96	1	31,96	X >= 20
30,73	32,87	1	32,87	X >= 20
4,14	36,54	1	36,54	X >= 20

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 W_i + \beta_3 X_i W_i + \xi_i$$



## ANOVA

	gl	SQ	MQ	F	valor-P
Regressão	3	31224,24	10408,08	600,75	3,17E-18
Resíduo	18	311,85	17,32		
Total	21	31536,09			

altamente significativo

	Erro			valor-P
	Coefficientes	padrão	Stat t	
Interseção	100,28	2,66	37,67	1,41E-18
X	2,70	0,25	10,87	2,43E-9
W	202,24	7,33	27,60	3,49E-16
XW	-10,90	0,35	-30,97	4,56E-17

# Regressão Padronizada

Muitas vezes, o coeficiente  $\beta_k$  pode ser utilizado como uma medida do poder da variável independente  $k$  em “explicar” a variável dependente  $Y$

Por exemplo:  $\hat{Y}_i = 10,5 + 0,4X_{1,i} + 5,9X_{2,i}$

Observe que a variação em 1 unidade de  $X_2$  gera uma mudança em 5,9 unidades em  $Y$ , ao passo que a mesma variação em  $X_1$  gera uma mudança de apenas 0,4. Assim, conclui-se que a variável  $X_2$  é mais importante para  $Y$  do que  $X_1$ . Será mesmo?

Isso é verdade quando todas as variáveis independentes possuem a mesma unidade de medida e quando possuem variâncias similares

No exemplo anterior, se as unidades das variáveis do modelo fossem:  $Y$  em mm,  $X_1$  em ton/ha e  $X_2$  em °C, quais as unidades de  $\beta_1$  e  $\beta_2$ ?

$$\beta_1 \rightarrow \text{mm.ha/ton}$$

$$\beta_2 \rightarrow \text{mm/}^\circ\text{C}$$

Como comparar estes parâmetros?

# Regressão Padronizada

Para obter um modelo cujos coeficientes sejam adimensionais, deve-se padronizar cada uma das variáveis dependente e independentes, ou seja:

$$Y'_i = \frac{Y_i - \bar{Y}}{S_Y} \quad X'_{k,i} = \frac{X_{k,i} - \bar{X}_k}{S_{X_k}}$$

Nesse caso, a reta de regressão estimada

$$\hat{Y}_i = b_0 + b_1 X'_{1,i} + b_2 X'_{2,i} + \dots + b_{p-1} X'_{p-1,i}$$

torna-se

$$\hat{Y}'_i = b'_1 X'_{1,i} + b'_2 X'_{2,i} + \dots + b'_{p-1} X'_{p-1,i} \quad \Rightarrow \quad b'_k = b_k \frac{S_{X_k}}{S_Y}$$

Estes coeficientes podem então ser comparados entre si

Em muitos pacotes estatísticos, estes coeficientes são conhecidos como “coeficientes beta”

# Construção do Modelo

---

Em muitos casos, dispomos de muitas variáveis independentes que podem ou não estar relacionadas com a variável dependente. Em geral, o objetivo de um estudo de regressão é determinar quais dessas variáveis independentes disponíveis melhor explicam ou predizem a variável em estudo

Pode-se imaginar que muitos modelos podem ser estimados. Nesse caso, deve-se buscar o melhor modelo que represente a relação entre as variáveis, ou seja, aquele que melhor se ajuste aos dados analisados

## Dicas:

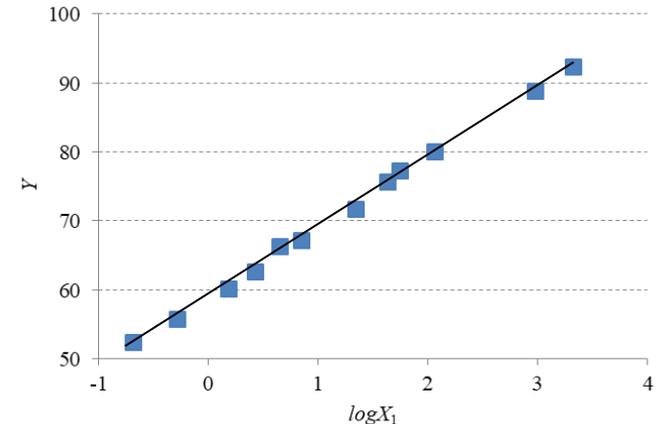
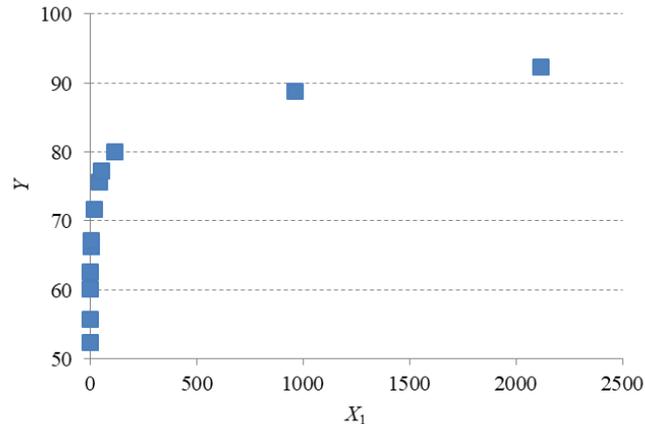
- quanto mais simples o modelo, melhor
- dê preferência por modelos lineares (ou linearizáveis)
- utilize conhecimentos prévios para escolha do tipo de relação entre cada variável independente e a dependente (linear, polinomial, logarítmico, exponencial, potencial, etc), construindo primeiramente um modelo conceitual ou analise modelos utilizados em trabalhos semelhantes
- evite métodos automáticos que “procuram” esta melhor relação. A escolha do tipo de relação deve ser fundamentada em conhecimentos prévios ou baseada em diagramas de dispersão. Na dúvida, utilize relações lineares
- após a estimação dos parâmetros, faça a análise dos resíduos para detectar quaisquer anomalias (relação inadequada, *outliers*, não normalidade, não constância da variância, etc) e tente minimizá-las

# Construção do Modelo

Quando se trabalha com um grande número de variáveis independentes, muitas vezes o processo de escolha de quais deverão compor o modelo final é bastante dificultado, especialmente quando há colinearidade entre estas variáveis

De modo geral, o primeiro passo é verificar se a relação entre a variável dependente e cada uma das variáveis independentes é linear. No caso da relação não ser linear, procura-se transformações de modo a linearizá-la. Se não houver nenhuma evidência, deixe-a na sua forma original

$Y$	$X_1$	$\log X_1$
52,31	0,21	-0,67
55,78	0,53	-0,28
60,05	1,55	0,19
62,66	2,69	0,43
66,32	4,53	0,66
67,16	7,05	0,85
71,69	22,24	1,35
75,59	43,40	1,64
77,17	55,43	1,74
80,02	116,31	2,07
88,78	964,13	2,98
92,32	2117,60	3,33



$$Y_i = \beta_0 + \beta_1 \log X_{1i} + \xi_i$$

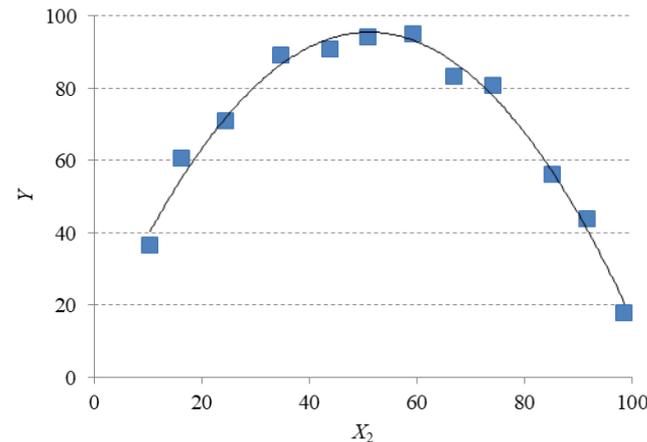
Nesse caso, utiliza-se somente a variável transformada descartando-se a original

# Construção do Modelo

Quando se trabalha com um grande número de variáveis independentes, muitas vezes o processo de escolha de quais deverão compor o modelo final é bastante dificultado, especialmente quando há colinearidade entre estas variáveis

De modo geral, o primeiro passo é verificar se a relação entre a variável dependente e cada uma das variáveis independentes é linear. No caso da relação não ser linear, procura-se transformações de modo a linearizá-la. Se não houver nenhuma evidência, deixe-a na sua forma original

$Y$	$X_1$	$X_2^2$
36,5	10,3	106,09
60,7	16,3	265,69
71,1	24,5	600,25
89,2	34,7	1204,09
90,8	43,8	1918,44
94,1	51,0	2601,00
95,2	59,2	3504,64
83,3	66,9	4475,61
80,8	74,2	5505,64
56,1	85,2	7259,04
43,8	91,7	8408,89
17,9	98,6	9721,96



$$Y_i = \beta_0 + \beta_1 X_{2,i} + \beta_2 X_{2,i}^2 + \xi_i$$

**OBS:** no caso de adotar modelos polinomiais, considere, a princípio, todos os termos do polinômio!

# Construção do Modelo

---

Quando se trabalha com um grande número de variáveis independentes, muitas vezes o processo de escolha de quais deverão compor o modelo final é bastante dificultado, especialmente quando há colinearidade entre estas variáveis

De modo geral, o primeiro passo é verificar se a relação entre a variável dependente e cada uma das variáveis independentes é linear. No caso da relação não ser linear, procura-se transformações de modo a linearizá-la. Se não houver nenhuma evidência, deixe-a na sua forma original

Todas as variáveis categóricas devem ser convertidas para variáveis indicadoras e, se necessário, incluir também as interações com todas as outras variáveis independentes do modelo

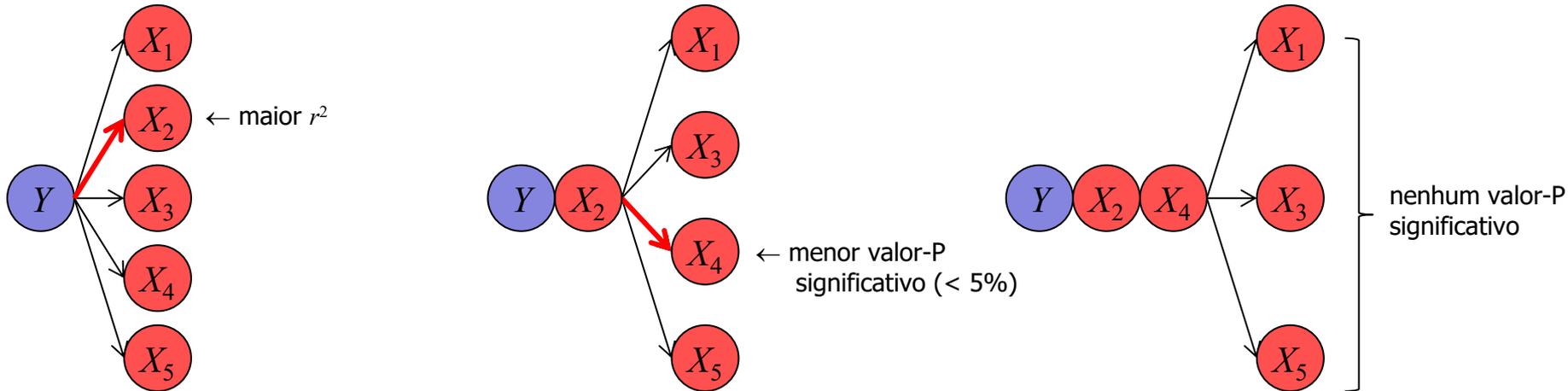
$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \xi_i$$

Se  $X_3 = \{0,1\}$  for uma variável indicadora, então o novo modelo será:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{1i} X_{3i} + \beta_5 X_{2i} X_{3i} + \xi_i$$

# Seleção de Variáveis

A seleção pode ser feita manualmente, identificando-se a variável independente com maior poder de explicação (maior  $r^2$  ou menor valor-P) e em seguida, acrescenta-se uma a uma, cada variável independente, testando-se a significância de cada variável independente adicionada.



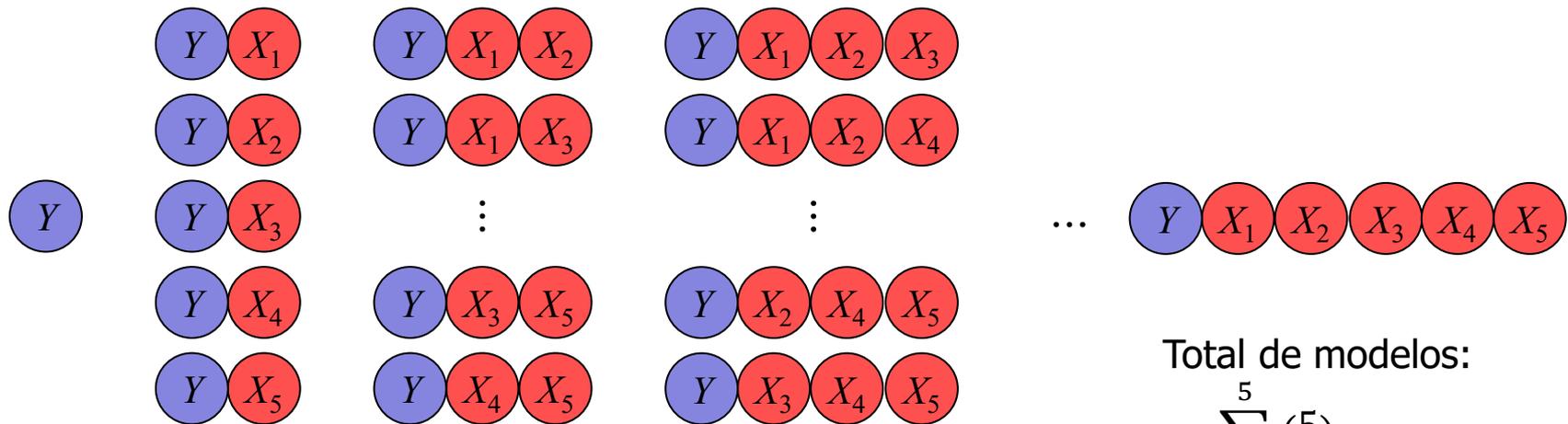
$$\text{Modelo Final: } Y_i = \beta_0 + \beta_2 X_{2,i} + \beta_4 X_{4,i} + \xi_i$$

Este processo não garante que o modelo final seja o melhor dentre todos os possíveis modelos. Esta seleção pode ser otimizada através de processos automáticos

Os mais comuns são: **busca exaustiva** e **stepwise**.

# Seleção de Variáveis – Busca Exaustiva

Na busca exaustiva, escolhe-se o melhor modelo simples (1 variável independente) e depois o melhor modelo com 2 variáveis (todos os pares são testados) e depois o modelo com 3 variáveis (todas as triplas são testadas), até que o modelo completo seja ajustado. Avalia-se os modelos obtidos (do mais simples ao mais completo) de forma a garantir que o acréscimo de variáveis independentes traga ganhos significativos.



Total de modelos:

$$\sum_{i=0}^5 \binom{5}{i} = 32$$

Este método é muito oneroso e inviável quando se trabalha com muitas variáveis independentes. Se fossem 10 variáveis, haveriam 1024 modelos a serem testados!

# Seleção de Variáveis - *Stepwise*

Há três maneiras de se aplicar o método *stepwise* ("passo a passo"):  
crescente (*forward*), decrescente (*backward*) ou ambos (*both*)

No modo *forward*, o modelo é inicializado sem nenhuma variável independente (modelo nulo:  $Y = \beta_0$ ) e, a cada passo, adiciona-se uma nova variável independente, testando-se o ganho no poder explicativo do novo modelo.

No modo *backward*, inicia-se o modelo com todas as variáveis independentes e, a cada passo, retira-se uma das variáveis do modelo, testando-se a perda no poder explicativo do novo modelo reduzido.

No modo *both*, a cada passo testa-se a entrada e a retirada de cada variável independente.

O teste utilizado para medir o ganho ou a perda do poder explicativo pode variar mas, em geral, utiliza-se o teste F para comparar os modelos completo e reduzido, ou o teste t quando apenas um parâmetro é adicionado ou retirado.

$$F = \frac{SQE_R - SQE_C}{p_{C-R}} \div \frac{SQE_C}{n-p} \sim F_{p_{C-R}, n-p}$$

# Seleção de Variáveis - *Stepwise*

Além do teste F que avalia o ganho significativo de um modelo em relação a outro, pode-se também adotar outros critérios para decidir qual modelo utilizar:

- Coeficiente de Determinação Múltiplo Ajustado

$$r_a^2 = 1 - \left( \frac{n-1}{n-p} \right) \frac{SQE}{SQTO} \quad \text{Quanto maior for o valor } r_a^2, \text{ melhor o modelo.}$$

- Critério  $C_p$  de Mallows

$$C_p = \frac{SQE_{R(p)}}{QME_C} - n + 2p \quad C_p \leq p \quad \text{Quanto mais próximo } C_p \text{ de } p, \text{ melhor o modelo.}$$

$SQE_{R(p)}$  é a soma dos quadrados do erro do modelo com  $p$  parâmetros e  $QME_C$  é o quadrado médio do modelo completo (com todas as variáveis).

- Índice AIC (*Akaike's Information Criterion*):

$$AIC = 2p + n \log(SQE / n)$$

Observe que este índice é uma combinação entre uma medida de ajuste ( $SQE$ ) e uma medida de simplicidade do modelo (dado pelo número de parâmetros  $p$ ). Quanto menor for o valor AIC, melhor o modelo.

Muitas variações deste índice podem ser encontradas na literatura.

# Seleção de Variáveis - *Stepwise*

---

- O método de seleção de variáveis *Stepwise* é particularmente útil quando se tem um grande número de variáveis independentes candidatas a compor o modelo de regressão. Nesse caso, o método *backward* não é recomendado pois, em geral, ocorre um "super ajuste" (*overfit*) dos dados quando o número de parâmetros do modelo se aproxima ao tamanho da amostra.
- É importante notar que o *Stepwise* não avalia a significância dos coeficientes do modelo final selecionado e nem a sua adequabilidade. Por essa razão, é imprescindível proceder à análise completa do modelo selecionado.
- É também bastante recomendável analisar os resíduos do modelo selecionado contra todas as variáveis que ficaram de fora para verificar se alguma, após alguma transformação ou retirada de *outliers*, poderia ser novamente avaliada. Nesse caso, repete-se todo o processo.

# Exemplo em R

Os dados estão disponíveis em RegrDados.dat

Y	X1	X2	X3	X4	X5
100,9716	110,6689	191,5408	0,035715	0,163621	110,631
112,1469	105,8706	243,3895	0,15264	0,106897	110,4889
104,326	111,5136	62,96696	0,132668	0,094935	110,9833
108,5016	96,00361	195,0345	0,050644	0,235669	109,6515
96,11517	100,5438	71,00401	0,006774	0,274924	109,5986
106,3803	100,9839	277,3767	0,023691	0,088838	110,0388
92,81556	103,8784	243,5599	0,000303	0,160092	110,1938
113,3299	108,8013	221,162	0,935501	0,219022	110,8796
102,8919	107,694	212,9797	1,302829	0,316011	110,4447
114,0049	105,906	233,7622	0,299974	0,169543	109,7264
94,63697	107,5253	173,7078	0,021178	0,158086	110,3228
111,4428	102,9983	216,4401	0,037315	0,135116	109,5703
105,0202	105,3302	163,8557	0,289124	0,163017	109,7318
88,41419	114,4864	165,1862	0,008662	0,145228	111,219
99,81212	109,5304	121,4531	0,08814	0,25679	110,839
87,03449	108,7498	335,4734	0,002011	0,238312	109,9678
106,95	102,1372	126,4867	0,08776	0,205099	109,607
110,5253	109,228	322,801	0,082994	0,097948	110,06
83,67311	111,4563	137,3347	0,001553	0,267389	110,2006
102,5173	109,5361	304,8186	0,023646	0,09666	110,5746

Y	X1	X2	X3	X4	X5
108,1522	95,8132	164,1165	0,01457	0,244236	109,308
103,6505	99,70563	275,3885	0,007832	0,262593	109,4251
111,4834	106,1368	201,3384	1,696478	0,226717	110,2725
123,1041	95,47518	217,71	0,780169	0,199064	108,8689
100,5872	106,9972	268,7112	0,046441	0,243702	109,9278
96,44818	117,2304	272,5158	0,027461	0,168976	111,2666
105,4046	109,6684	247,6312	0,050508	0,195319	110,1356
107,6394	102,7008	122,6451	0,093624	0,127112	109,5073
119,6807	93,15921	179,5055	0,336576	0,188007	109,0429
110,9345	94,74519	271,6365	0,096466	0,292361	109,4735
104,8678	102,6694	282,3483	0,051442	0,080458	109,7799
105,8734	100,8227	109,4809	0,031293	0,175168	109,9754
95,95295	116,9337	219,9231	0,057211	0,207979	110,7389
125,1698	101,5443	409,9216	0,270551	0,003029	109,9989
93,63811	111,3147	464,3571	0,05907	0,153045	110,6434
110,7853	98,80057	91,17997	0,158132	0,165678	109,6072
108,5595	112,7691	259,943	0,557714	0,176864	110,8162
113,2232	104,3571	191,8647	1,701136	0,270866	109,5895
99,92746	99,8217	237,0379	0,03948	0,323259	109,7102
110,344	103,6897	406,2605	0,069096	0,087841	110,2778

# Exemplo em R

```
# Entrada dos dados
```

```
dados <- read.csv("RegrDados.dat", header = TRUE, sep="\t", dec = ".", na.strings = NA)
```

```
# Plotando gráficos de dispersão e correlações
```

```
upanel <- function(x, y, ...)
```

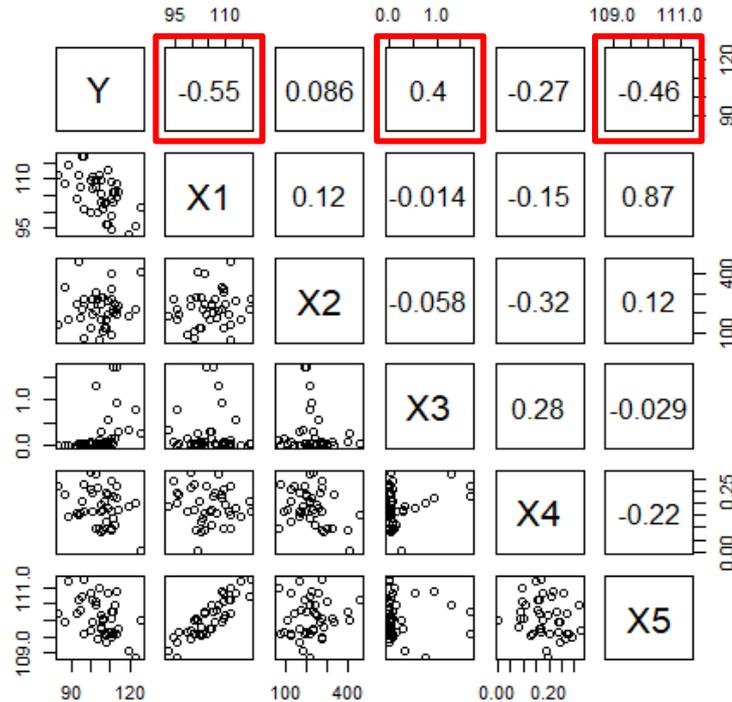
```
{
```

```
  par(usr = c(0, 1, 0, 1))
```

```
  text(0.5, 0.5, format(cor(x, y), digits=2), cex = 1.5)
```

```
}
```

```
pairs(dados, upper.panel=upanel)
```



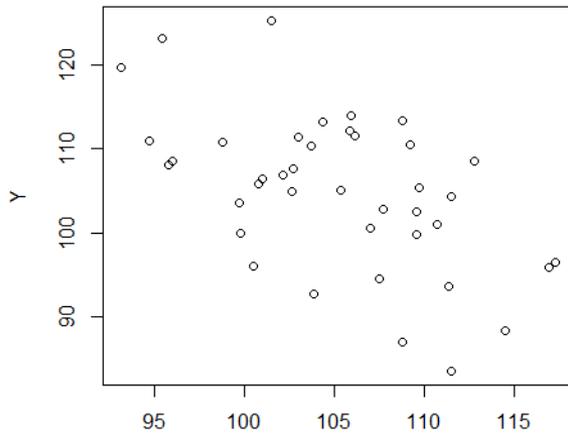
Qual são as melhores variáveis que explicam  $Y$ ?

Se avaliarmos apenas a correlação:  $X1$ ,  $X3$  e  $X5$

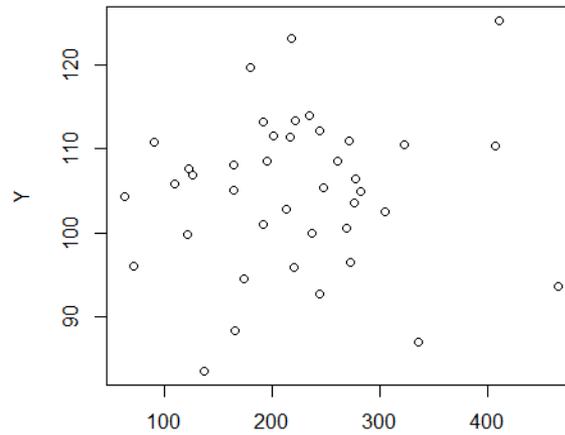
Mas todas as relações das variáveis independentes com a  $Y$  são lineares?

# Exemplo

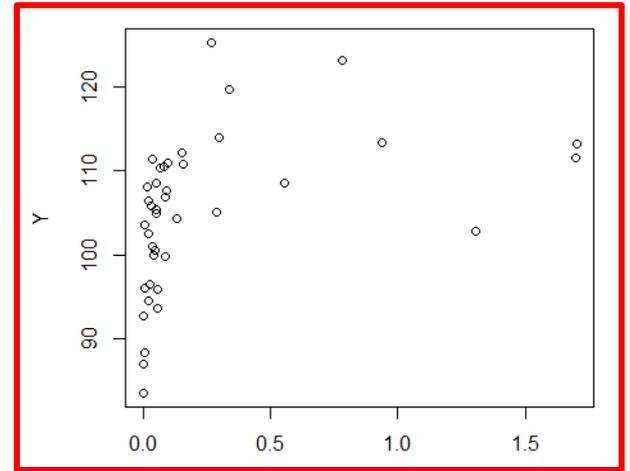
`plot(Y~X1,data=dados)` `plot(Y~X2,data=dados)` `plot(Y~X3,data=dados)` `plot(Y~X4,data=dados)` `plot(Y~X5,data=dados)`



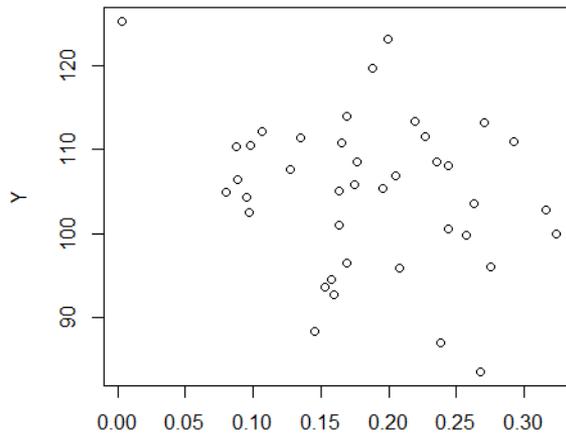
X1



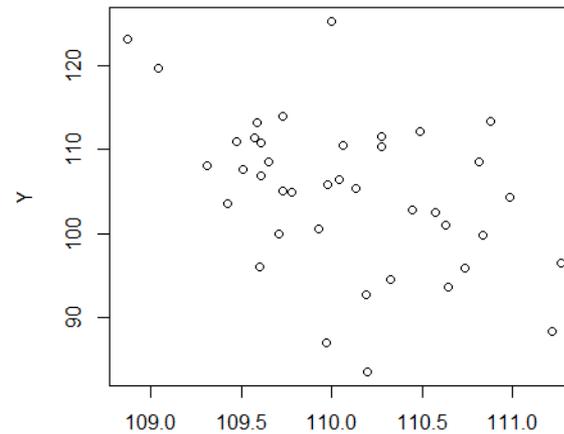
X2



X3



X4

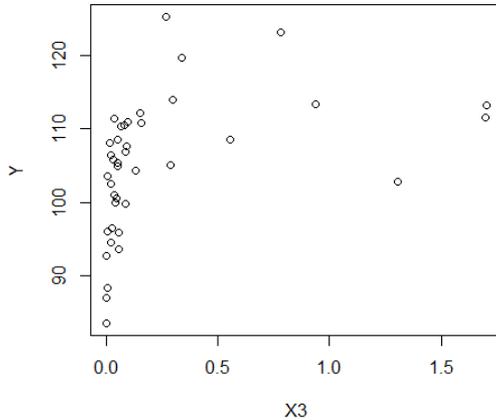


X5

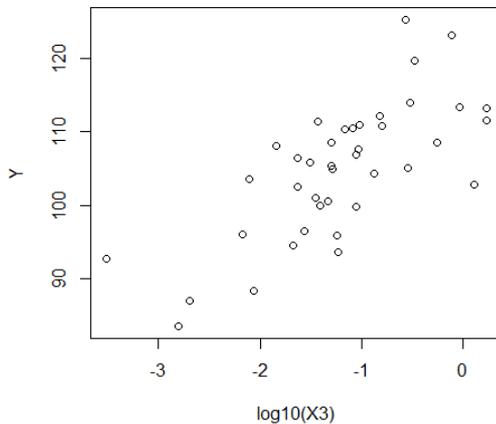
relação parece  
não ser linear

# Exemplo

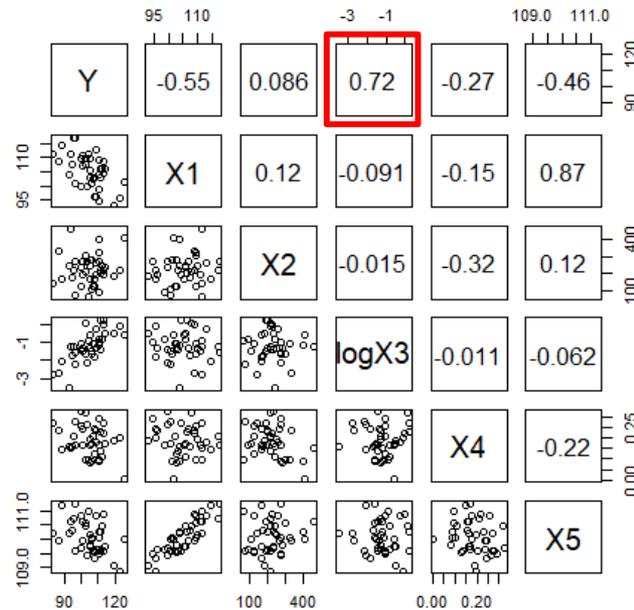
```
plot(Y~X3,data=dados)
```



```
plot(Y~log10(X3),data=dados)
```



```
#linearizando a variável X3  
dados$X3 <- log10(dados$X3)  
names(dados)[4]<-"logX3"  
pairs(dados,upper.panel=upanel)
```



**OBS:** se para linearização fosse necessária a transformação da variável  $Y$ , esta mesma transformação deveria ser adequada para as demais variáveis independentes, caso contrário, não seria possível utilizar esta variável com as demais

# Exemplo

---

Calculando-se o *VIF*...

```
reg<-lm(Y ~ X1 + X2 + logX3 + X4 + X5, data=dados)
library(car)
round(vif(reg),digits=2)
```

• X1	X2	logX3	X4	X5	
• 4.18	1.12	1.01	1.17	4.25	 todos valores < 10

Conclusão: não há evidências de colinearidade

# Exemplo

## Aplicando-se o Stepwise...

```
library(MASS)
reg<-lm(Y ~ X1 + X2 + logX3 + X4 + X5, data=dados)
regsel<-stepAIC(reg,direction="both")
```

- Start: AIC=105.36
- $Y \sim X1 + X2 + \log X3 + X4 + X5$

	Df	Sum of Sq	RSS	AIC	
• - X5	1	5.30	418.04	103.87	← diminui AIC
• - X2	1	9.78	422.52	104.30	
• <none>			412.74	105.36	
• - X1	1	173.12	585.85	117.37	
• - X4	1	303.15	715.89	125.39	
• - logX3	1	1447.39	1860.13	163.58	

- Step: AIC=103.87
- $Y \sim X1 + X2 + \log X3 + X4$

	Df	Sum of Sq	RSS	AIC	
• - X2	1	10.27	428.31	102.84	← diminui AIC
• <none>			418.04	103.87	
• + X5	1	5.30	412.74	105.36	
• - X4	1	298.44	716.48	123.42	
• - X1	1	923.03	1341.06	148.49	
• - logX3	1	1443.42	1861.45	161.61	

- Step: AIC=102.84
- $Y \sim X1 + \log X3 + X4$  ← modelo final

	Df	Sum of Sq	RSS	AIC
• <none>			428.31	102.84
• + X2	1	10.27	418.04	103.87
• + X5	1	5.78	422.52	104.30
• - X4	1	368.85	797.16	125.69
• - X1	1	913.38	1341.69	146.51
• - logX3	1	1440.77	1869.08	159.77

# Exemplo

## Resumo do modelo selecionado:

summary(regsel)

- Residuals:

- Min 1Q Median 3Q Max
- -7.4194 -1.7517 0.1031 2.7467 5.3421

- Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	207.82477	10.09345	20.590	< 2e-16 ***
X1	-0.81676	0.09322	-8.762	1.87e-10 ***
logX3	7.51212	0.68264	11.005	4.53e-13 ***
X4	-43.25904	7.76925	-5.568	2.63e-06 ***

← todos os coeficientes são significativos

- ---

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

- Residual standard error: 3.449 on 36 degrees of freedom
- Multiple R-squared: 0.8683, Adjusted R-squared: 0.8573
- F-statistic: 79.1 on 3 and 36 DF, p-value: 6.513e-16

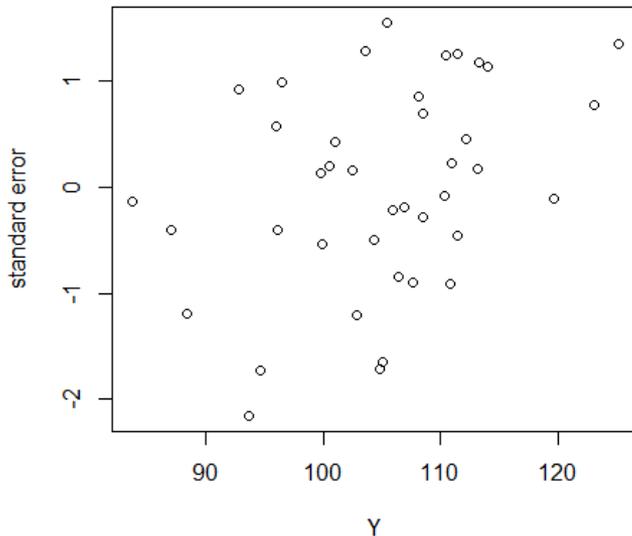
# Exemplo

## Avaliando a qualidade do modelo selecionado

```
shapiro.test(regsel$residuals)
```

- Shapiro-Wilk normality test
  - data: regsel\$residuals
  - $W = 0.96599$ , p-value = 0.2668
- ← resíduos são normalmente distribuídos

```
erro padr <- (summary(regsel))$sigma  
plot(dados$Y, regsel$residuals/erro padr, xlab="Y", ylab="standard error")
```



Aparentemente nenhum *outlier*  
(|erro padronizado| > 2,5)  
Valores de  $Y$  menores que 90 e maiores que 115 foram pouco amostrados!

```
library(lmtest)  
bptest(regsel)
```

- studentized Breusch-Pagan test
- data: regsel
- BP = 4.3227, df = 3, p-value = 0.2287

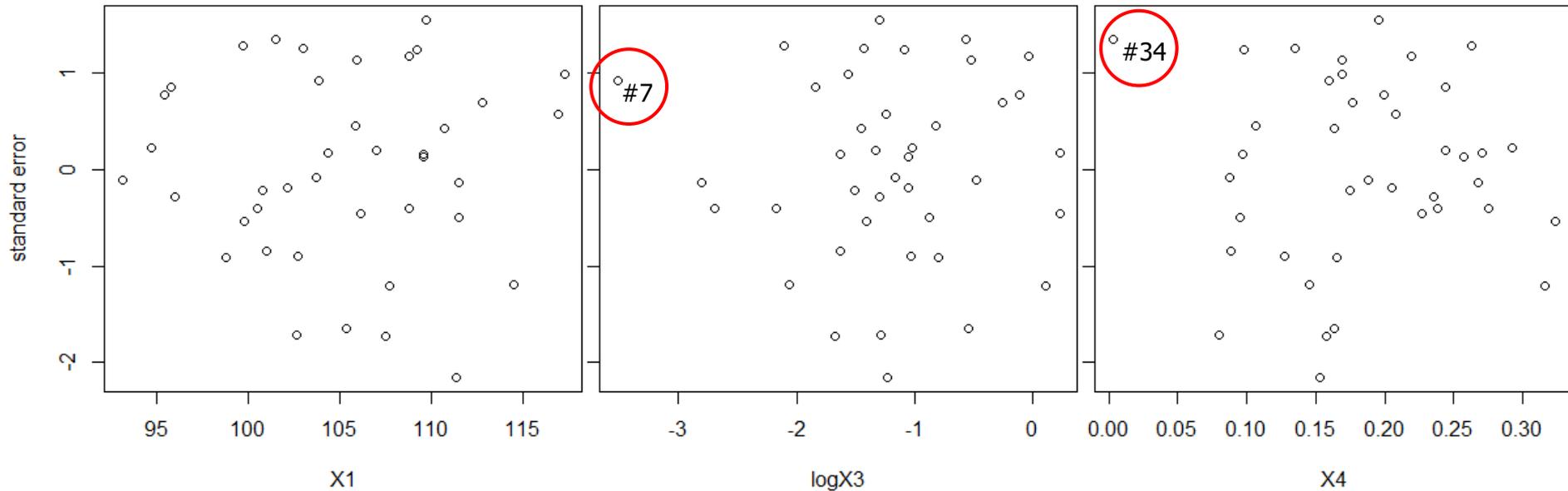
↑  
variância é constante

# Exemplo

## Avaliando a qualidade do modelo selecionado

```
plot(dados$X1,regsel$residuals/erropadr,xlab="X1",ylab="standard error")  
plot(dados$logX3,regsel$residuals/erropadr,xlab="logX3",ylab="standard error")  
plot(dados$X4,regsel$residuals/erropadr,xlab="X4",ylab="standard error")
```

### Variáveis selecionadas

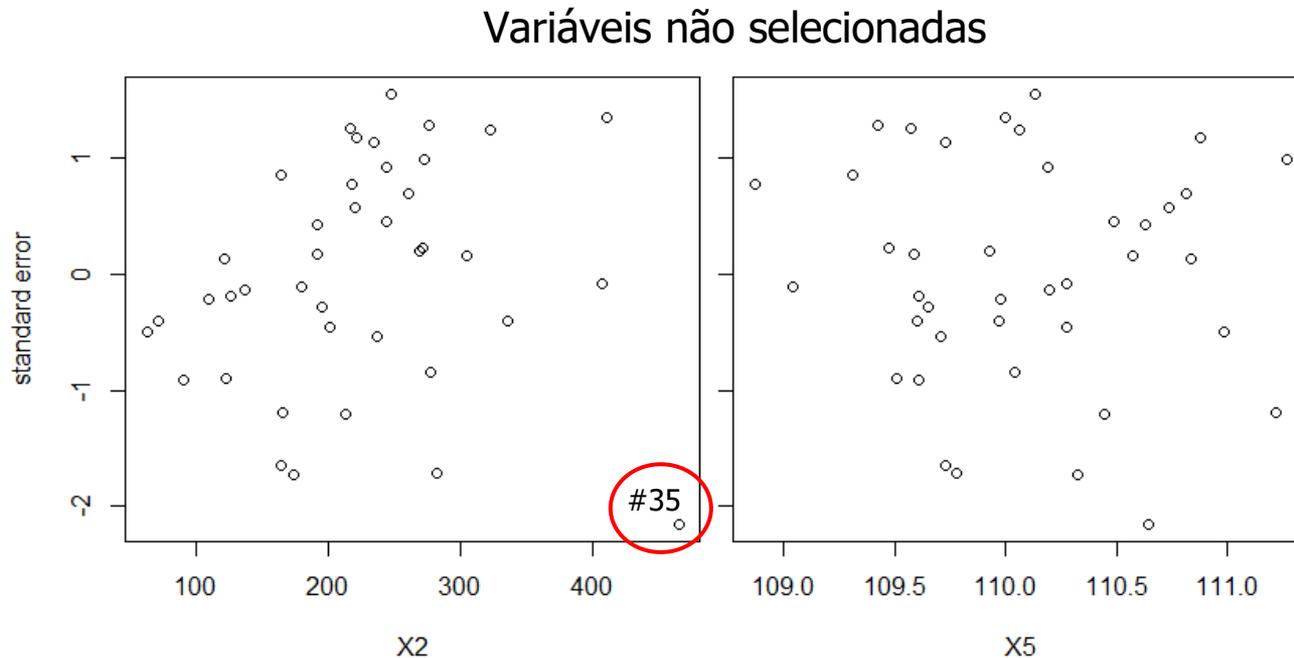


Alguns pontos parecem estar isolados e com erros mais altos: vale a pena uma investigação se podem estar influenciando os resultados...

# Exemplo

É importante também avaliar os resíduos em relação às variáveis descartadas pela seleção

```
plot(dados$X2,regsel$residuals/erropadr,xlab="X2",ylab="standard error")  
plot(dados$X5,regsel$residuals/erropadr,xlab="X5",ylab="standard error")
```



A variável X2 parece ter alguma relação com o erro caso a amostra isolada fosse descartada

# Exemplo

Eliminando-se cada uma das amostras e verificando o impacto no AIC e do modelo final

```
#dados completos
reg<-lm(Y ~ X1 + X2 + logX3 + X4 + X5, data=dados)
regsel<-stepAIC(reg,direction="both",trace=FALSE)
print(paste("Dados Completos  ", " AIC:", extractAIC(regsel)," modelo final:",deparse(formula(regsel))))
#dados sem amostras
for (i in c(7,34,35)) {
  reg<-lm(Y ~ X1 + X2 + logX3 + X4 + X5, data=dados[-i,])
  regsel<-stepAIC(reg,direction="both",trace=FALSE)
  print(paste("Dados sem amostra",i," AIC:", extractAIC(regsel)," modelo final:",deparse(formula(regsel))))
}
```

Dados	AIC	Modelo final
Completo	102,84	$Y \sim X1 + \log X3 + X4$
sem #7	100,21	$Y \sim X1 + \log X3 + X4$
sem #34	98,79	$Y \sim X1 + \log X3 + X4$
sem #35	91,79	$Y \sim X1 + X2 + \log X3 + X4$

Repetir toda a análise sem a amostra 35!

 Além de diminuir o AIC, acrescentou a variável X2

# Exemplo

```
reg<-lm(Y ~ X1 + X2 + logX3 + X4 + X5, data=dados[-35,]) ← sem a amostra 35
regsel<-stepAIC(reg,direction="both",trace=FALSE)
summary(regsel)
```

- Residuals:

Min	1Q	Median	3Q	Max
-6.5691	1.6460	0.3935	2.2728	4.4922

- Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	200.665796	9.189850	21.836	< 2e-16 ***
X1	-0.786335	0.083630	-9.403	5.52e-11 ***
X2	0.015347	0.006468	2.373	0.0235 *
logX3	7.548661	0.604932	12.479	3.03e-14 ***
X4	-38.551520	7.273529	-5.300	7.00e-06 ***

← significativo apenas a 5%!

- 

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

- Residual standard error: 3.056 on 34 degrees of freedom
- Multiple R-squared: 0.8983, Adjusted R-squared: 0.8863
- F-statistic: 75.04 on 4 and 34 DF, p-value: 2.2e-16

# Exemplo

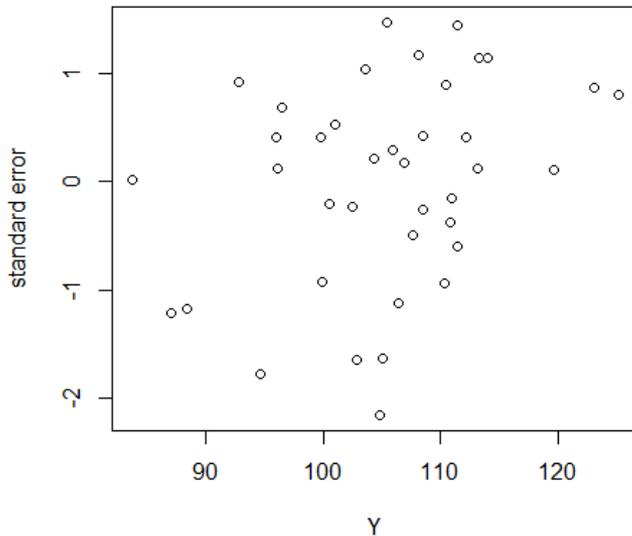
## Avaliando a qualidade do modelo selecionado

```
shapiro.test(regsel$residuals)
```

- Shapiro-Wilk normality test
  - data: regsel\$residuals
  - $W = 0.95815$ , p-value = 0.1547
- ← resíduos são normalmente distribuídos

```
erropadr <- (summary(regsel))$sigma
```

```
plot(dados[-35,]$Y, regsel$residuals/erropadr, xlab="Y", ylab="standard error")
```



Não mudou muito!

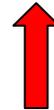
Aparentemente nenhum *outlier*  
(|erro padronizado| > 2,5)

Valores de  $Y$  menores que 90 e maiores que 115 foram pouco amostrados!

```
library(lmtest)
```

```
bptest(regsel)
```

- studentized Breusch-Pagan test
- data: regsel
- BP = 2.0092, df = 4, p-value = 0.7341

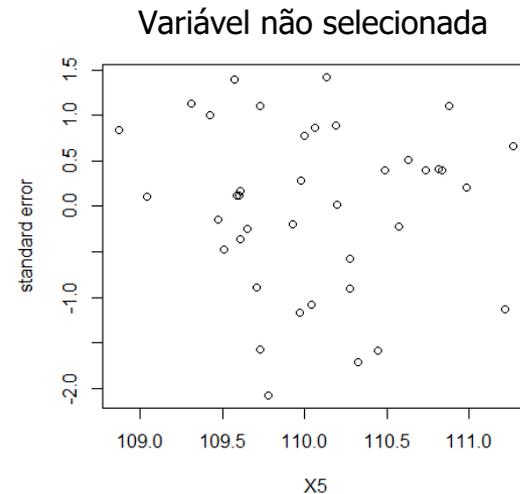
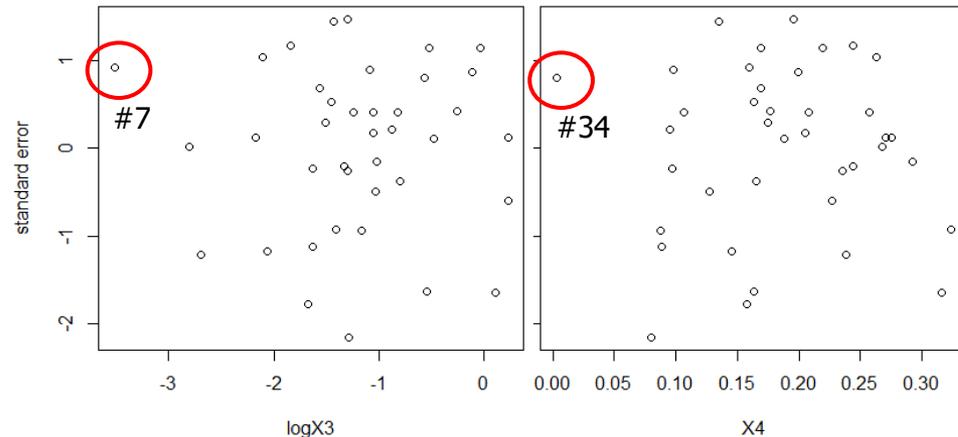
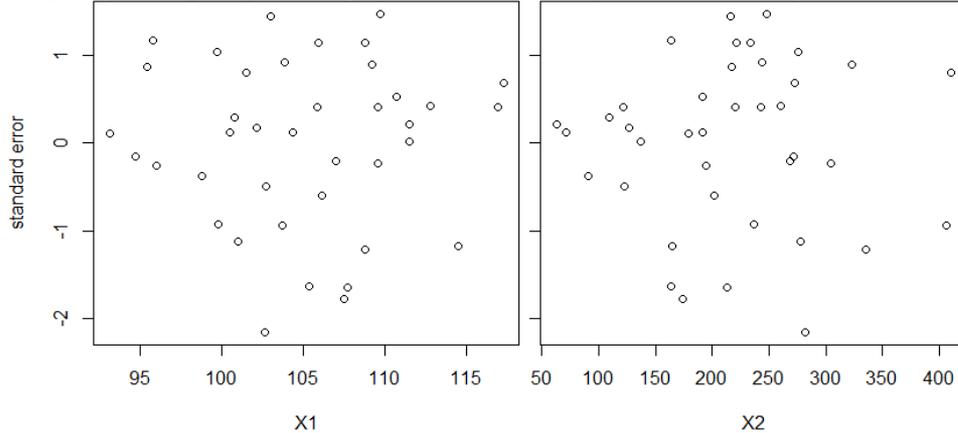


variância é constante

# Exemplo

## Avaliando a qualidade do modelo selecionado

```
plot(dados[-35,]$X1,regsel$residuals/erropadr,xlab="X1",ylab="standard error")  
plot(dados[-35,]$X2,regsel$residuals/erropadr,xlab="X2",ylab="standard error")  
plot(dados[-35,]$logX3,regsel$residuals/erropadr,xlab="logX3",ylab="standard error")  
plot(dados[-35,]$X4,regsel$residuals/erropadr,xlab="X4",ylab="standard error")  
plot(dados[-35,]$X5,regsel$residuals/erropadr,xlab="X5",ylab="standard error")
```



As amostras 7 e 34 ainda continuam isoladas  
Vale a pena descartá-las também?

# Exemplo

Ao se eliminar as amostras 7, 34 e 35, verificou-se que o modelo selecionado pelo *stepwise* incluiu a variável X2 mas a mesma não foi significativa a 5%

```
#dados sem as amostras 7, 34 e 35
regsel<-lm(Y ~ X1 + logX3 + X4, data=dados[-c(7,34,35),])
summary(regsel)
```

- Residuals:

- Min 1Q Median 3Q Max
- -5.8101 -1.9095 0.3236 1.7606 4.8961

- Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	198.295309	9.362345	21.180	< 2e-16 ***
X1	-0.763356	0.085146	-8.965	3.06e-10 ***
X2	0.013021	0.006799	1.915	0.0645 . ← não significativo!
logX3	7.819945	0.688386	11.360	9.17e-13 ***
X4	-35.337544	7.770705	-4.548	7.36e-05 ***

- ---

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

- Residual standard error: 3.051 on 32 degrees of freedom
- Multiple R-squared: 0.884 Adjusted R-squared: 0.8695
- F-statistic: 60.95 on 4 and 32 DF, p-value: 1.633e-14

# Exemplo

Assim, optou-se por ajustar o modelo apenas com as variáveis X1, logX3 e X4

```
#dados sem as amostras 7, 34 e 35
regsel<-lm(Y ~ X1 + logX3 + X4, data=dados[-c(7,34,35),])
summary(regsel)
```

- Residuals:

Min	1Q	Median	3Q	Max
-5.7769	-1.9588	-0.1146	1.8759	5.0786

- Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	200.32427	9.67089	20.714	< 2e-16 ***
X1	-0.75235	0.08832	-8.519	7.63e-10 ***
logX3	7.77793	0.71531	10.874	1.91e-12 ***
X4	-37.90940	7.95713	-4.764	3.69e-05 ***

 todos os coeficientes são significativos

- ---

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

- Residual standard error: 3.172 on 33 degrees of freedom
- Multiple R-squared: 0.8707                      Adjusted R-squared: 0.8589
- F-statistic: 74.06 on 3 and 33 DF, p-value: 9.663e-15

# Exemplo

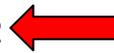
## Avaliando a qualidade do modelo selecionado

```
shapiro.test(regsel$residuals)
```

- Shapiro-Wilk normality test

- data: regsel\$residuals

- W = 0.96652, p-value = 0.3222



resíduos são normalmente distribuídos

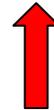
```
library(lmtest)
```

```
bptest(regsel)
```

- studentized Breusch-Pagan test

- data: regsel

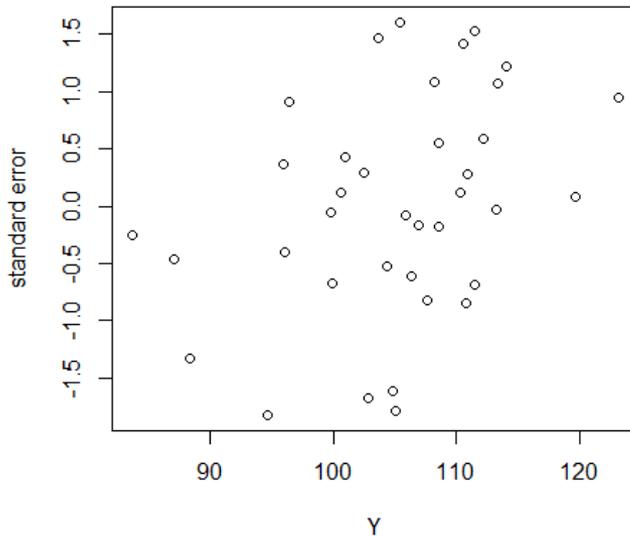
- BP = 1.0795, df = 3, p-value = 0.782



variância é constante

```
erropadr <- (summary(regsel))$sigma
```

```
plot(dados[-c(7,34,35),]$Y,regsel$residuals/erropadr,xlab="Y",ylab="standard error")
```



Não mudou muito!

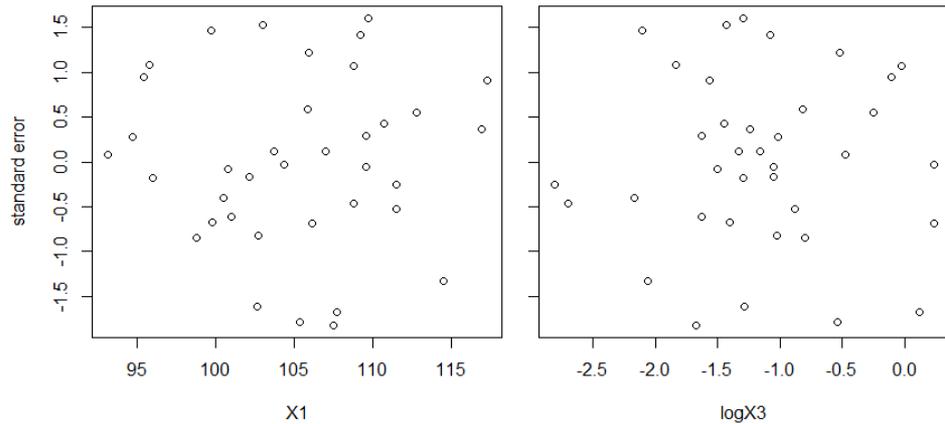
Aparentemente nenhum *outlier*  
(|erro padronizado| > 2,5)

Valores de *Y* menores que 90 e maiores que 115 foram pouco amostrados!

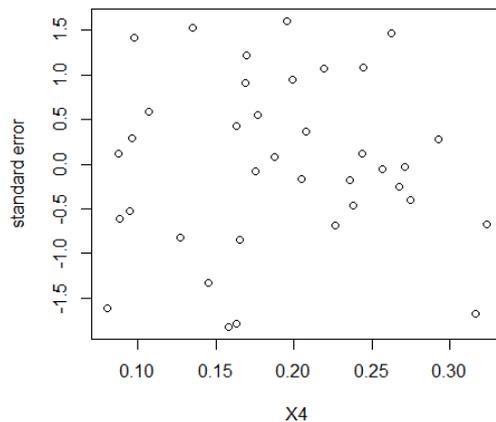
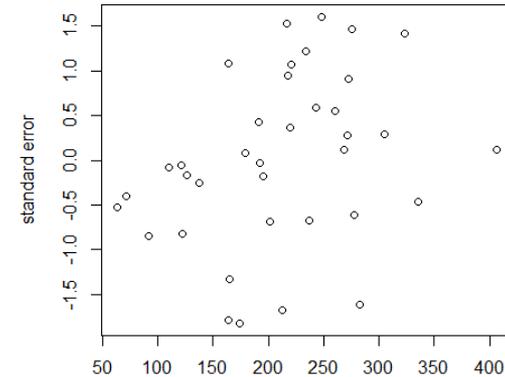
# Exemplo

## Avaliando a qualidade do modelo selecionado

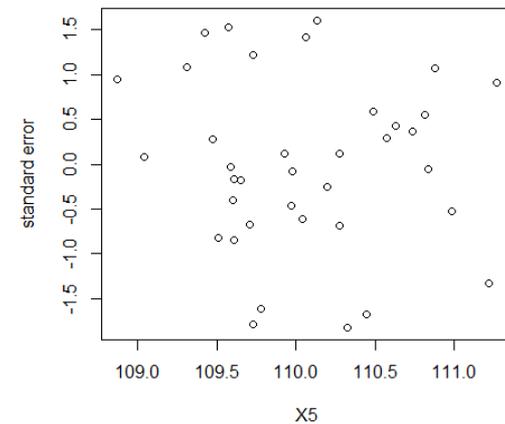
```
plot(dados[-c(7,34,35),]$X1,regsel$residuals/erropadr,xlab="X1",ylab="standard error")  
plot(dados[-c(7,34,35),]$logX3,regsel$residuals/erropadr,xlab="logX3",ylab="standard error")  
plot(dados[-c(7,34,35),]$X4,regsel$residuals/erropadr,xlab="X4",ylab="standard error")  
plot(dados[-c(7,34,35),]$X2,regsel$residuals/erropadr,xlab="X2",ylab="standard error")  
plot(dados[-c(7,34,35),]$X5,regsel$residuals/erropadr,xlab="X5",ylab="standard error")
```



## Variáveis não selecionadas



Aparentemente  
tudo OK



# Exemplo

Afinal, qual modelo escolher?

Modelo	$Y \sim X1 + \log X3 + X4$	$Y \sim X1 + X2 + \log X3 + X4$	$Y \sim X1 + \log X3 + X4$
Dados	completo	sem #35	sem #7, #34 e #35
AIC	102,84	91,79	89,19
$r_a^2$	0,8573	0,8863	0,8589
Intercepto	207,8248	200,6658	200,3243
X1	-0,8168	-0,7863	-0,7524
X2		0,0153	
logX3	7,5121	7,5487	7,7779
X4	-43,2590	-38,5515	-37,9094

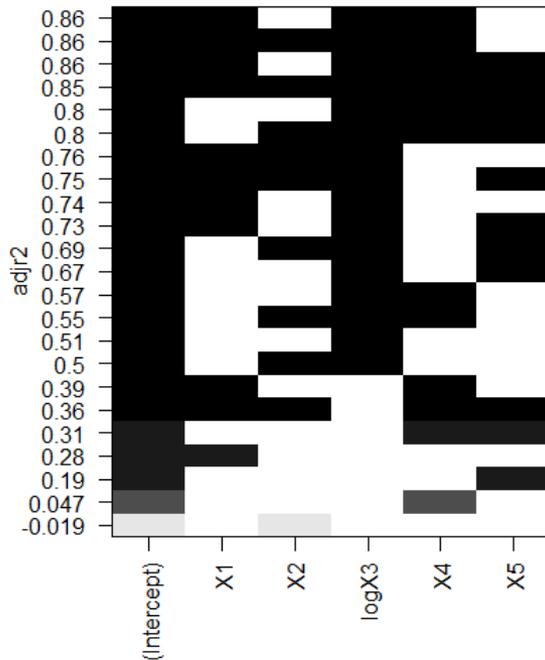
Todos os modelos apresentaram desempenho semelhante com ligeira vantagem (maior  $r_a^2$ ) do modelo com a variável X2 e sem a amostra 35

O último modelo apresentou o menor AIC e a análise dos resíduos não indicou nenhum problema importante. No entanto, esse modelo só poderia ser adotado se a eliminação das 3 amostras fosse justificada

# Exemplo

Qual melhor modelo usando a busca exaustiva?

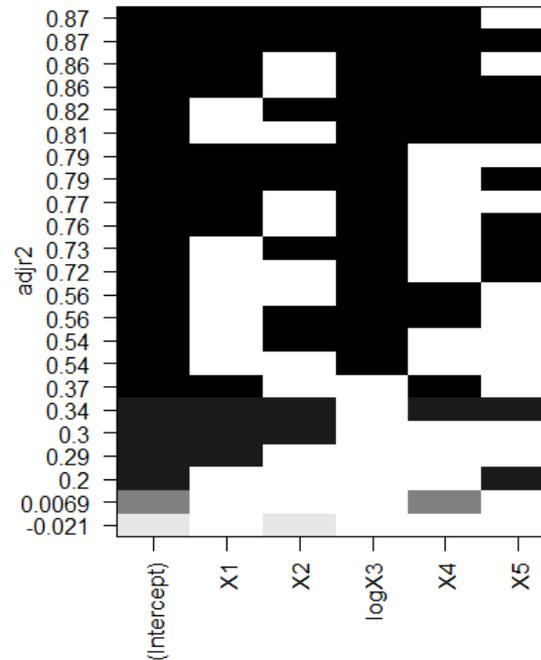
```
library(leaps)  
regsets<-regsubsets(Y ~ X1 + X2 + logX3 + X4 + X5,  
  data=dados,nbest=6)  
plot(regsets,scale="adjr2")
```



Melhor Modelo:

$$Y \sim \beta_0 + \beta_1 X_{1,i} + \beta_3 \log X_{3,i} + \beta_4 X_{4,i} + \xi_i$$

```
library(leaps)  
regsets<-regsubsets(Y ~ X1 + X2 + logX3 + X4 + X5,  
  data=dados[-c(7,34,35),],nbest=6)  
plot(regsets,scale="adjr2")
```



Melhor Modelo:

$$Y \sim \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \beta_3 \log X_{3,i} + \beta_4 X_{4,i} + \xi_i$$

**OBS:** Não avalia a significância dos coeficientes dos modelos!

# Considerações Finais

---

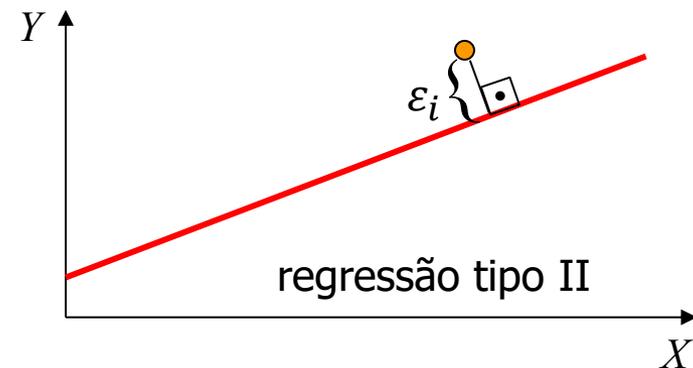
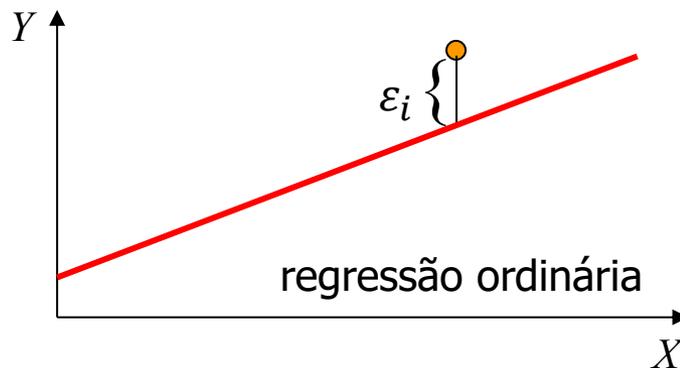
- A análise inicia-se pela avaliação dos diagramas de dispersão de cada variável independente versus dependente, buscando-se anomalias (*outliers*) e/ou falta de linearidade nas relações.
- A análise dos diagramas de dispersão entre variáveis independentes pode ser negligenciada mas pode indicar a presença de colinearidade. Neste caso, algumas variáveis já podem ser provisoriamente descartadas nesta fase. Ao final da construção do modelo, é sempre útil testar se essas variáveis descartadas realmente não contribuem na explicação da variável dependente.
- Métodos automáticos de busca (*stepwise*, exaustivo) podem ser utilizados para encontrar o “melhor” modelo. Como estes métodos são “sub-ótimos”, diferentes métodos podem selecionar diferentes modelos.
- A análise de resíduos para detecção de *outliers* e não-linearidade, e testes formais de normalidade e homoscedasticidade dos resíduos devem ser feitos ao final do processo de busca do melhor modelo. Qualquer intervenção (retirada de *outliers*, transformação de variáveis, inclusão de novas variáveis independentes, inclusão de interação entre variáveis, etc) faz com que todo o processo tenha que ser refeito.
- **NÃO** é necessário (**é irrelevante!**) testar a normalidade das variáveis dependente e independentes. Testes de normalidade são feitos sempre sobre os resíduos.

# Outras abordagens...

- Regressão linear tipo II (*Model II regression*)

Nesse tipo de regressão, considera-se que também as variáveis independentes sejam variáveis aleatórias

Nesse caso, os erros (ou desvios) podem ser medidos ao longo da perpendicular (ou normal) à linha de regressão. Dessa forma, os coeficientes são estimados minimizando-se a soma dos quadrados dos desvios normais



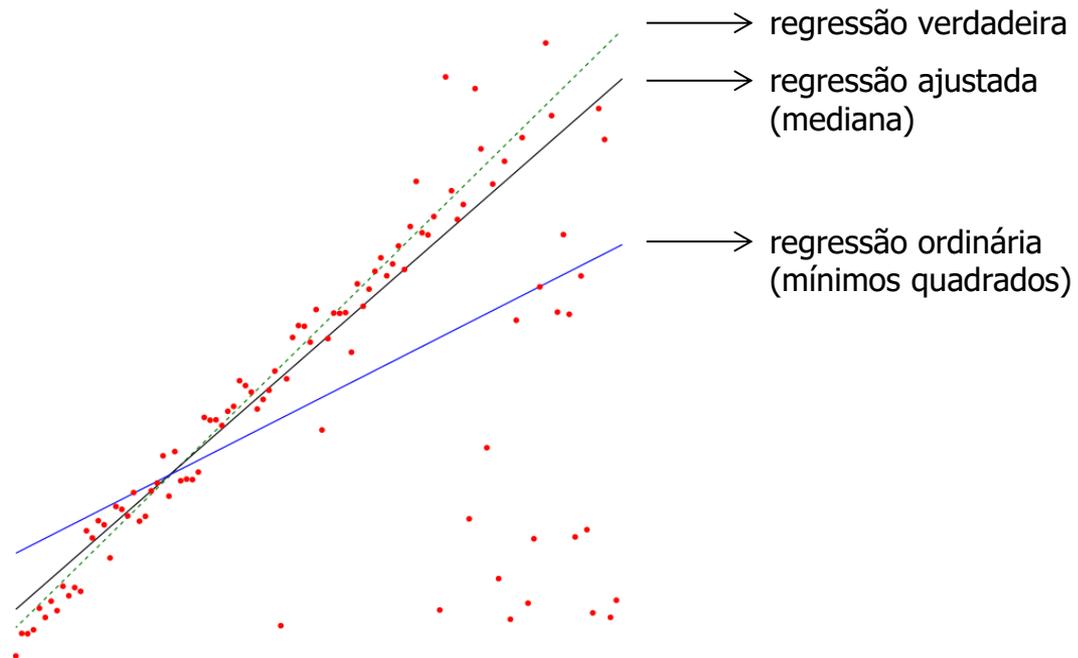
- Regressão não linear

Nesse caso, a estimação dos coeficientes da regressão é feita por métodos de aproximações sucessivas, buscando-se minimizar alguma função de erro

Diferentes métodos podem resultar em diferentes soluções e podem chegar a resultados sub-ótimos

# Outras abordagens...

- Regressão baseada no estimador de Theil-Sen\* (método da mediana simples)  
Nessa regressão, o coeficiente angular  $b_1$  é calculado através da mediana dos coeficientes angulares de todos os pares de pontos da amostra  
Uma vez calculado o coeficiente angular  $b_1$ , o coeficiente linear  $b_0$  é estimado através da mediana de  $Y_i - b_1 X_i$  para todos os pontos da amostra  
Esse tipo de regressão é menos sensível a presença de *outliers*

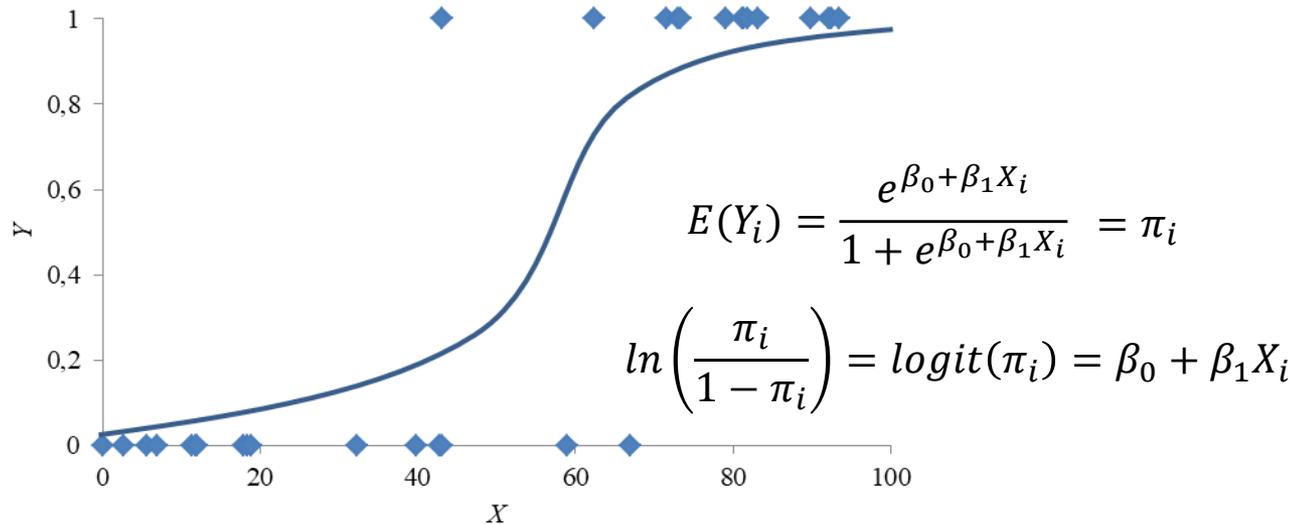


\* [https://en.wikipedia.org/wiki/Theil%E2%80%93Sen\\_estimator](https://en.wikipedia.org/wiki/Theil%E2%80%93Sen_estimator)

# Outras abordagens...

- Variável dependente binária:  $Y \sim \text{Bernoulli}$ ,  $P(Y_i = 1) = \pi_i$

Regressão Logística



- Variável dependente que represente proporção  $Y_i = p_i = [0,1]$

Regressão Beta

funções de ligação  $g(E(Y_i))$ : logit e probit

$$\text{logit}(p_i) = \ln\left(\frac{p_i}{1 - p_i}\right) \quad \text{probit}(p_i) = F^{-1}(p_i) = z_i \quad p_i = P(Z < z_i)$$

# Outras abordagens...

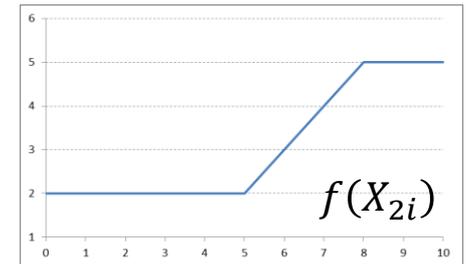
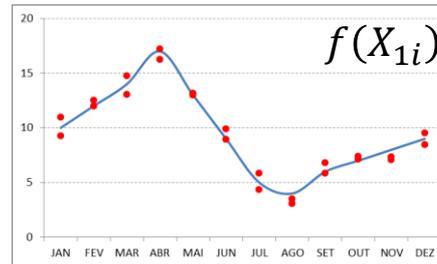
- Modelos Lineares Generalizados

É uma flexibilização da regressão linear ordinária para permitir que a variável dependente tenha resíduos com distribuição diferente da gaussiana

Inclui a Regressão Ordinária (clássica), Regressão Logística, Regressão Beta, Regressão de Poisson, Regressão Multinomial

Também inclui Modelos Aditivos Generalizados (GAM) que permitem o uso de funções de suavização (p.ex. médias móveis)

$$g(E(Y_i)) = \beta_0 + f(X_{1i}) + f(X_{2i}) + \varepsilon_i$$



- Modelos autoregressivos com (ARMAX) ou sem variáveis exógenas (ARMA)

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \beta_3 X + \xi$$

- Regressão espacial

*Spatial Lag Models* (SAR): atribuem a autocorrelação espacial à variável dependente

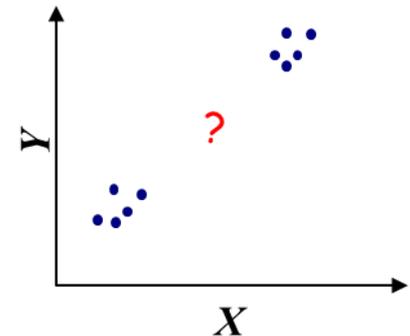
*Spatial Error Models* (CAR): atribuem a autocorrelação espacial ao erro

# Tamanho de amostra para Regressão

Assim como em outras análises, determinar o tamanho da amostra conveniente evita o desperdício de tempo, força de trabalho, custos, etc. no processo de coleta e análise de dados.

Se a amostra for muito pequena, pode-se chegar a resultados aparentemente muito bons ( $r^2$  muito altos) mas ainda assim não significativos, indicando que o modelo ajustado poderia ter sido obtido casualmente.

Como vimos, tão importante como o tamanho da amostra, é sua distribuição com relação a cada variável analisada. A presença de "buracos" de observações podem prejudicar a análise dos resultados.



Apesar de não existir uma regra prática, muitos pesquisadores alegam que deve haver pelo menos 10 observações por variável independente. Por exemplo, se estivermos usando 4 variáveis independentes, deveríamos ter um tamanho mínimo de amostra de 40.

Pode-se encontrar na literatura diversos trabalhos que discutem o assunto, propondo fórmulas para o cálculo do tamanho da amostra mas o mais importante é concentrar-se em representar bem o fenômeno estudado e, caso os resultados não pareçam ser adequados, investigar se a causa não pode ser a amostragem insuficiente.