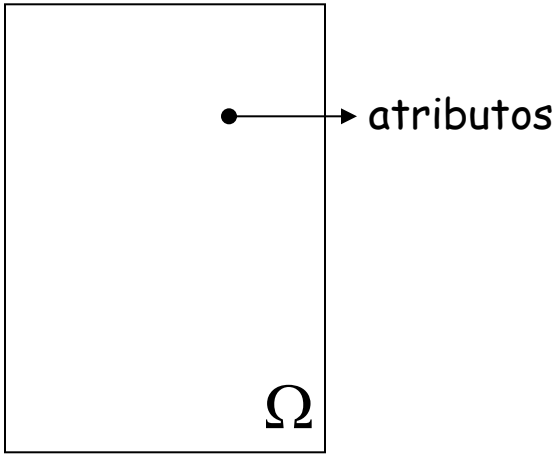
Estatística: Aplicação ao Sensoriamento Remoto

SER 204 - ANO 2023

Análise de Regressão

Camilo Daleles Rennó
camilo.renno@inpe.br
http://www.dpi.inpe.br/~camilo/estatistica/

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 (Ω)

atributos quantitativos:
. altura total
. diâmetro da copa
. diâmetro do tronco (DAP)
. biomassa
. etc

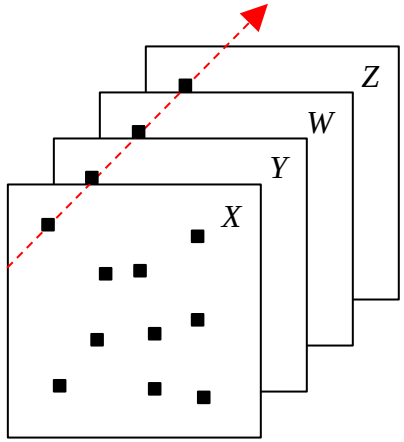
Estes atributos constituem as variáveis em estudo.

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

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

Relacionamento entre Variáveis

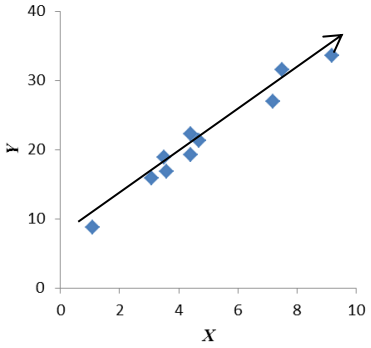
Em imagens ou mapas, o relacionamento aparece pela posição geográfica

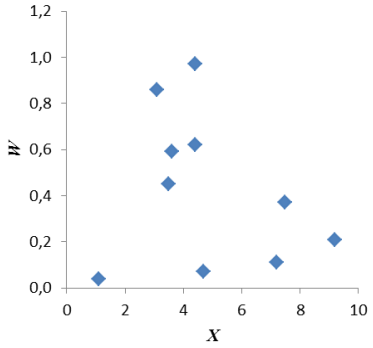


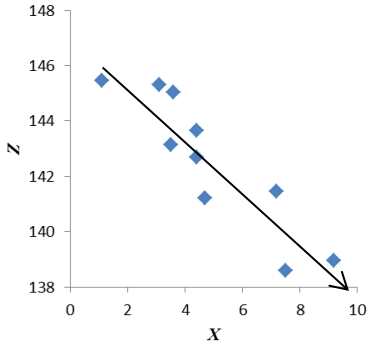
Amostra	X	Y	W	Z
1	3,5	18,9	0,45	143,2
2	7,5	31,5	0,37	138,6
3	4,4	22,2	0,62	142,7
4	1,1	8,7	0,04	145,5
5	4,4	19,2	0,97	143,7
6	4,7	21,3	0,07	141,2
7	7,2	27,0	0,11	141,5
8	3,6	16,8	0,59	145,1
9	9,2	33,6	0,21	139,0
10	3,1	15,9	0,86	145,3

← mesma posição geográfica

Diagrama de dispersão





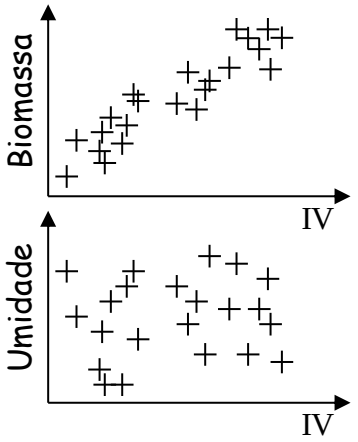


Muitos estudos buscam entender as relações de dependência entre variáveis de modo a construir modelos que permitam prever o comportamento de uma variável conhecendo-se os valores de outra ou outras variáveis

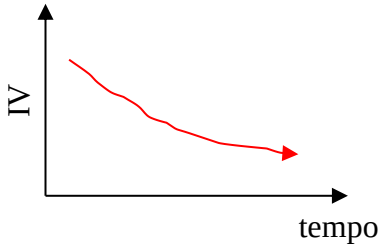
Relacionamento entre Variáveis

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?



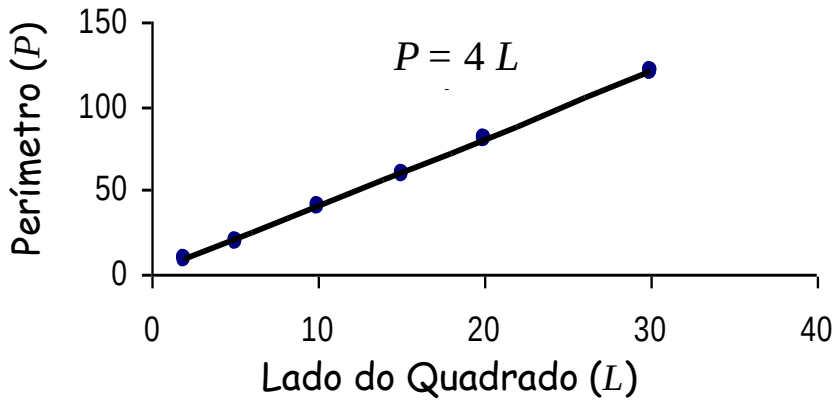
Quanto à biomassa, espera-se que tenha havido uma diminuição
Quanto à umidade, nada podemos afirmar

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

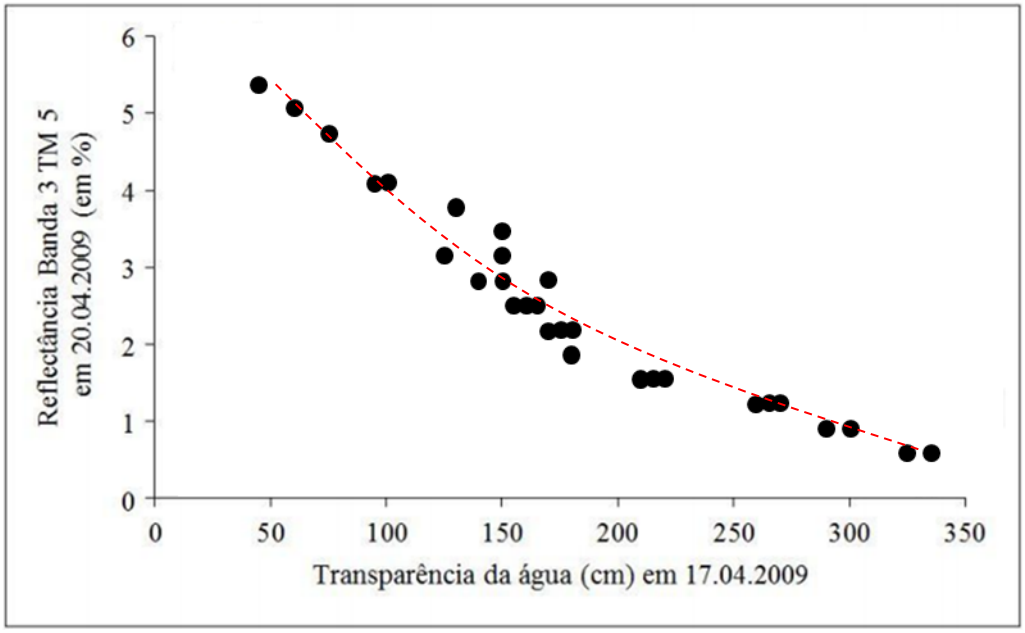


Todos os pontos caem perfeitamente sobre a linha que representa a relação funcional entre L e P

Relação funcional x Relação estatística

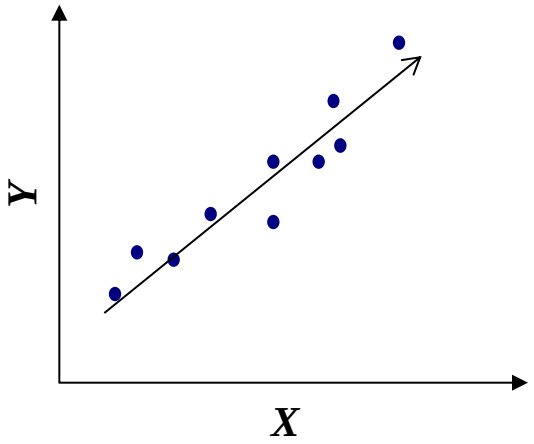
2) Estatística: não há uma relação perfeita como no caso da relação funcional.
As observações em geral não caem exatamente na linha que descreve a relação.

Ex: relação entre transparência da água e a reflectância na banda 3 TM5

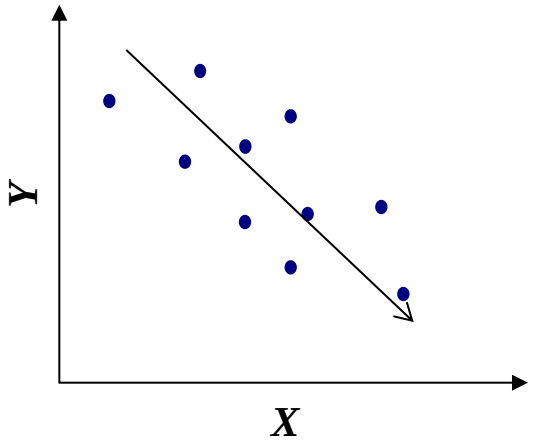


Fonte: Adaptado de Santos, F.C.; Pereira Filho, W.; Toniolo, G.R.. Transparência associada à reflectância da água do reservatório Passo Real. In: XVII SBSR, 2015. p. 6653-6659

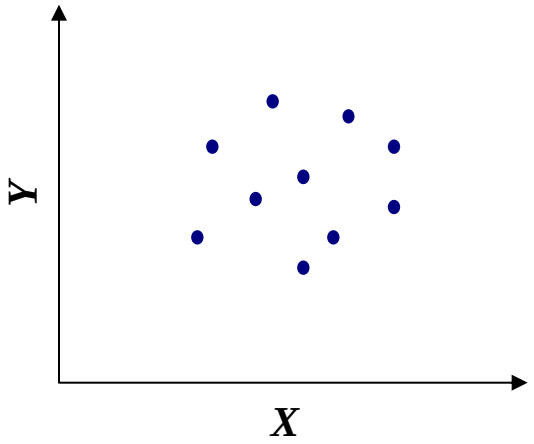
Grau de Relacionamento



Relação direta ou positiva



Relação inversa ou negativa



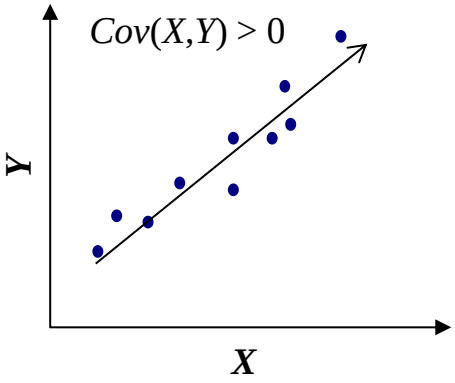
Ausência de relação

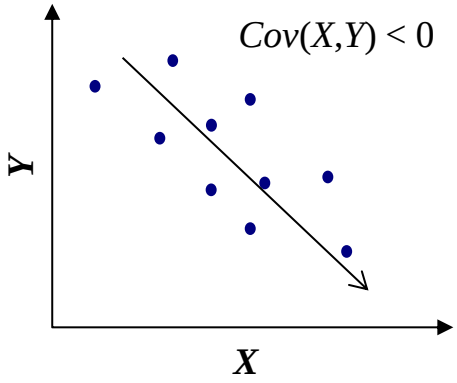
Como caracterizar o grau de relacionamento (ou associação) entre pares de variáveis?

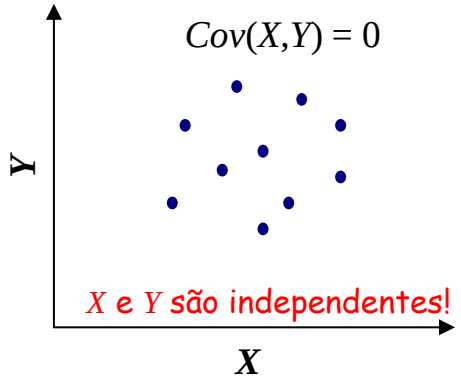
Covariância
Coeficiente de Correlação

Covariância

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



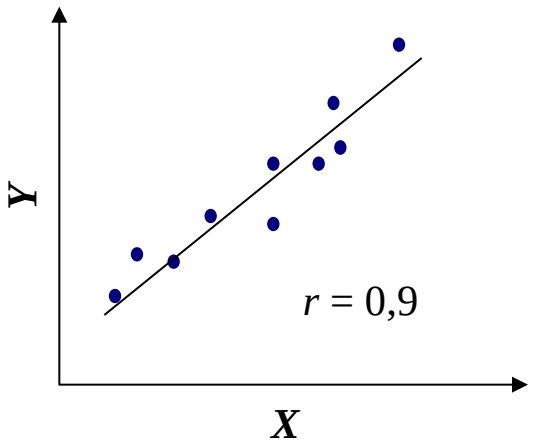


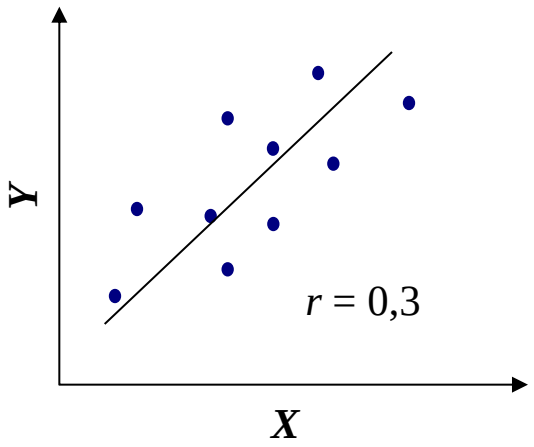


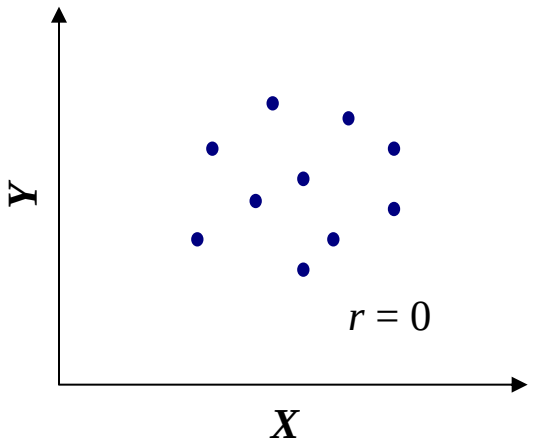
Quanto maior a covariância (em módulo), mais próximos estarão os pontos entorno da reta que representa a tendência principal da nuvem de pontos

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

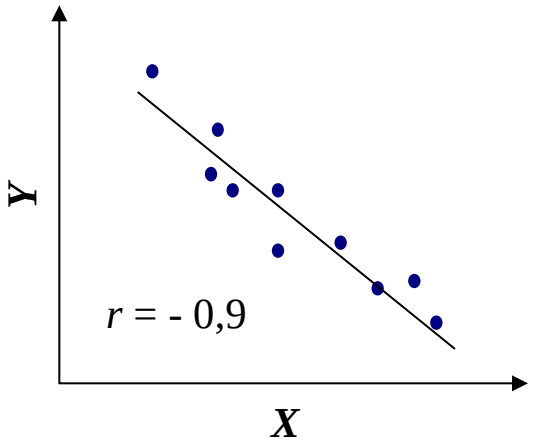
Coeficiente de Correlação







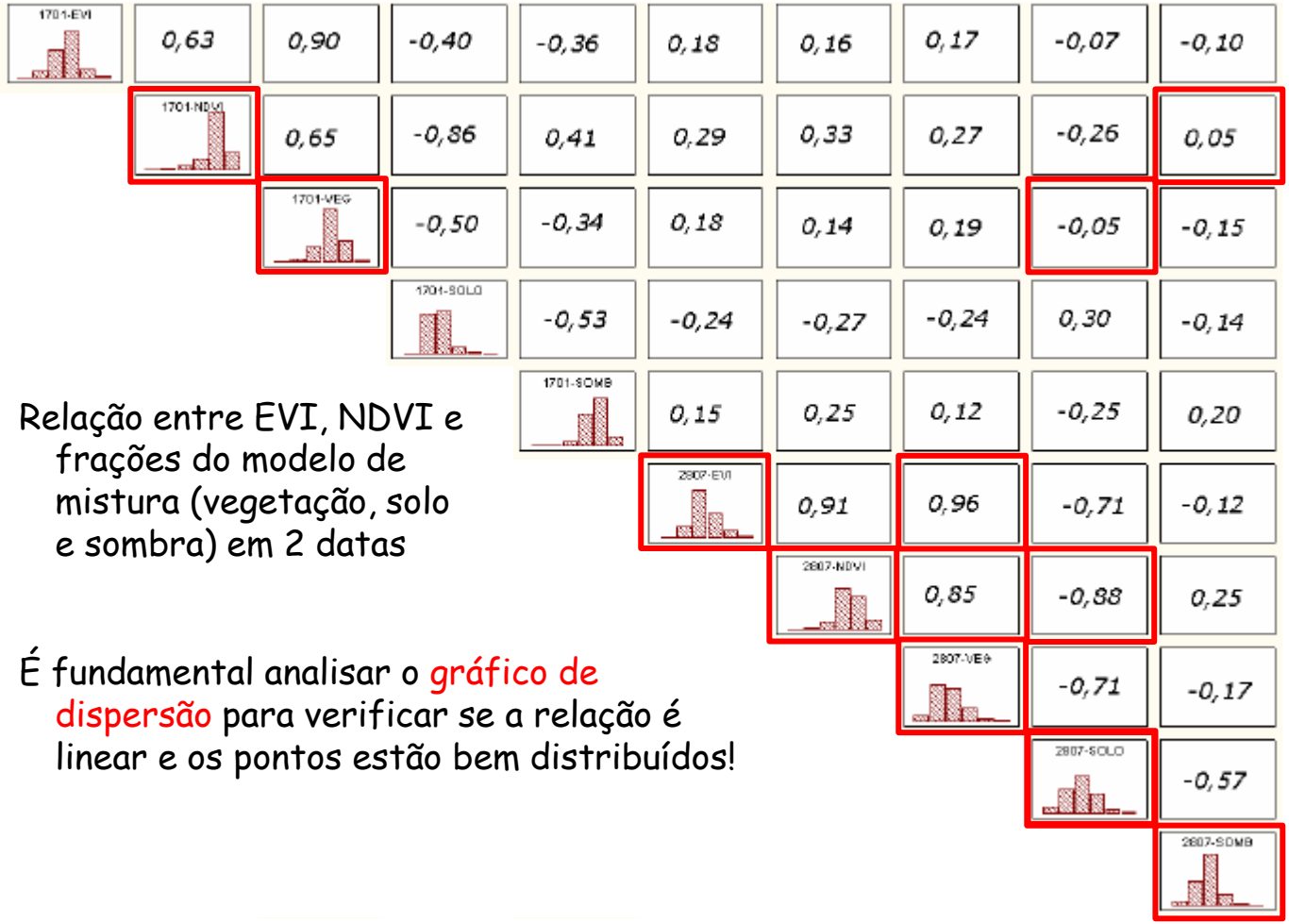
Coeficiente de Correlação (de Pearson)
mede o grau de relação linear entre X e Y



$$r = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}} \qquad -1 \le r \le 1 \quad \text{(adimensional)}$$

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

Coeficiente de Correlação

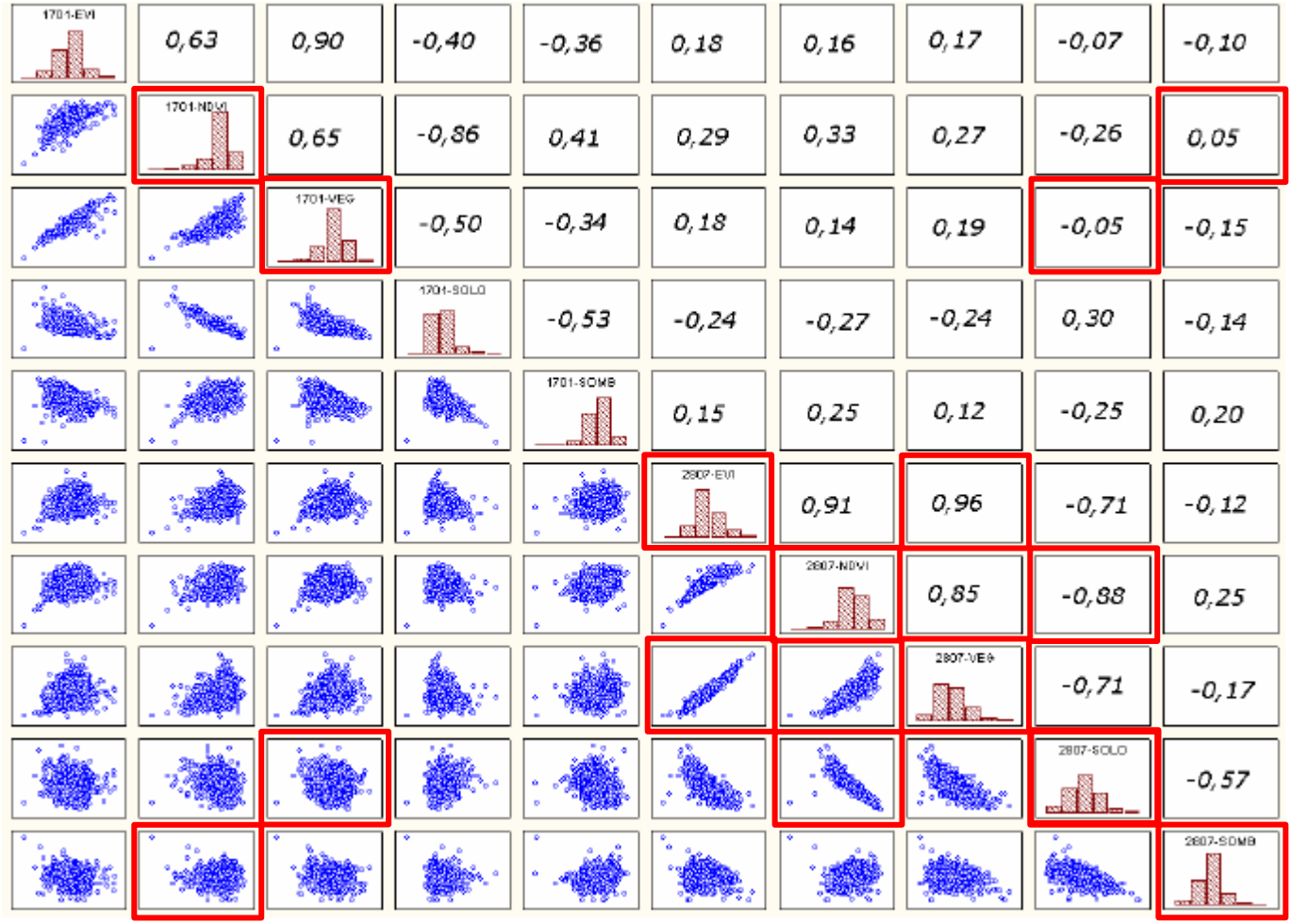


Relação entre EVI, NDVI e frações do modelo de mistura (vegetação, solo e sombra) em 2 datas

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

*Fonte: Souza e Guimarães et al. (2005) - http://marte.sid.inpe.br/col/ltid.inpe.br/sbsr/2004/11.20.17.44/doc/2665.pdf

Coeficiente de Correlação

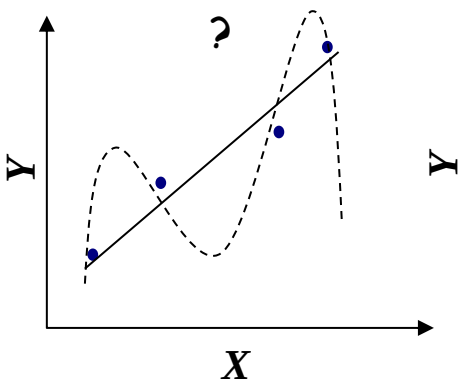


*Fonte: Souza e Guimarães et al. (2005) - http://marte.sid.inpe.br/col/ltid.inpe.br/sbsr/2004/11.20.17.44/doc/2665.pdf

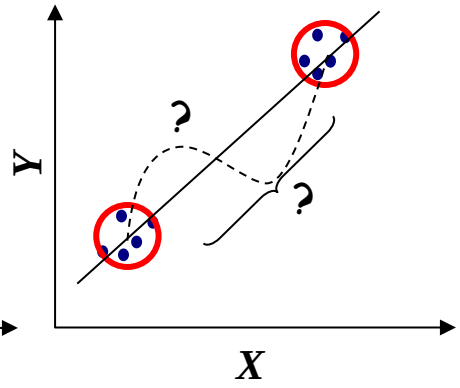
Coeficiente de Correlação

Interpretações errôneas do coeficiente de correlação

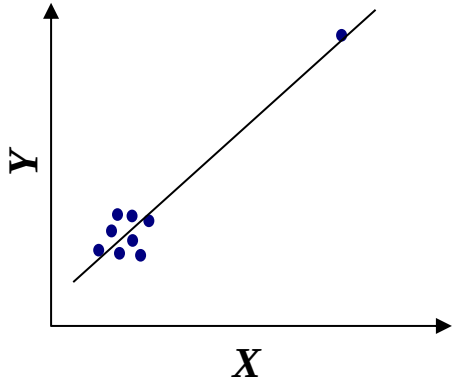
- Um alto coeficiente de correlação (em módulo) nem sempre indica que a equação de regressão estimada está bem ajustada aos dados.



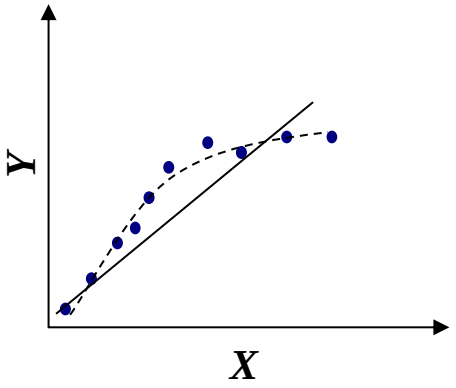
Poucos pontos



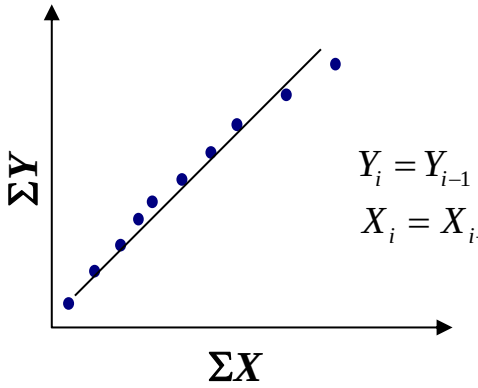
Grupos de pontos



Ponto de alavanca (*leverage*)



Relação quase linear



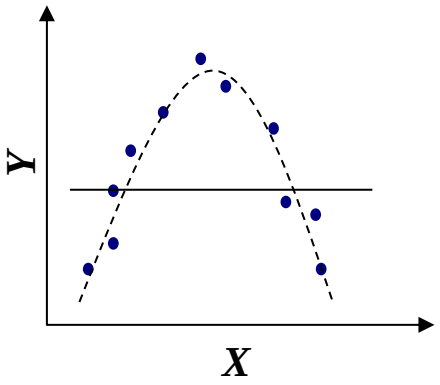
$$Y_i = Y_{i-1} + \Delta y_i \qquad \Delta y_i \geq 0$$
$$X_i = X_{i-1} + \Delta x_i \qquad \Delta x_i \geq 0$$

Variáveis cumulativas

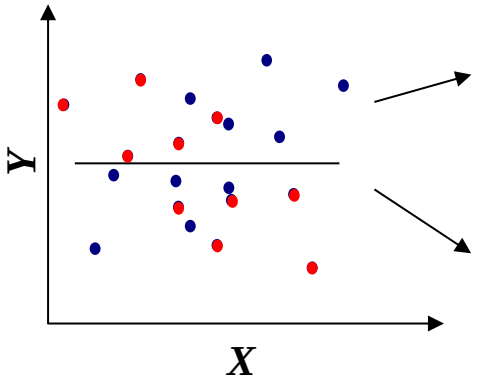
Coeficiente de Correlação

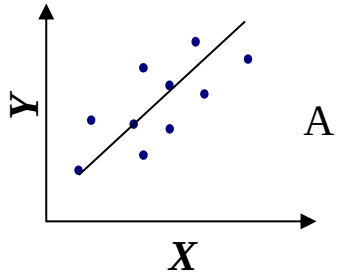
Interpretações errônea do coeficiente de correlação

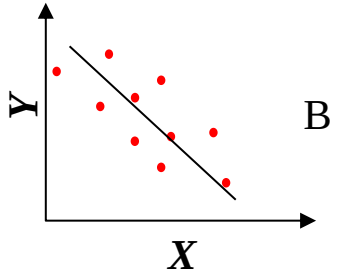
- Um coeficiente de correlação próximo de zero nem sempre indica que X e Y não são relacionadas.



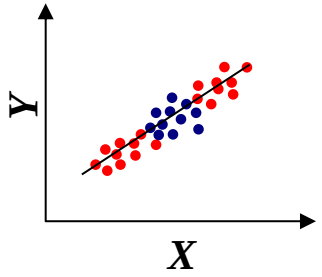
Relação não linear



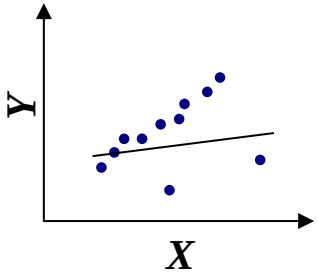




Mistura de grupos com relações diferentes



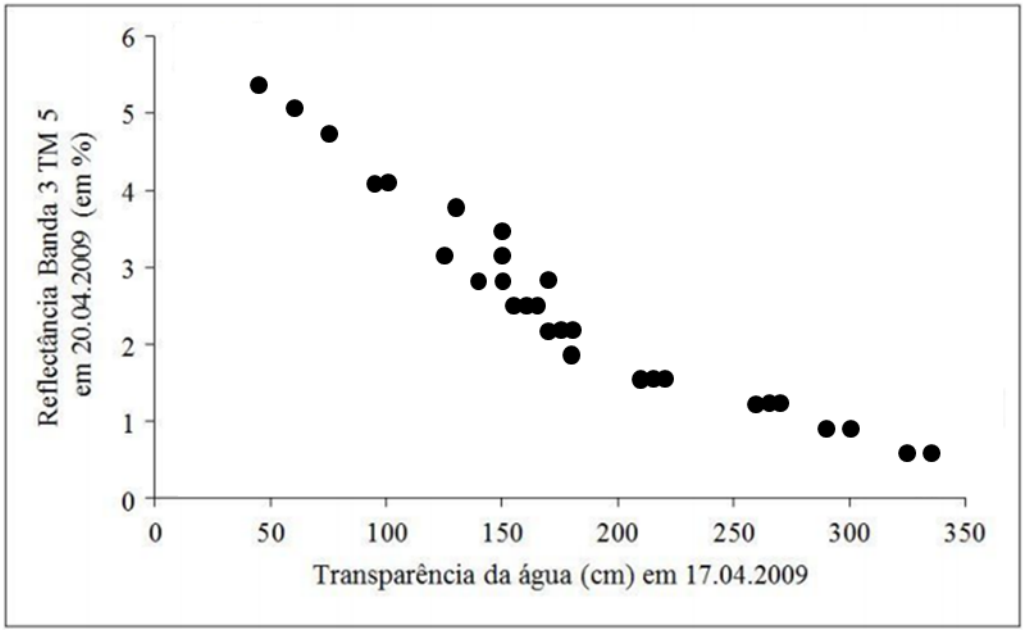
Amostragem não representativa



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"



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

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.

Neter, J, et al, Applied Linear Statistical Models, McGraw Hill, 1996

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

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

- ε_i → componente aleatório (erro ou resíduo)
- X_i → variável independente* (variável explicativa)
- Y_i → variável dependente (variável resposta)

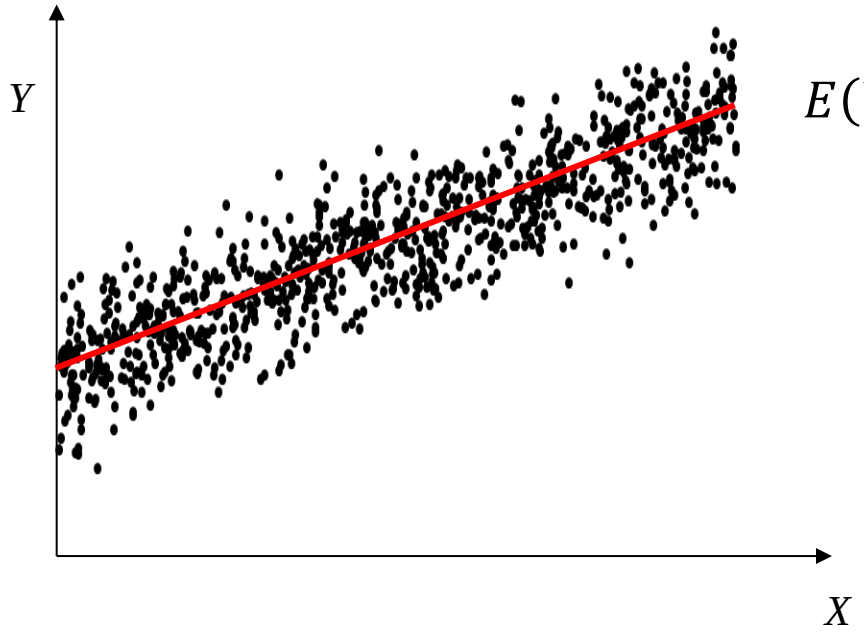
Pressuposições:

$E(\varepsilon_i) = 0$

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

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

β_0 e β_1 são parâmetros (fixos)

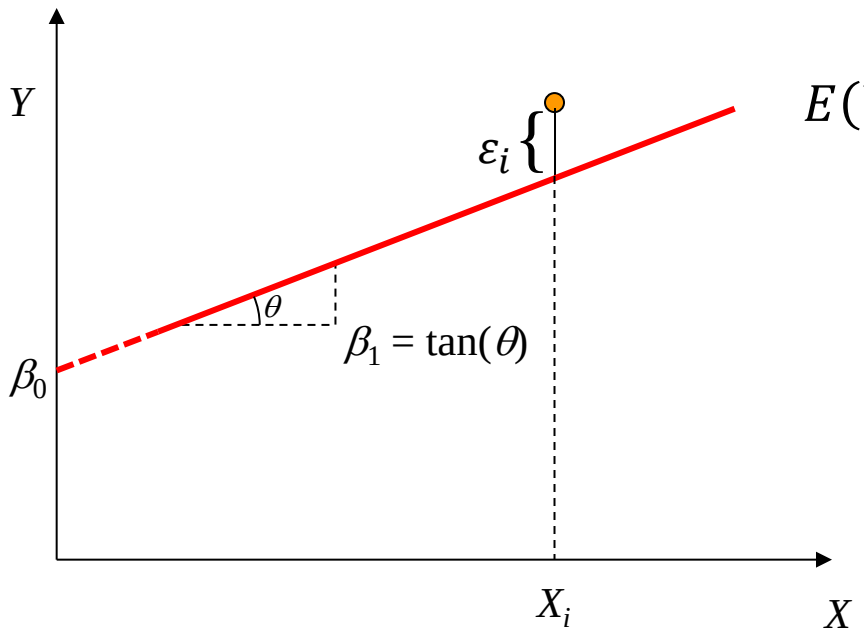


$$E(Y_i) = \beta_0 + \beta_1 X_i$$

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



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

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

Intercepto populacional

Inclinação populacional

β_0 representa o valor de $E(Y_i)$ quando $X_i = 0$

β_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 β_0 e β_1

Em geral não se conhece os valores de β_0, β_1 e σ^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:

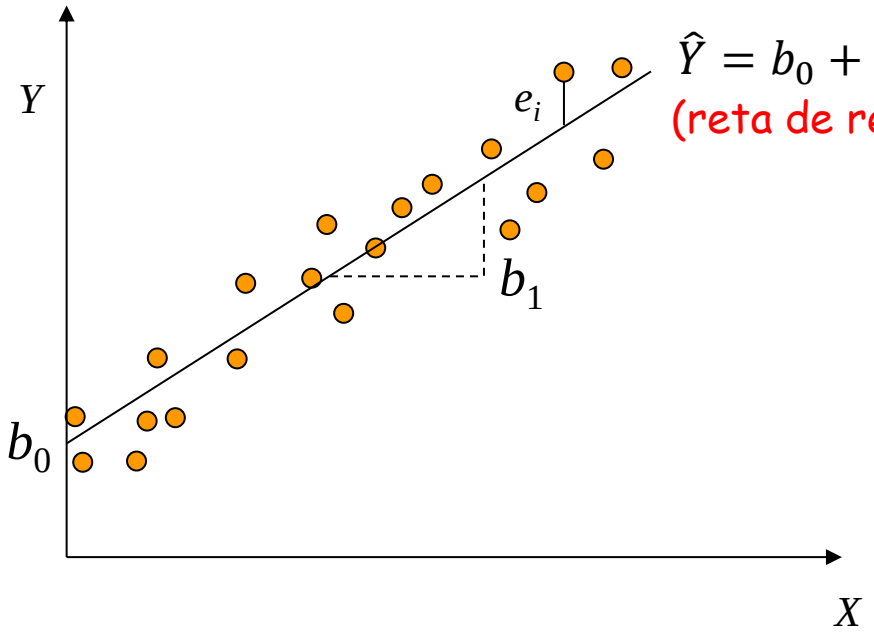
$$\varepsilon_i^2 = [Y_i - E(Y_i)]^2 \qquad \varepsilon_i^2 = [Y_i - (\beta_0 + \beta_1 X_i)]^2$$

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 β_0 e β_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 β_0 e β_1 e igualando-se as expressões encontradas a zero.

Estimação dos parâmetros β_0 e β_1



$\hat{Y} = b_0 + b_1 X$
(reta de regressão estimada)

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

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

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 (σ^2)

A variância dos erros ε_i, denotada por σ^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):

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

A soma dos quadrados dos erros tem $n-2$ graus de liberdade, pois 2 graus de liberdade foram perdidos por estimar β_0 e β_1.

Portanto, o estimador de σ^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 $E[QME] = \sigma^2$

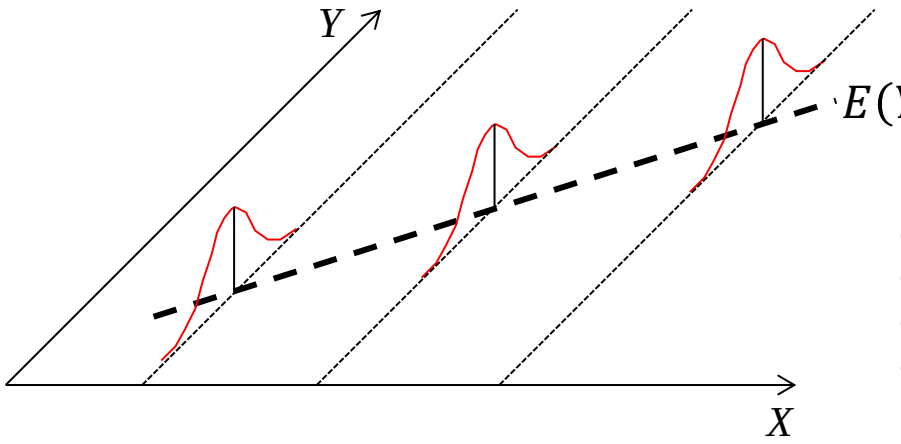
Inferência em Análise de Regressão

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

Para a estimação dos parâmetros β_0, β_1 e σ^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 ε_i.

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

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

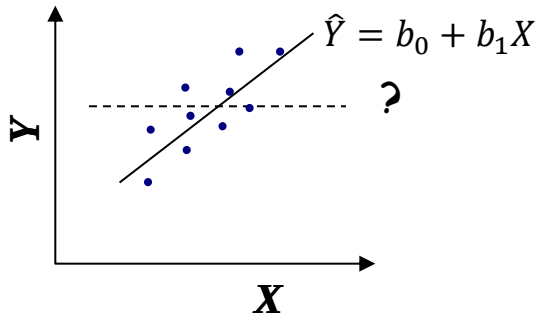


$$E(Y_i) = \beta_0 + \beta_1 X_i$$

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 β_1



$E(Y_i) = \beta_0$?

existe uma relação entre Y e X ?

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

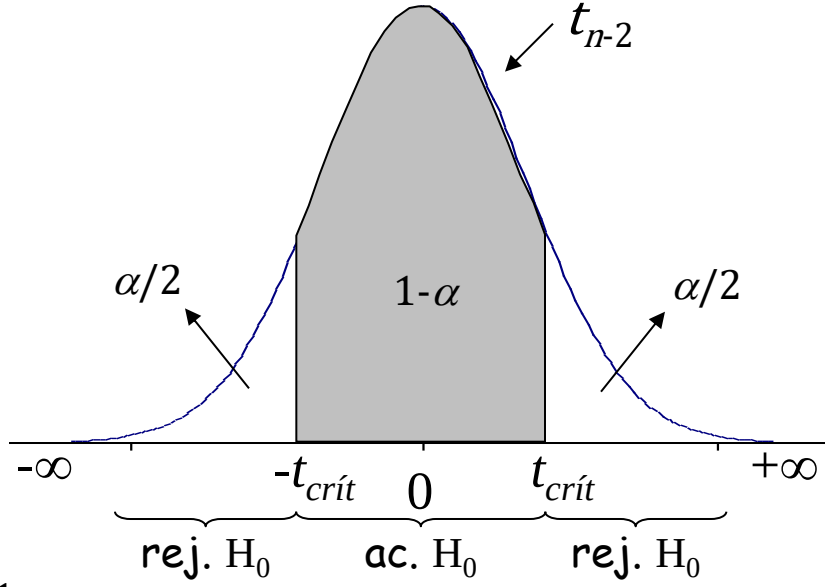
$H_0: \beta_1 = 0$

$H_1: \beta_1 \neq 0$

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

se H_0 verdadeira:

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

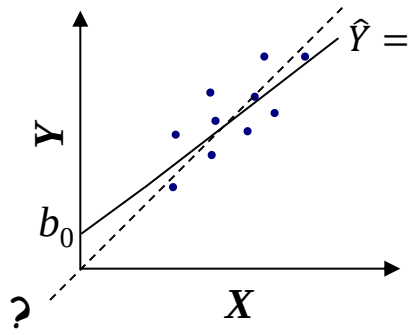


Região Crítica:

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

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 β_0



$E(Y_i) = \beta_1 X_i$?

a regressão passa pela origem ?

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

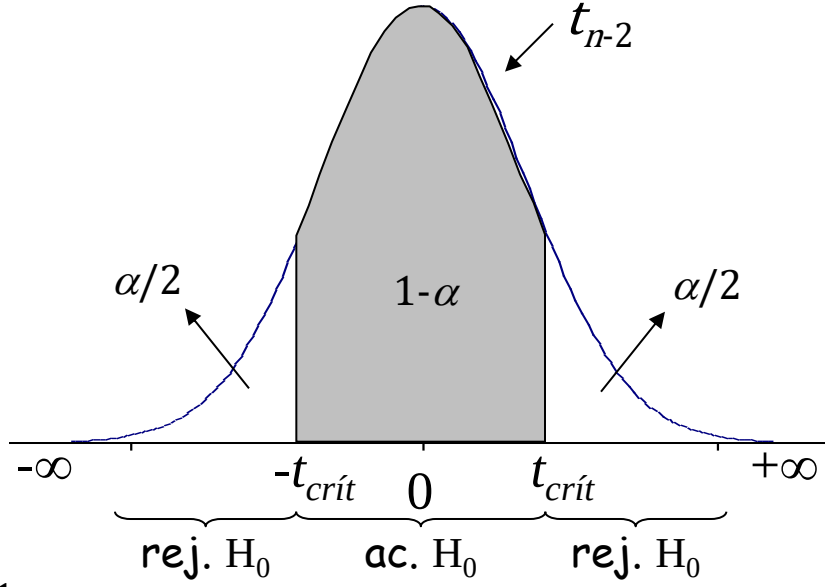
$H_0: \beta_0 = 0$

$H_1: \beta_0 \neq 0$

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

se H_0 verdadeira:

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



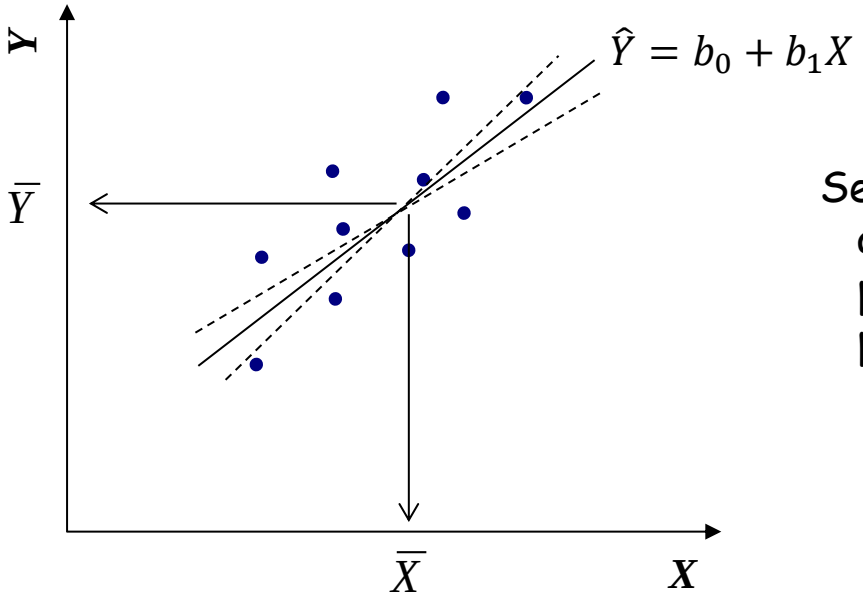
Região Crítica:

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

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 $\mathrm{E}(Y_h)$

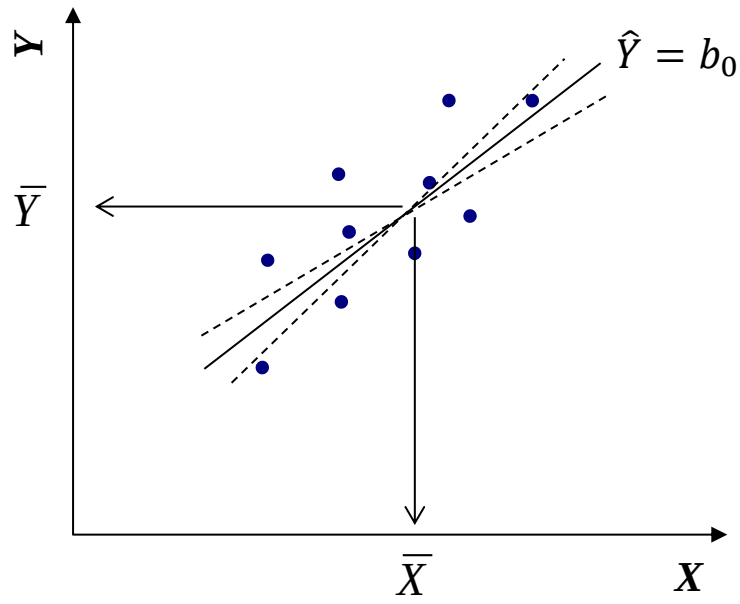
Considerando um determinado valor de X_h, quais as incertezas relacionadas às estimativas de $\mathrm{E}(Y_h)$?



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

Inferências para $\mathrm{E}(Y_h)$

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

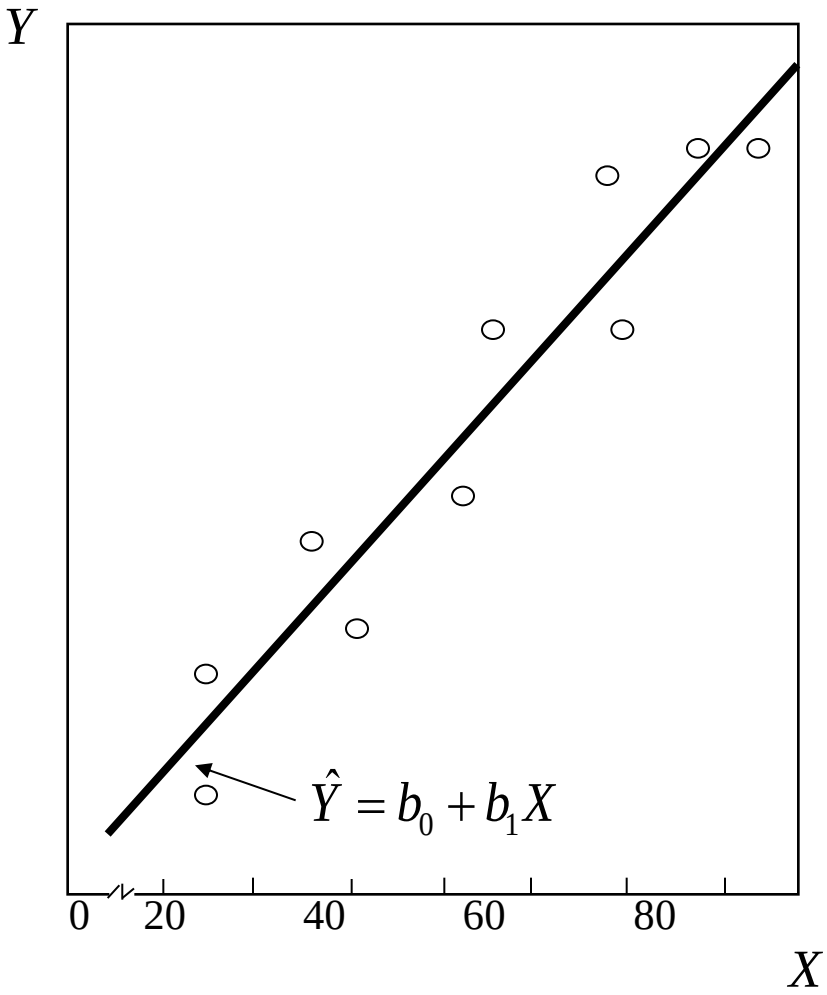


$$\frac{\hat{Y}_h - \mathrm{E}(Y_h)}{s(\hat{Y}_h)} \sim t_{n-2}$$

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

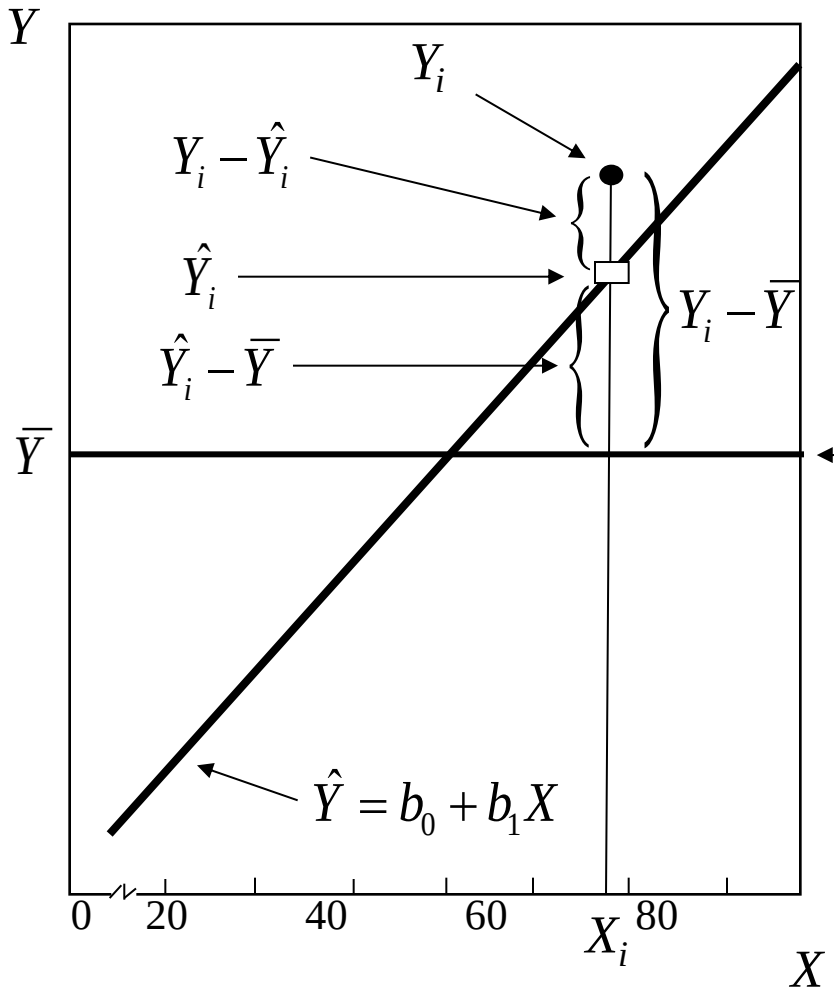
Interpretação: quanto mais distante X_h estiver de $\overline{X}$, maiores serão as incertezas nas estimativas de $\mathrm{E}(Y_h)$. Por isso, extrapolações para faixa de valores de X extremos ou não observados devem ser evitadas!

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

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

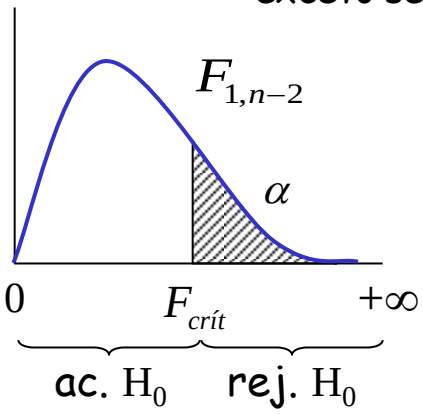
$$\mathrm{E}(QME) = \sigma^2$$

estimador tendencioso, exceto se $\beta_1 = 0$

H_0: $\beta_1 = 0$
H_1: $\beta_1 \neq 0$

se H_0 verdadeira:

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



Região Crítica:

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

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 β_1

Coeficiente de Determinação

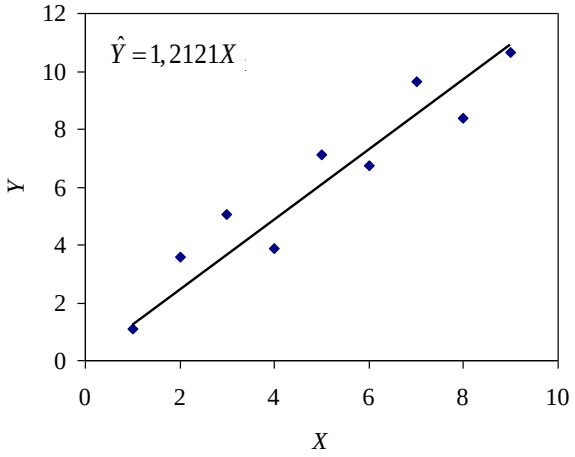
$$r^2 = \frac{SQReg}{SQTO} = \frac{SQTO - SQE}{SQTO} = 1 - \frac{SQE}{SQTO} \qquad 0 \le r^2 \le 1$$

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 ($\beta_0 = 0$)



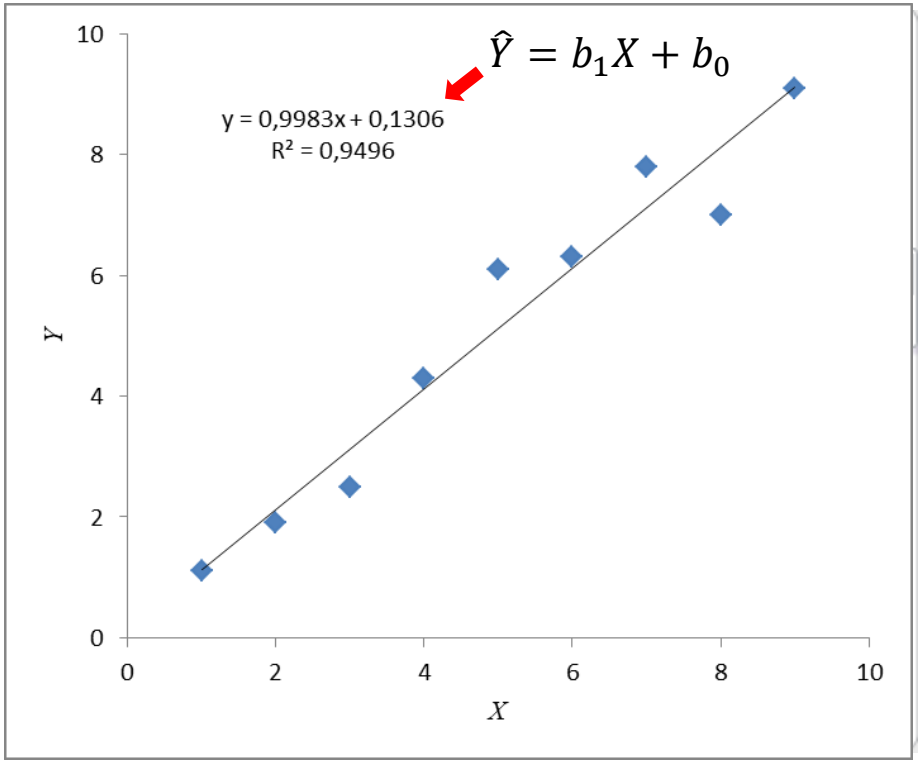
$$Y_i = \beta_1 X_i + \xi_i \qquad b_1 = \frac{\sum_{i=1}^{n} X_i Y_i}{\sum_{i=1}^{n} X_i^2} \qquad s^2(b_1) = \frac{QME}{\sum_{i=1}^{n} X_i^2}$$

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

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

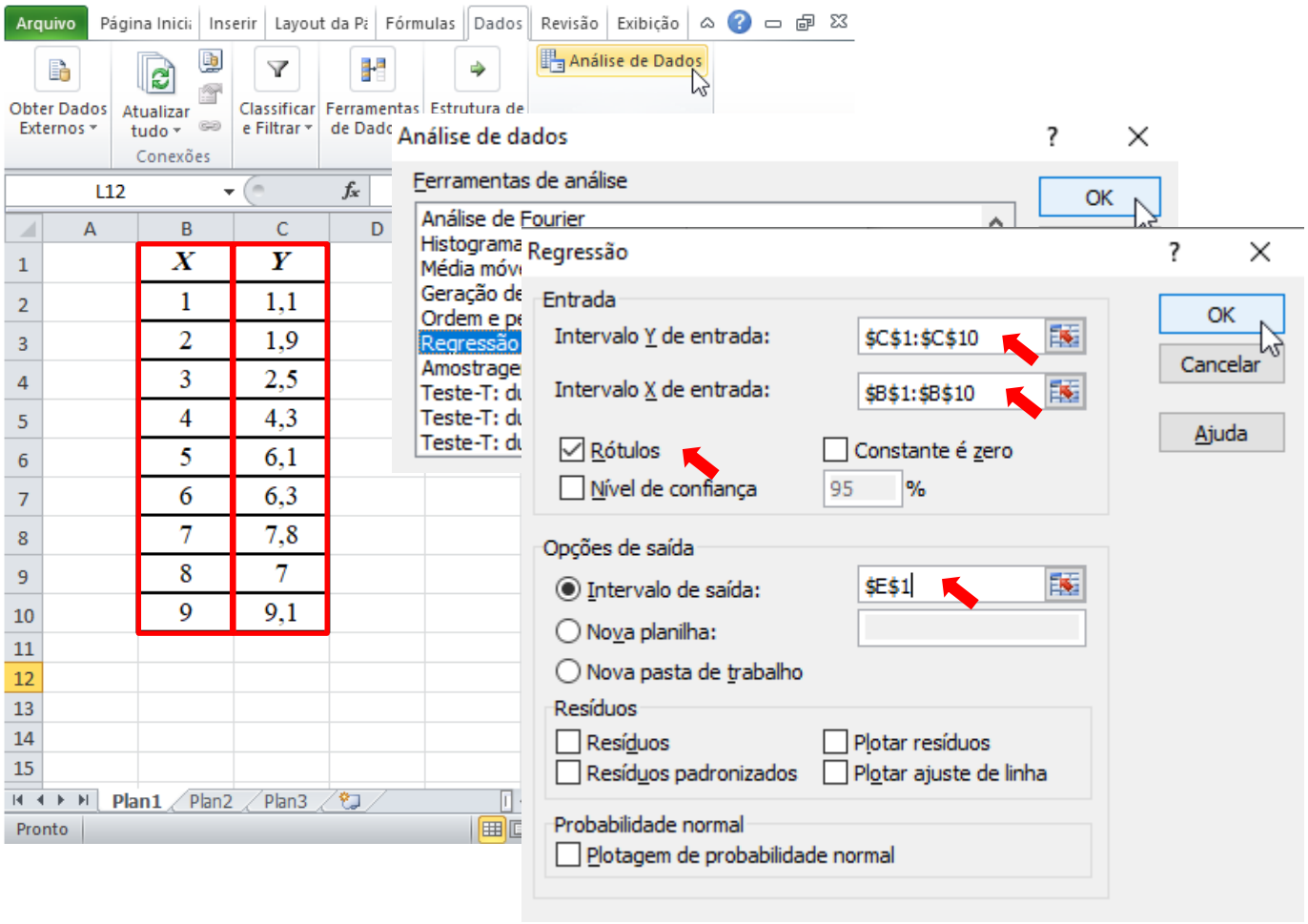
Análise de Regressão no EXCEL

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



Análise de Regressão no EXCEL

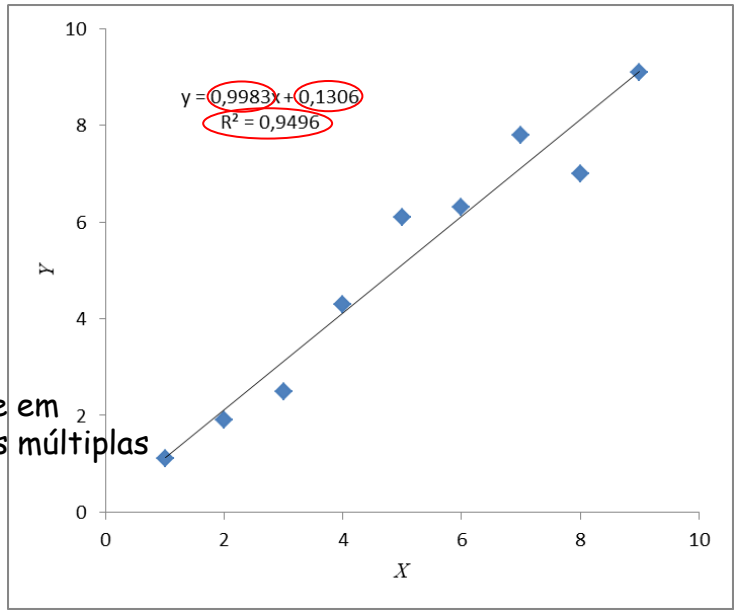
X	Y
1	1,1
2	1,9
3	2,5
4	4,3
5	6,1
6	6,3
7	7,8
8	7,0
9	9,1

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			

QME

valor-P

bilateral

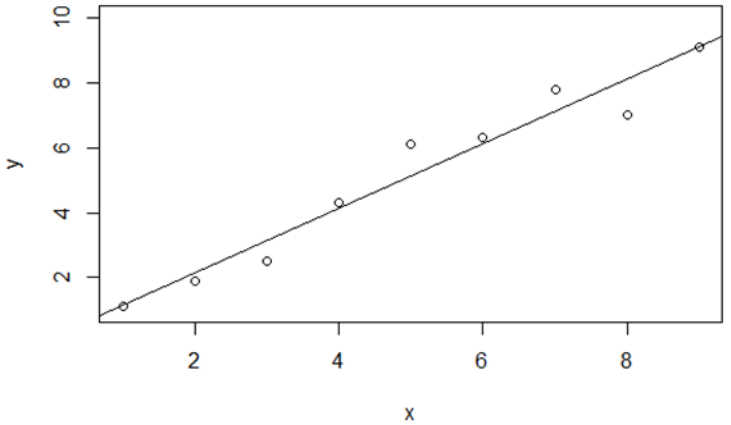
	Coeficientes	*Erro padrão*	*Stat t*	*valor-P*	*95% inferiores*	*95% superiores*
Interseção	0,1306	0,4893	0,2668	0,7973	-1,0265	1,2876
X	0,9983	0,0870	11,4816	8,55E-06	0,7927	1,2039

OBS: Para regressão linear simples: teste F é equivalente ao teste t bilateral para β_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(1.1,1.9,2.5,4.3,6.1,6.3,7.8,7,9.1)
>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: 0.6[illegible]  √QME  es of freedom
Multiple R-squared:  0.9496,  r²  justed 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

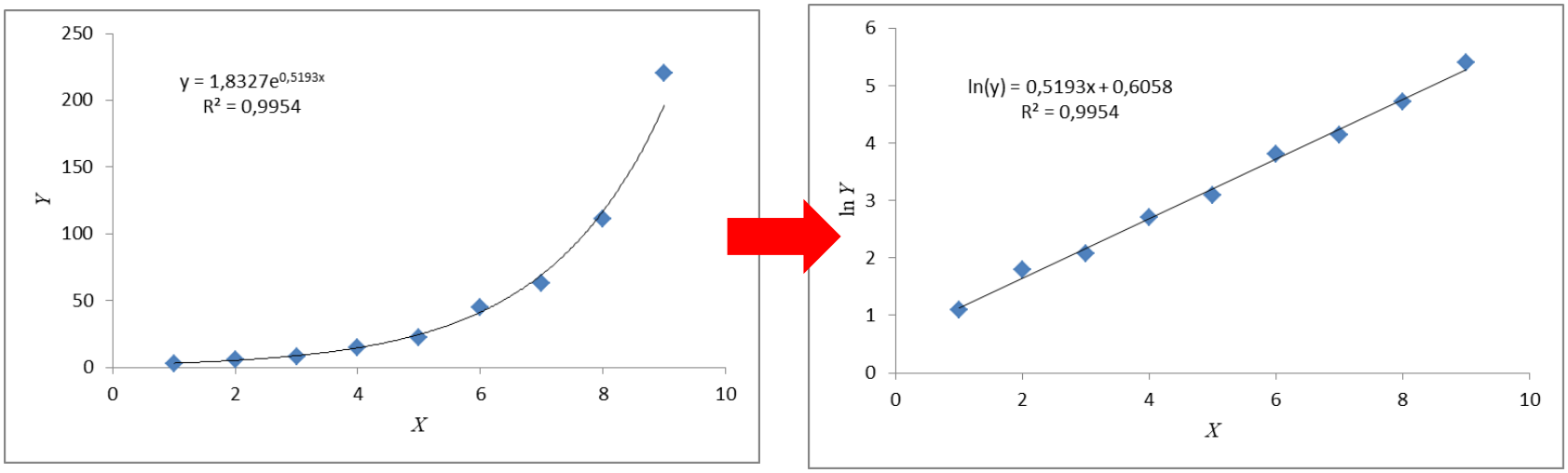
Modelo Padrão: $Y_i = \beta_0 + \beta_1 X_i + \xi_i$

exponencial

$$Y_i = \beta_0 e^{\beta_1 X_i} \xi_i \qquad \ln Y_i = \ln \beta_0 + \beta_1 X_i + \ln \xi_i \qquad Y_i' = \beta_0' + \beta_1 X_i + \xi_i'$$

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

exponencial



Modelos Linearizáveis

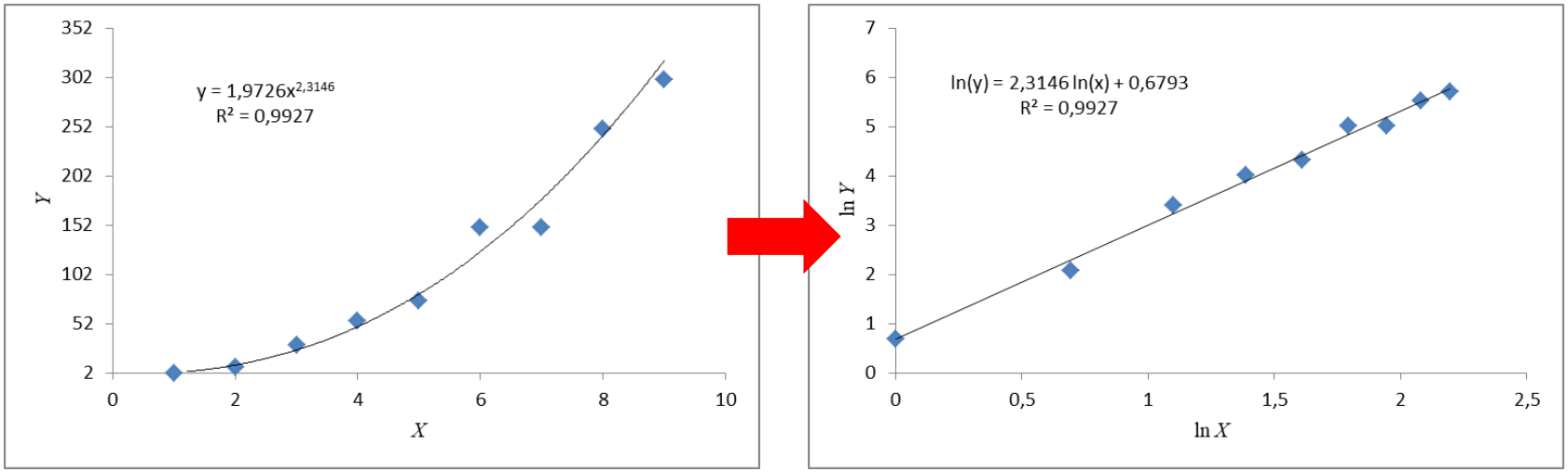
Modelo Padrão: $Y_i = \beta_0 + \beta_1 X_i + \xi_i$

potencial

$$Y_i = \beta_0 X_i^{\beta_1} \xi_i \qquad \ln Y_i = \ln \beta_0 + \beta_1 \ln X_i + \ln \xi_i \qquad Y_i' = \beta_0' + \beta_1 X_i' + \xi_i'$$

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

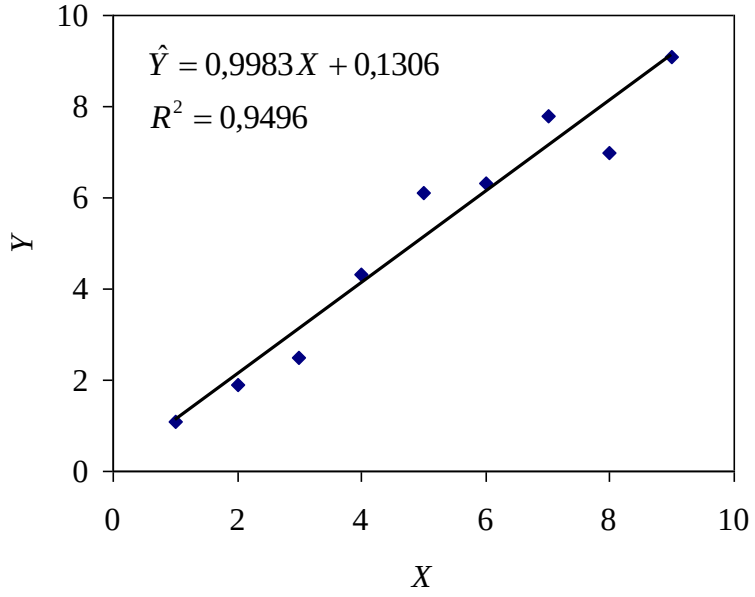
potencial

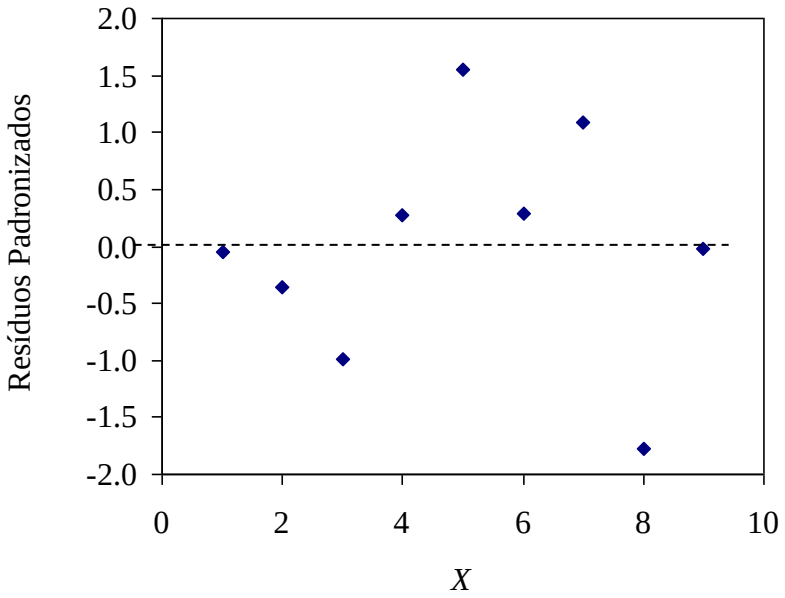


A adoção de uma série de transformações (logaritmo, potência, inverso, etc) em X, Y ou ambos permite que o modelo linear simples possa ser utilizado para representar relações mais complexas $\Rightarrow$ Modelos Lineares Generalizados

Análise de Resíduos

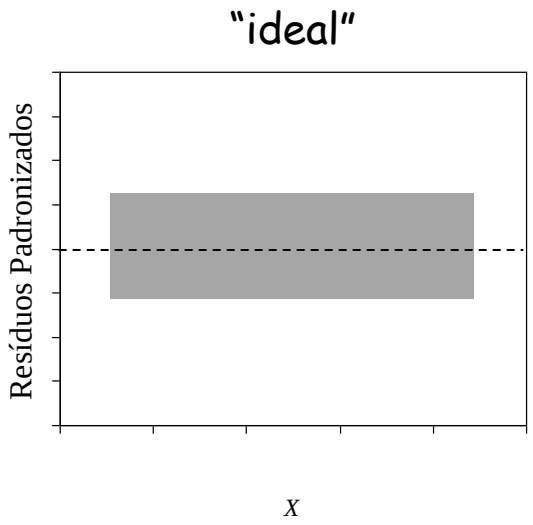
É uma etapa fundamental na Análise de Regressão pois auxilia na identificação de problemas que afetam diretamente a interpretação dos resultados

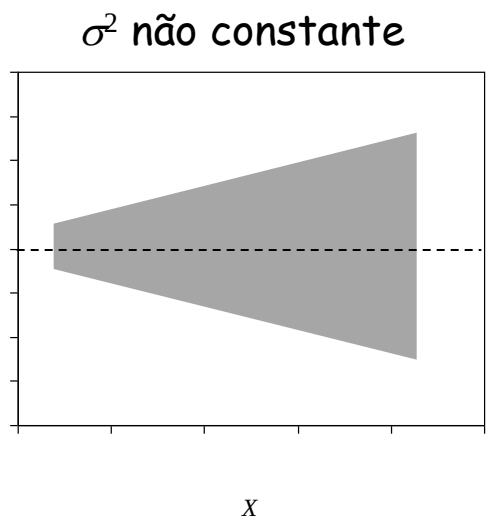


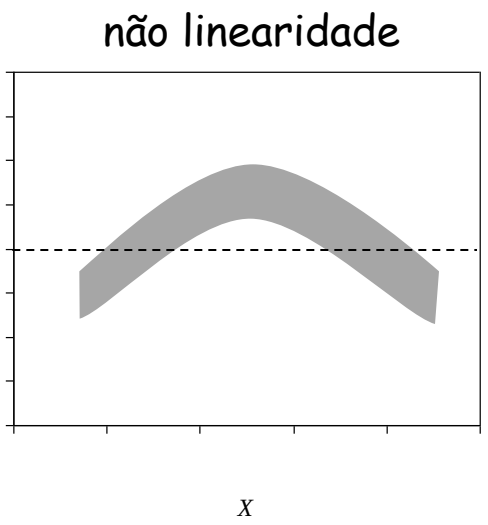


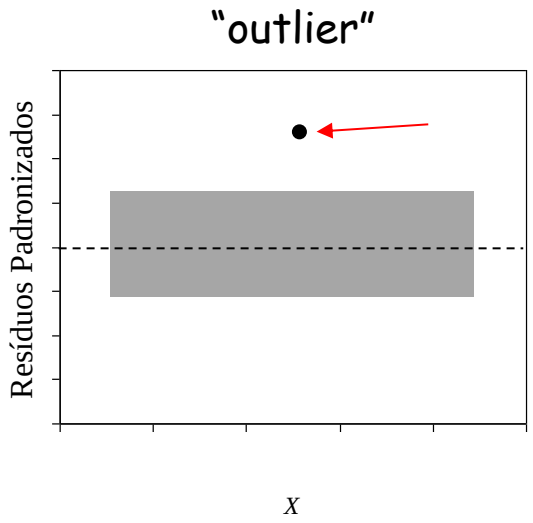
Resíduo Padronizado = $e_i / \sqrt{QME}$

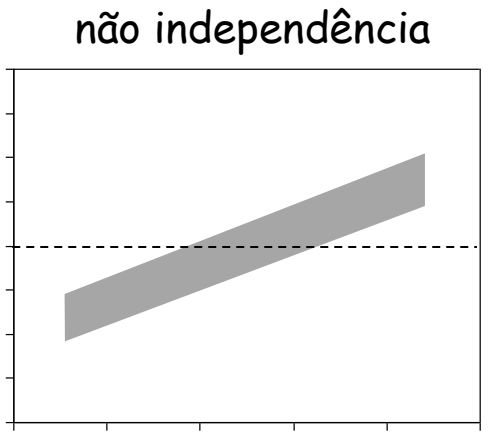
Análise de Resíduos











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

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

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

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

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

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

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

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

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

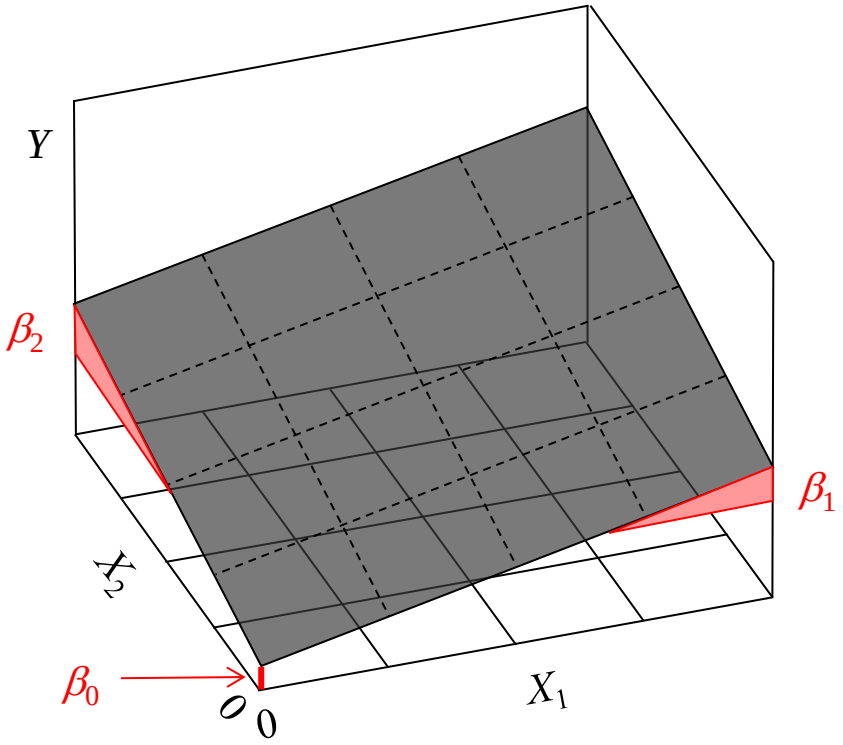
$$Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \beta_3 X_{3,i} + \xi_i$$

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

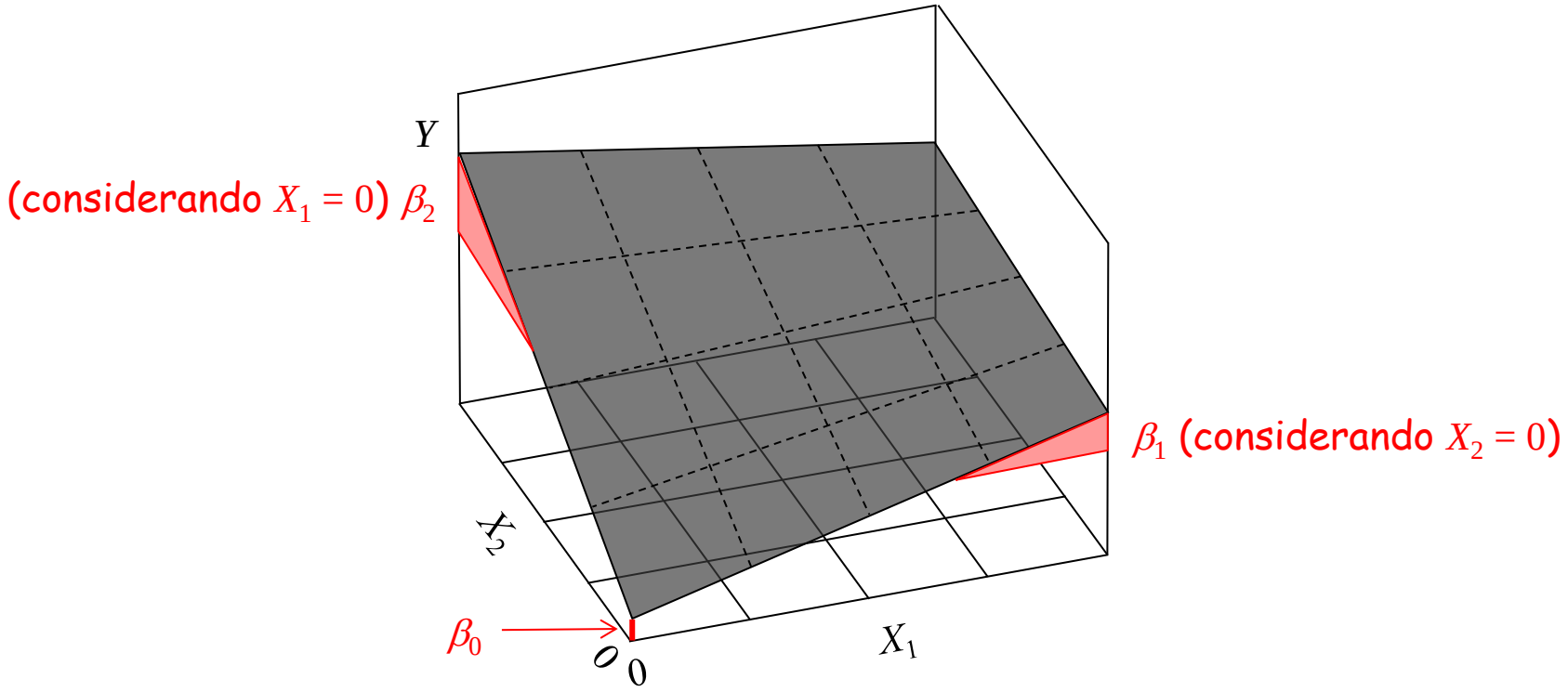
$$Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \xi_i$$



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

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

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

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

primeira coluna de 1's para representar o β_0

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

ANOVA x Análise de Regressão

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

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

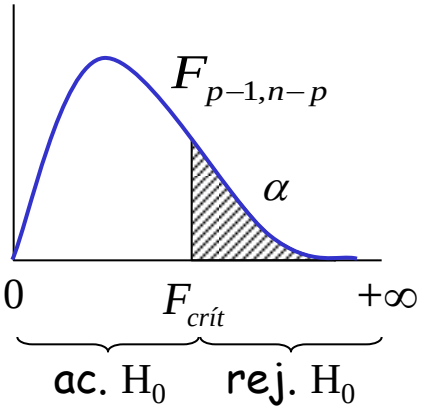
$$\mathrm{E}(QMReg) \geq \sigma^2$$

$$\mathrm{E}(QME) = \sigma^2$$

H_0: $\beta_k = 0$ $(k = 1, \ldots, p-1)$
H_1: pelo menos um dos $\beta_k \neq 0$

se H_0 verdadeira:

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



Região Crítica:

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

OBS: o coeficiente β_0 não tem nenhuma influência sobre este teste
o teste F não é capaz de identificar qual ou quais β_k são diferentes de zero, nem quais β_k são diferentes entre si

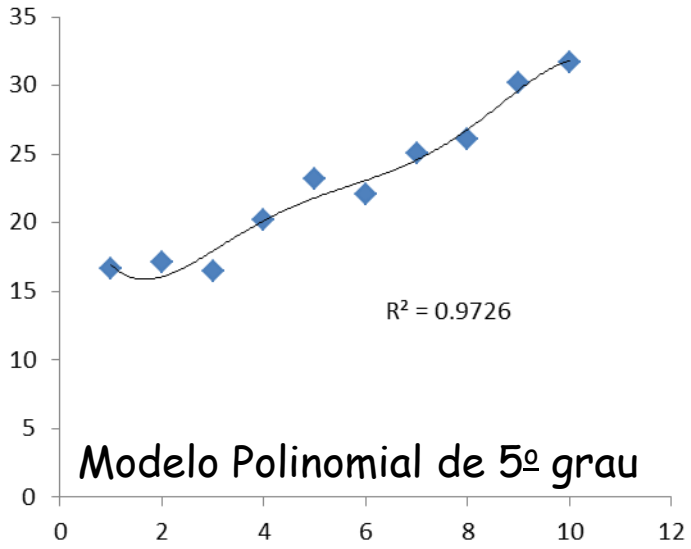
Coeficiente de Determinação Múltiplo

$$SQTO = SQReg + SQE$$

$$r^2 = \frac{SQReg}{SQTO}$$

$$= 1 - \frac{SQE}{SQTO}$$

Atenção: r^2 é fortemente influenciado pelo número de parâmetros considerados no modelo.
Quanto maior o número de parâmetros ($p \to n$), 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 ($p \to n$), 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 β_k

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

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

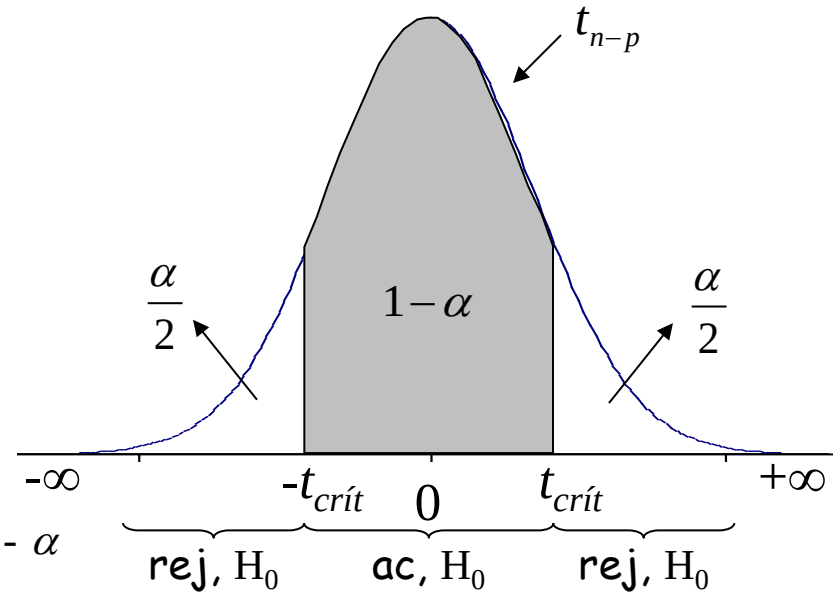
$$H_0 : \beta_k = 0 \qquad E(Y_i) = \sum_{\substack{j=0 \\ j \neq k}}^{p-1} \beta_j X_{j,i}$$

$$H_1 : \beta_k \neq 0$$

Todos os outros β_j estão no modelo, menos o β_k

se H_0 verdadeira:

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



Região Crítica:

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

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_1	X_2	X_3	X_4
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

	Coeficientes	*Erro padrão*	*Stat t*	*valor-P*
Interseção	64,4359	4,8424	13,3067	1,04E-09
X_1	-0,2129	0,3081	-0,6908	0,5002
X_2	-0,4741	0,0160	-29,5575	1,04E-14
X_3	0,2659	0,1553	1,7123	0,1074
X_4	-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 β_k, 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 β_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

$$\mathrm{H}_0 : \beta_1 = \beta_3 = 0$$
$$\mathrm{H}_1 : \beta_1 \neq 0 \text{ e/ou } \beta_3 \neq 0$$

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

Teste de Hipótese para múltiplos β_k

Y	X_1	X_2	X_3	X_4
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			

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

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

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 β_k

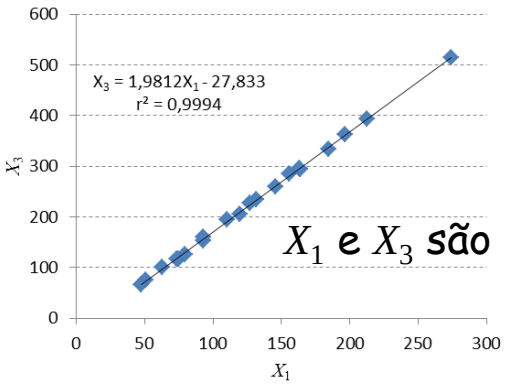
Y	X_1	X_2	X_3	X_4
11,70	126,92	174,56	226,69	364,26
16,34	75,02	129,40	117,43	329,68
16,76	51,00	106,17	75,41	592,57
16,83	47,75	110,50	66,58	471,11
22,02	145,83	148,78	258,84	1151,11
23,43	62,91	113,04	99,85	327,56
24,75	73,34	97,81	117,23	850,26
29,96	79,87	92,83	126,21	695,32
30,31	131,55	139,24	235,10	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:



X_1 e X_3 são colineares!

Mas afinal, quem eu retiro, X_1 ou X_3?

Multicolinearidade

Y	X_1	X_2	X_3	X_4
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 β_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_1	X_2	X_3	X_4
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*

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

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

$$r_1^2 = 0{,}9994$$

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

Multicolinearidade no R

Y	X_1	X_2	X_3	X_4
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

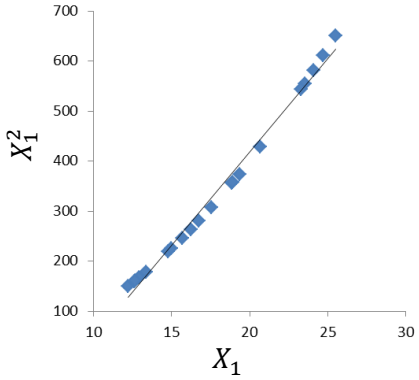
Pode-se então eliminar a variável que apresente o maior *VIF* repetindo-se a análise

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

Ignorando a Multicolinearidade

Y	X_1	X_2	X_1^2
100,38	20,69	15,37	428,08
115,89	16,24	15,52	263,74
116,59	17,55	18,67	308,00
122,59	14,78	19,64	218,45
105,87	19,34	16,86	374,04
121,00	16,73	19,96	279,89
128,75	14,99	21,04	224,70
82,83	24,11	12,51	581,29
104,27	12,23	11,59	149,57
114,52	15,69	15,17	246,18
122,14	18,86	20,00	355,70
109,86	13,35	13,17	178,22
96,16	24,72	18,77	611,08
102,91	12,59	11,38	158,51
103,16	12,92	8,25	166,93
60,45	25,52	4,91	651,27
93,08	18,90	10,49	357,21
87,26	23,55	13,98	554,60
141,60	12,65	26,97	160,02
98,16	23,30	18,32	542,89

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



ANOVA

	gl	*SQ*	*MQ*	*F*	*valor-P*
Regressão	3	6074,07	2024,69	304.15	2,57E-14
Resíduo	16	106,51	6,66		
Total	19	6180,58			

	Coeficientes	*Erro padrão*	*Stat t*	*valor-P*
Interseção	55.6623	12,7288	4,3729	0,0005
X_1	4,0979	1,4886	2,7527	0,0141
X_2	2,3083	0,1269	18,1920	4,1E-12
X_1^2	-0,1729	0,0399	-4,3290	0,0005

	VIF
X_1	125,41
X_2	1,18
X_1^2	127,02

Mas como pode o coeficiente ser significativo e apresentar *VIF* alto?

Atenção: Nem sempre *VIF* altos devem ser descartados! Termos polinomiais e interações podem apresentar *VIF* alto pois espera-se que sejam relacionados com a variável independente original.

Eliminando-se variáveis independentes

Y	X_1	X_2	X_3	X_4
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			

	Coeficientes	*Erro padrão*	*Stat t*	*valor-P*
Interseção	64,4359	4,8424	13,3067	1,04E-09
X_1	-0,2129	0,3081	-0,6908	0,5002
X_2	-0,4741	0,0160	-29,5575	1,04E-14
X_3	0,2659	0,1553	1,7123	0,1074
X_4	-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_2	X_3	X_4	$\hat{Y}$
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

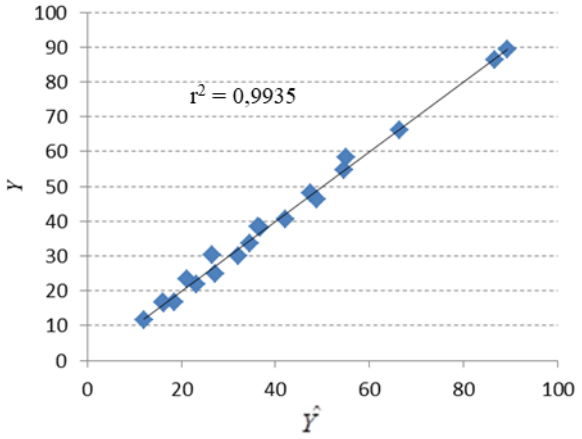
$$Y_i = \beta_0 + \beta_2 X_{2,i} + \beta_3 X_{3,i} + \beta_4 X_{4,i} + \xi_i$$

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			

	Coeficientes	*Erro padrão*	*Stat t*	*valor-P*
Interseção	61,4478	2,1408	28,7030	3,44E-15
X_2	-0,4734	0,0157	-30,0637	1,66E-15
X_3	0,1587	0,0040	40,0206	1,81E-17
X_4	-0,0077	0,0015	-5,1698	9,31E-05

todos significativos a 5%



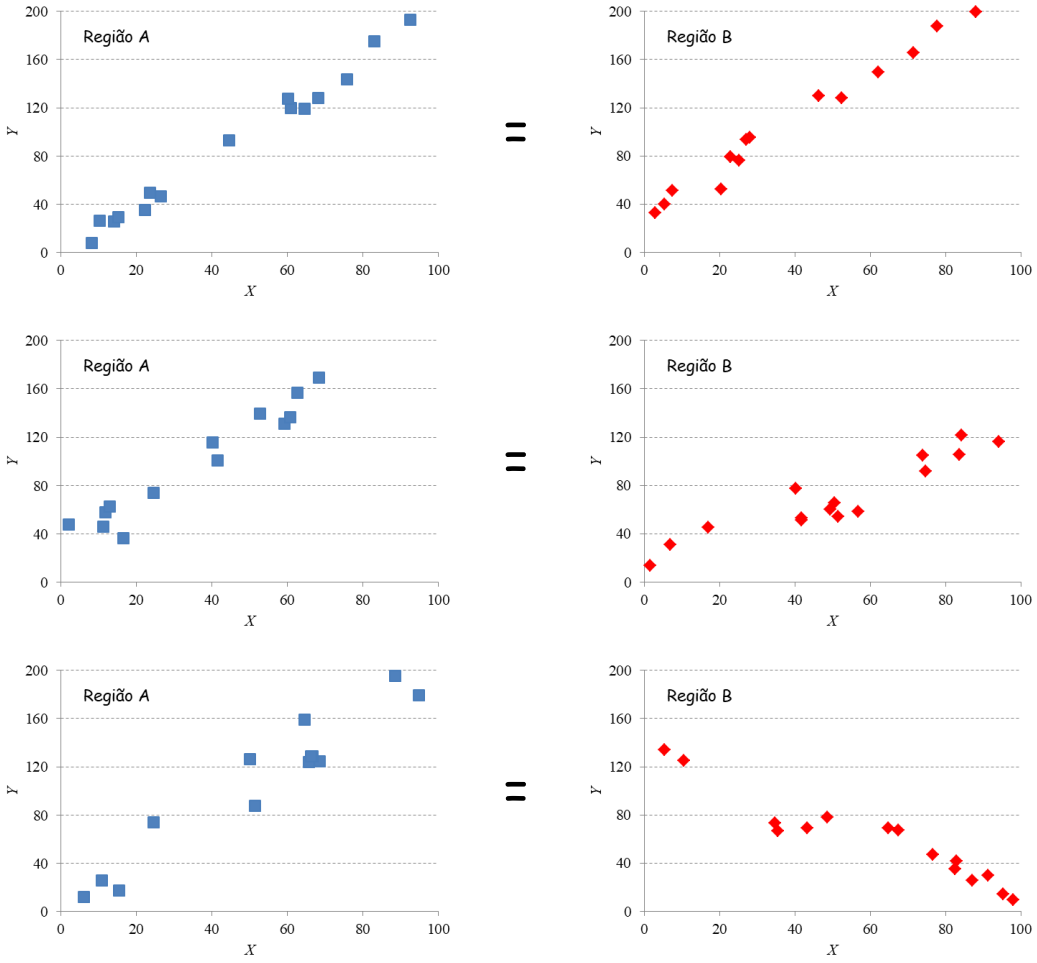
Coeficiente de correlação múltiplo

$$r = \sqrt{r^2}$$
$$r = 0{,}9964$$

(evite usar este índice!)

Comparando funções de regressão

Muitas vezes deseja-se saber se dois conjuntos amostrais adquiridos em duas regiões distintas resultam na mesma função de regressão, ou seja, se Y e X se relacionam da mesma forma nas duas regiões.



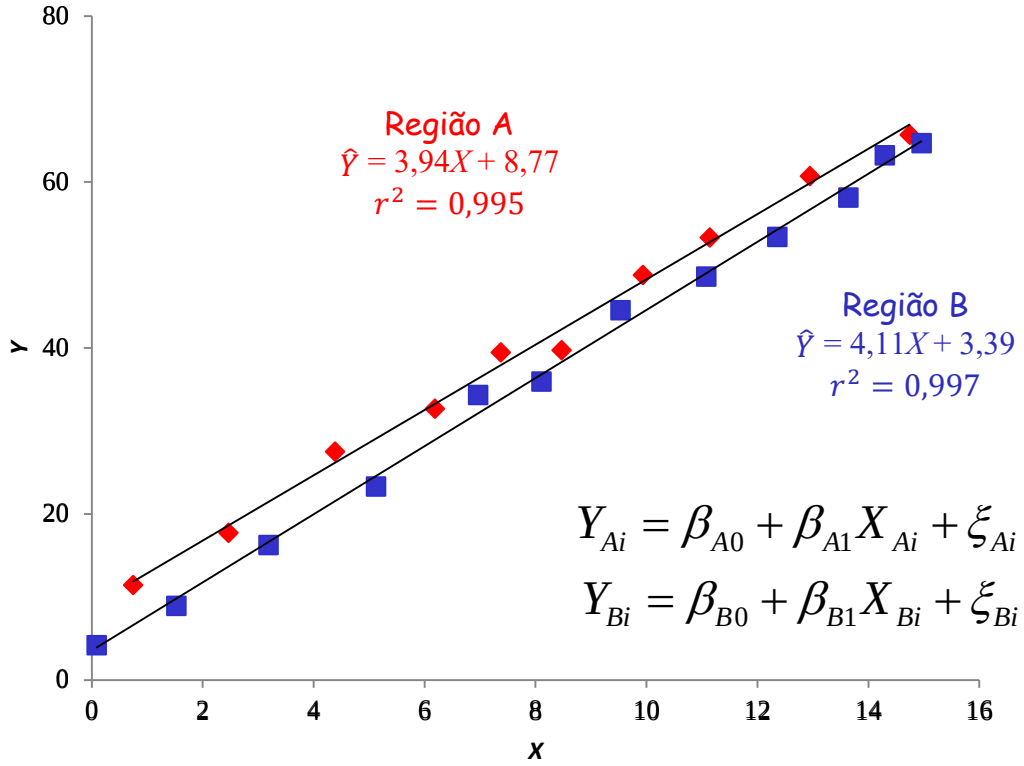
Para testar esta hipótese, é possível gerar uma única regressão usando uma variável indicadora (*dummy*) a fim de identificar a origem de cada ponto amostral.

Vamos analisar 2 exemplos a seguir.

Exemplo 1

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



as relações parecem ser mesmo lineares!

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

Exemplo 1

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 ($W_i = 1$):

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

Exemplo 1

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

Exemplo 1

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

	Coeficientes	*Erro 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

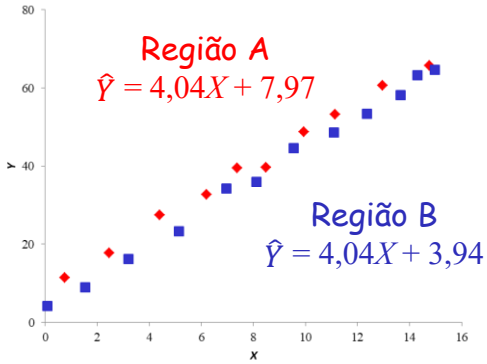
não significativo ($\beta_3 = 0$)

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

Exemplo 1

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			

	Coeficientes	*Erro 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%

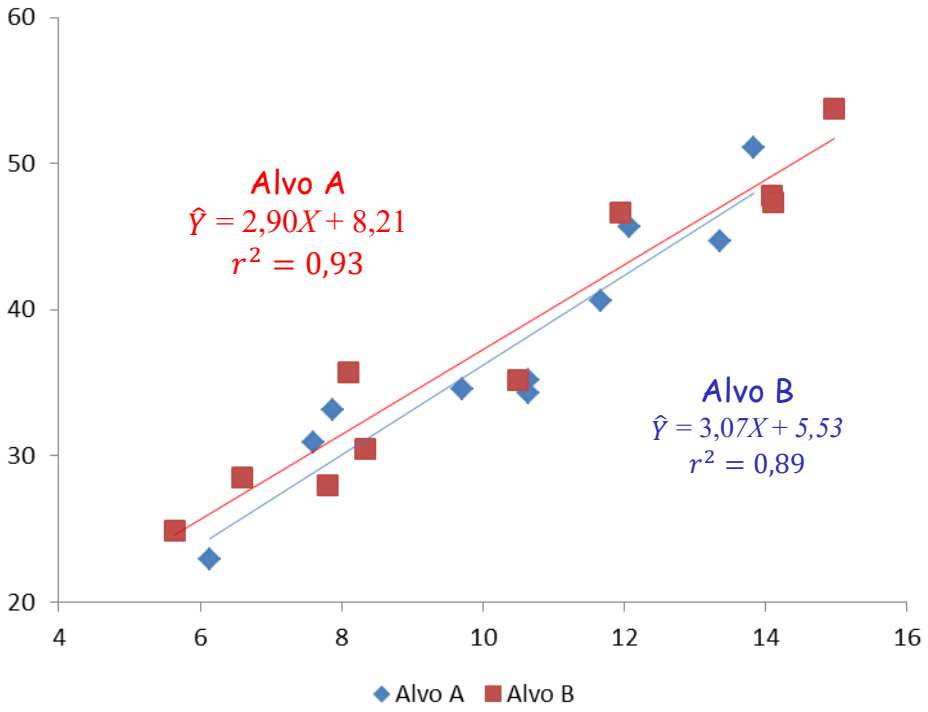
Conclusão: a 5% de significância, as regressões de ambas regiões possuem o mesmo coeficiente angular. Elas diferem-se apenas pelo intercepto.

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

Exemplo 2

Alvo A	
Y	X
30,96	7,61
22,94	6,13
51,08	13,84
45,71	12,07
44,71	13,35
35,16	10,65
33,13	7,88
34,54	9,71
40,64	11,68
34,32	10,65

Alvo B	
Y	X
53,79	14,97
35,70	8,09
47,79	14,09
46,67	11,94
35,23	10,50
47,40	14,11
24,93	5,64
27,98	7,81
28,52	6,60
30,47	8,34



as relações parecem ser mesmo lineares!

Exemplo 2

Y	X	W	XW
30,96	7,61	0	0
22,94	6,13	0	0
51,08	13,84	0	0
45,71	12,07	0	0
44,71	13,35	0	0
35,16	10,65	0	0
33,13	7,88	0	0
34,54	9,71	0	0
40,64	11,68	0	0
34,32	10,65	0	0
53,79	14,97	1	14,97
35,7	8,09	1	8,09
47,79	14,09	1	14,09
46,67	11,94	1	11,94
35,23	10,5	1	10,5
47,4	14,11	1	14,11
24,93	5,64	1	5,64
27,98	7,81	1	7,81
28,52	6,6	1	6,6
30,47	8,34	1	8,34

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

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

ANOVA

	gl	*SQ*	*MQ*	*F*	*Valor-P*
Regressão	3	1420,54	473,51	57,56	8,59E-09
Resíduo	16	131,63	8,23		
Total	19	1552,17			

	Coeficientes	*Erro padrão*	*Stat t*	*valor-P*
Interseção	5,53	4,01	1,38	0,186
X	3,07	0,38	8,14	4,41E-07
W	2,68	5,02	0,53	0,601
XW	-0,17	0,47	-0,35	0,729

← elimina-se o com maior valor-P

Exemplo 2

Y	X	W
30,96	7,61	0
22,94	6,13	0
51,08	13,84	0
45,71	12,07	0
44,71	13,35	0
35,16	10,65	0
33,13	7,88	0
34,54	9,71	0
40,64	11,68	0
34,32	10,65	0
53,79	14,97	1
35,7	8,09	1
47,79	14,09	1
46,67	11,94	1
35,23	10,5	1
47,4	14,11	1
24,93	5,64	1
27,98	7,81	1
28,52	6,6	1
30,47	8,34	1

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

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

ANOVA

	gl	*SQ*	*MQ*	*F*	*Valor-P*
Regressão	2	1419,52	709,76	90,96	8,32E-10
Resíduo	17	132,65	7,80		
Total	19	1552,17			

	Coeficientes	*Erro padrão*	*Stat t*	*valor-P*
Interseção	6,64	2,44	2,72	0,015
X	2,96	0,22	13,48	1,66E-10
W	0,97	1,25	0,77	0,449

← elimina-se também

Exemplo 2

Y	X
30,96	7,61
22,94	6,13
51,08	13,84
45,71	12,07
44,71	13,35
35,16	10,65
33,13	7,88
34,54	9,71
40,64	11,68
34,32	10,65
53,79	14,97
35,7	8,09
47,79	14,09
46,67	11,94
35,23	10,5
47,4	14,11
24,93	5,64
27,98	7,81
28,52	6,6
30,47	8,34

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

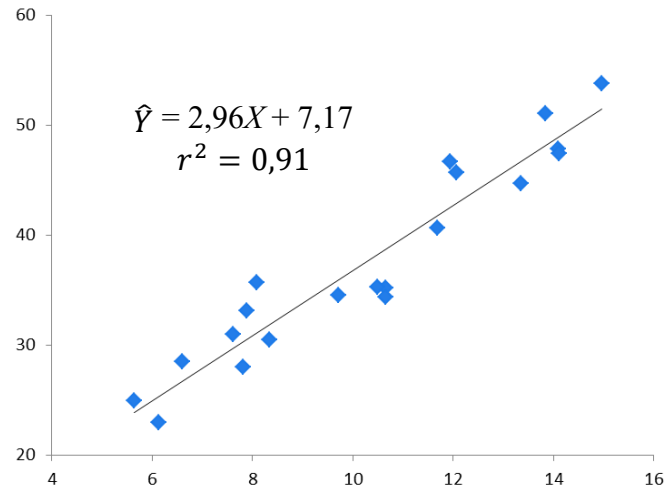
ANOVA

	gl	*SQ*	*MQ*	*F*	*Valor-P*
Regressão	1	1414,85	1414,85	185,45	6,42E-11
Resíduo	18	137,33	7,63		
Total	19	1552,17			

	Coeficientes	*Erro padrão*	*Stat t*	*valor-P*
Interseção	7,17	2,32	3,09	0,0063
X	2,96	0,22	13,62	6,42E-11

Conclusão: a 5% de significância, ambas regiões possuem o mesmo modelo de regressão

Vantagem: maior amostra!



Regressão Padronizada

Muitas vezes, o coeficiente β_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 β_1 e β_2?

$\beta_1 \rightarrow$ mm.ha/ton $\qquad$ $\beta_2 \rightarrow$ mm/°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:

$$Y_i' = \frac{Y_i - \overline{Y}}{s_Y} \qquad X'_{k,i} = \frac{X_{k,i} - \overline{X}_k}{s_{X_k}}$$

Nesse caso, a reta de regressão estimada

$$\hat{Y}_i = b_0 + b_1 X_{1,i} + b_2 X_{2,i} + \ldots + b_{p-1} X_{p-1,i}$$

torna-se

$$\hat{Y}_i' = b_1' X'_{1,i} + b_2' X'_{2,i} + \ldots + b'_{p-1} X'_{p-1,i} \quad \Rightarrow \quad b_k' = b_k \frac{s_{X_k}}{s_Y}$$

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:

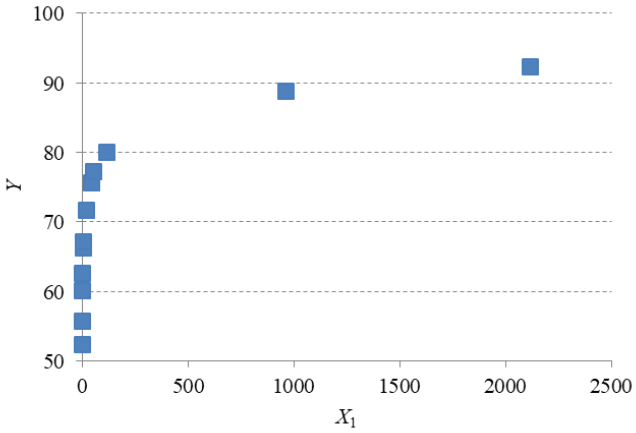
- quanto mais simples o modelo, melhor.
- dê preferência por modelos lineares (ou linearizáveis).
- utilize conhecimentos prévios para escolha do tipo de relação entre cada variável independente e a dependente (linear, polinomial, logarítmico, exponencial, potencial, etc), construindo primeiramente um 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.

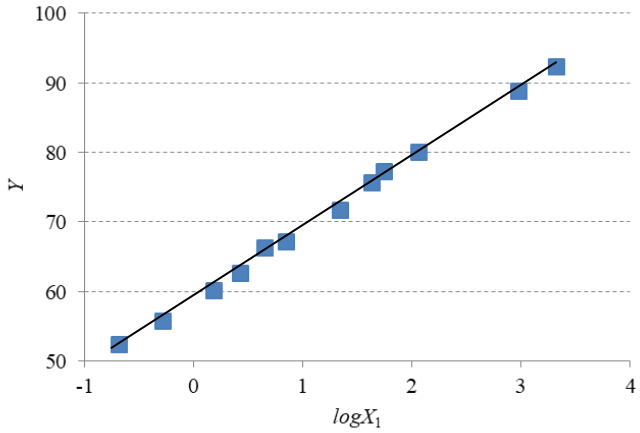
Construção do Modelo

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	$logX_1$
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





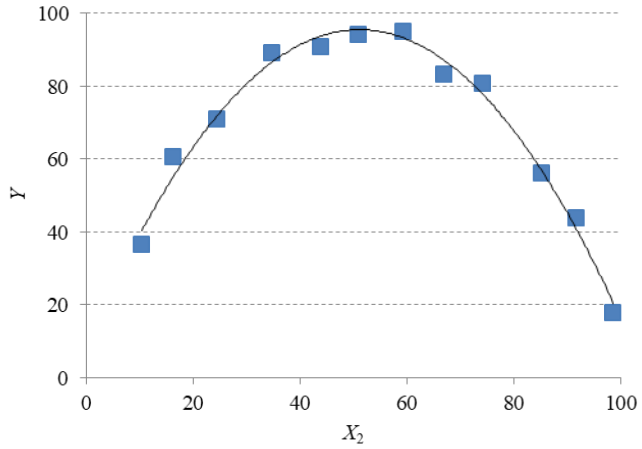
$$Y_i = \beta_0 + \beta_1 logX_{1,i} + \xi_i$$

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_2	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

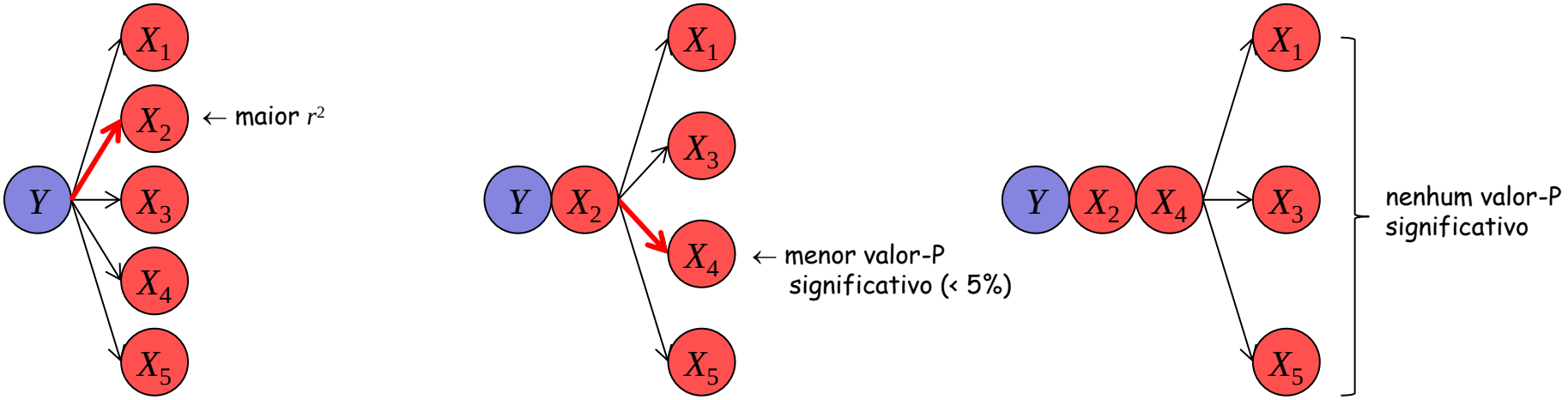


$$Y_i = \beta_0 + \beta_1 X_{2,i} + \beta_2 X_{2,i}^2 + \xi_i$$

OBS: cuidado ao propor transformações em Y pois esta afeta a relação com todas as variáveis independentes numa regressão múltipla!

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.

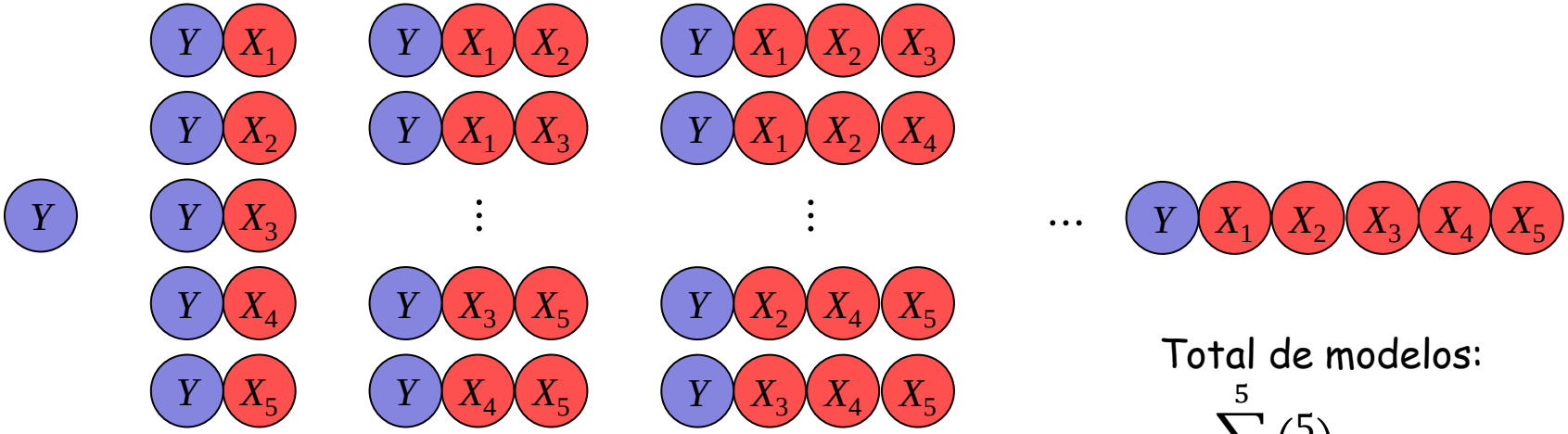


Modelo Final: $Y_i = \beta_0 + \beta_2 X_{2,i} + \beta_4 X_{4,i} + \xi_i$

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 de busca. 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

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

Quanto maior for o valor r_a^2, melhor o modelo.

- Critério C_p de Mallows

$$C_p = \frac{SQE_{R(p)}}{QME_C} - n + 2p \qquad C_p \leq p$$

Quanto mais próximo C_p de p, 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
http://www.dpi.inpe.br/~camilo/estatistica/r/RegrDados.dat

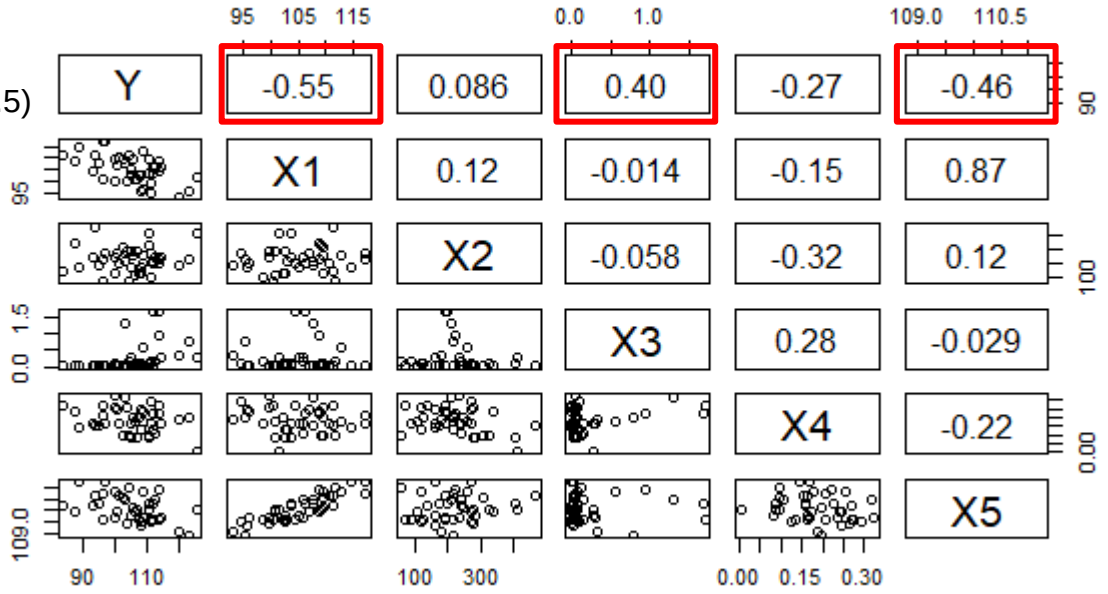
Y	X1	X2	X3	X4	X5
100,9716	110,6689	191,5408	0,035715	0,163621	110,631
112,1469	105,8706	243,3895	0,15264	0,106897	110,4889
104,326	111,5136	62,96696	0,132668	0,094935	110,9833
108,5016	96,00361	195,0345	0,050644	0,235669	109,6515
96,11517	100,5438	71,00401	0,006774	0,274924	109,5986
106,3803	100,9839	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)
```



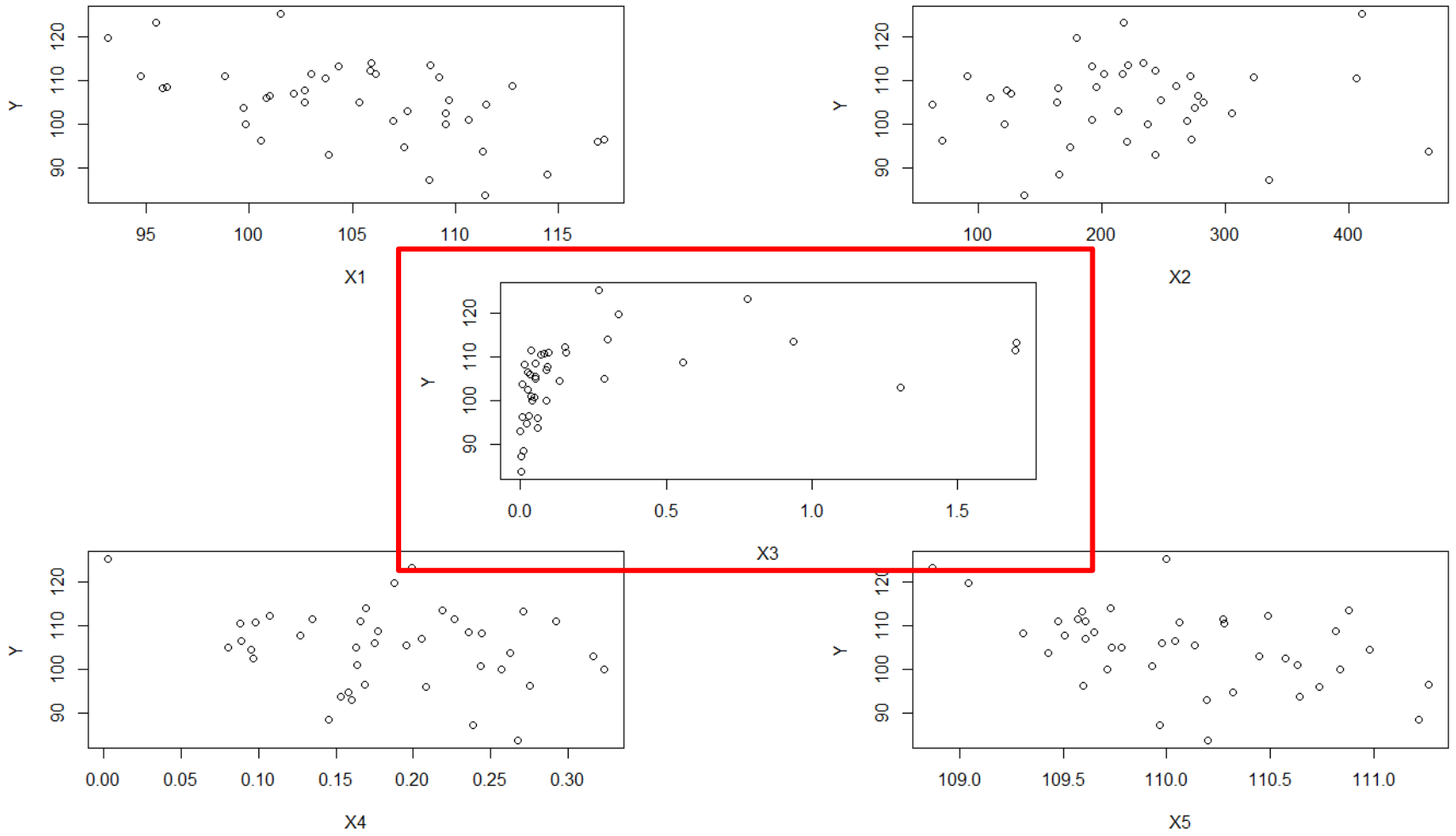
Qual são as melhores variáveis que explicam Y?

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

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

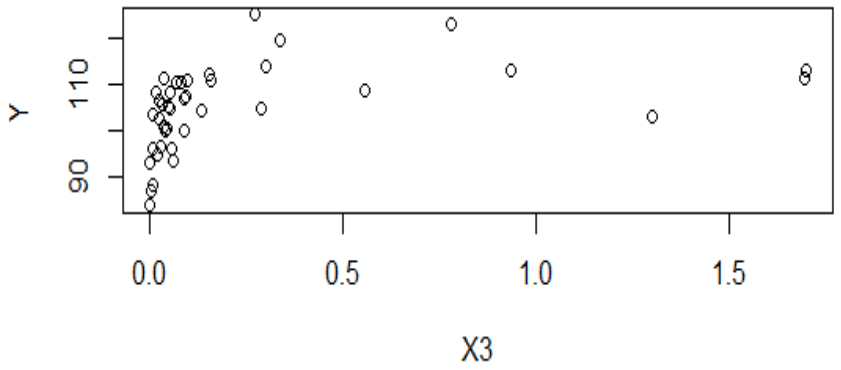
Exemplo

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)

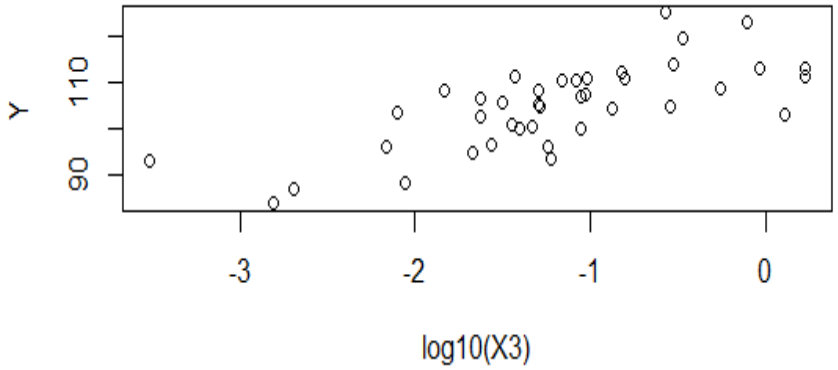


Exemplo

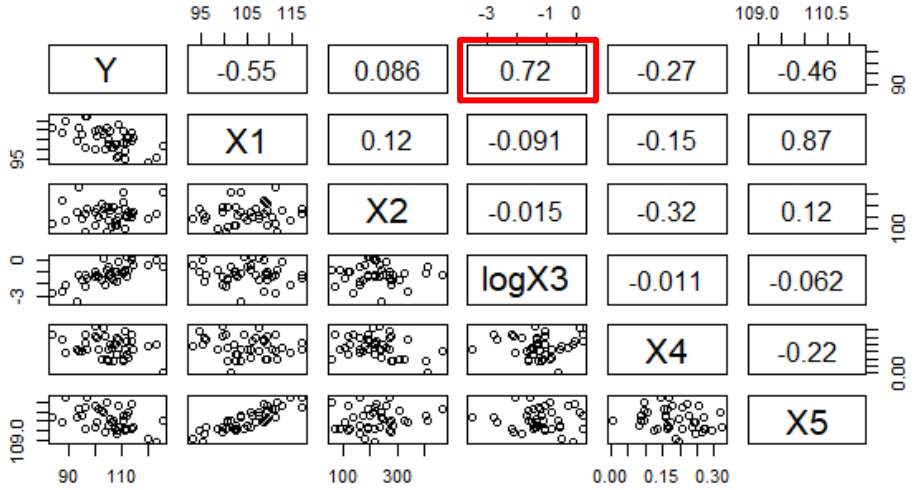
plot(Y~X3,data=dados)



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 ~ X1 + X2 + logX3 + X4 + X5

	Df	Sum of Sq	RSS	AIC
- X5	1	5.30	418.04	103.87
- 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

diminui AIC

- Step: AIC=103.87
- Y ~ X1 + X2 + logX3 + X4

	Df	Sum of Sq	RSS	AIC
- X2	1	10.27	428.31	102.84
<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

diminui AIC

- Step: AIC=102.84
- Y ~ X1 + logX3 + 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

OBS: Não avalia a significância dos coeficientes dos modelos!

Exemplo

Resumo do modelo selecionado:

summary(regsel)

- Call:
- lm(formula = Y ~ X1 + logX3 + X4, data = dados)

- 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

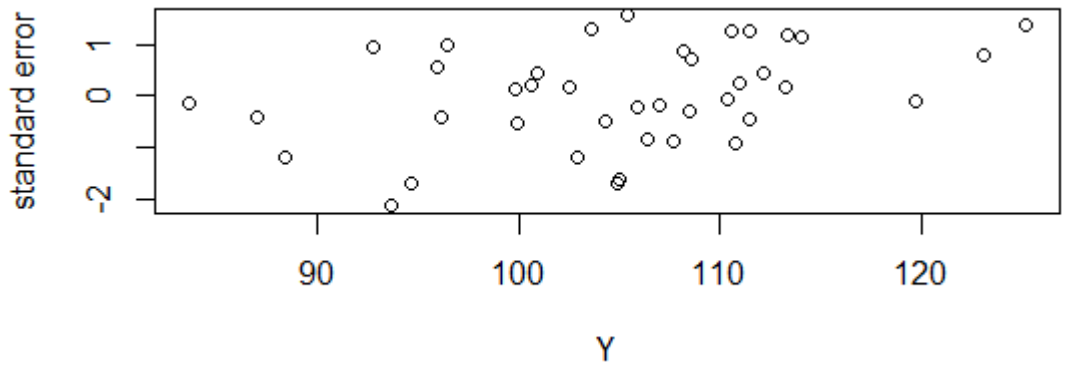
resíduos são normalmente distribuídos

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

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

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

variância é constante



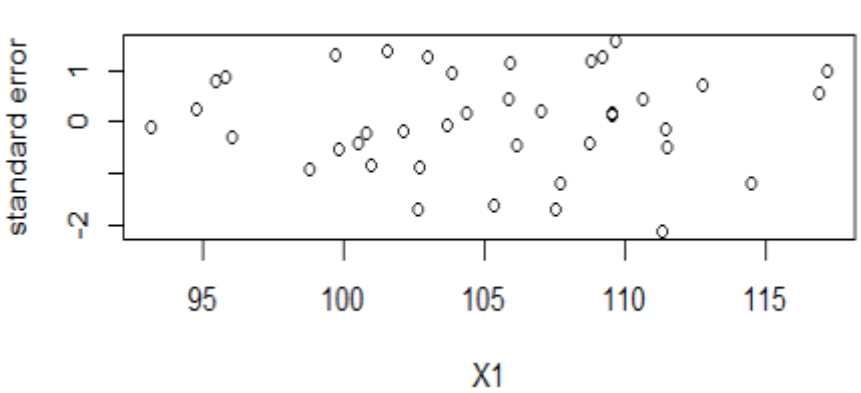
Aparentemente nenhum outlier (|erro padronizado| > 2,5)

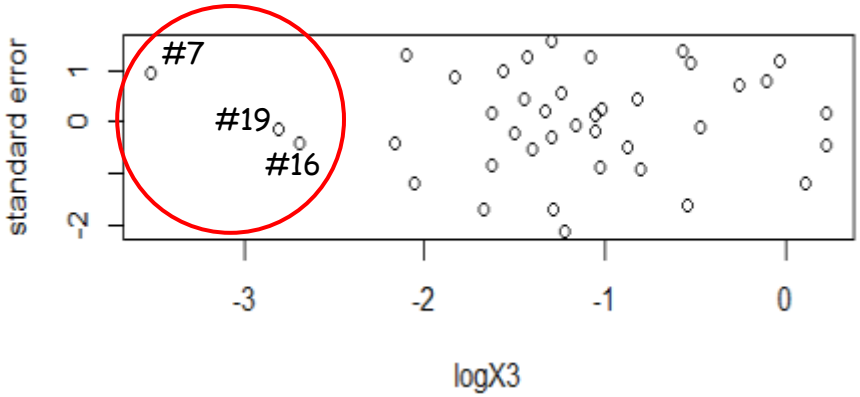
Valores de Y menores que 90 e maiores que 115 foram pouco amostrados!

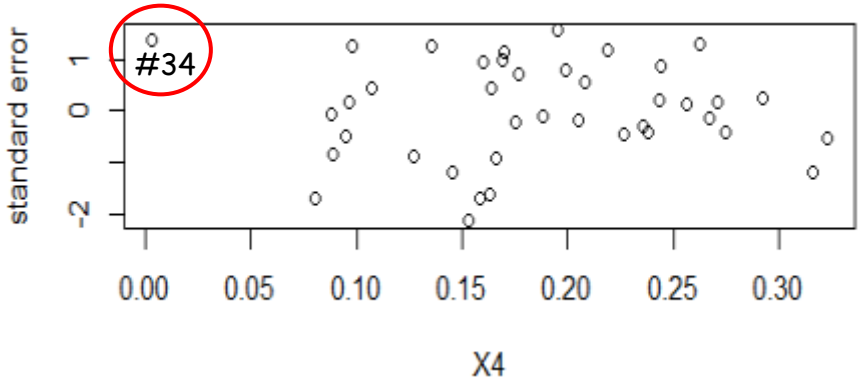
Exemplo

Avaliando a qualidade do modelo selecionado

```
plot(dados$X1,regsel$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")
```





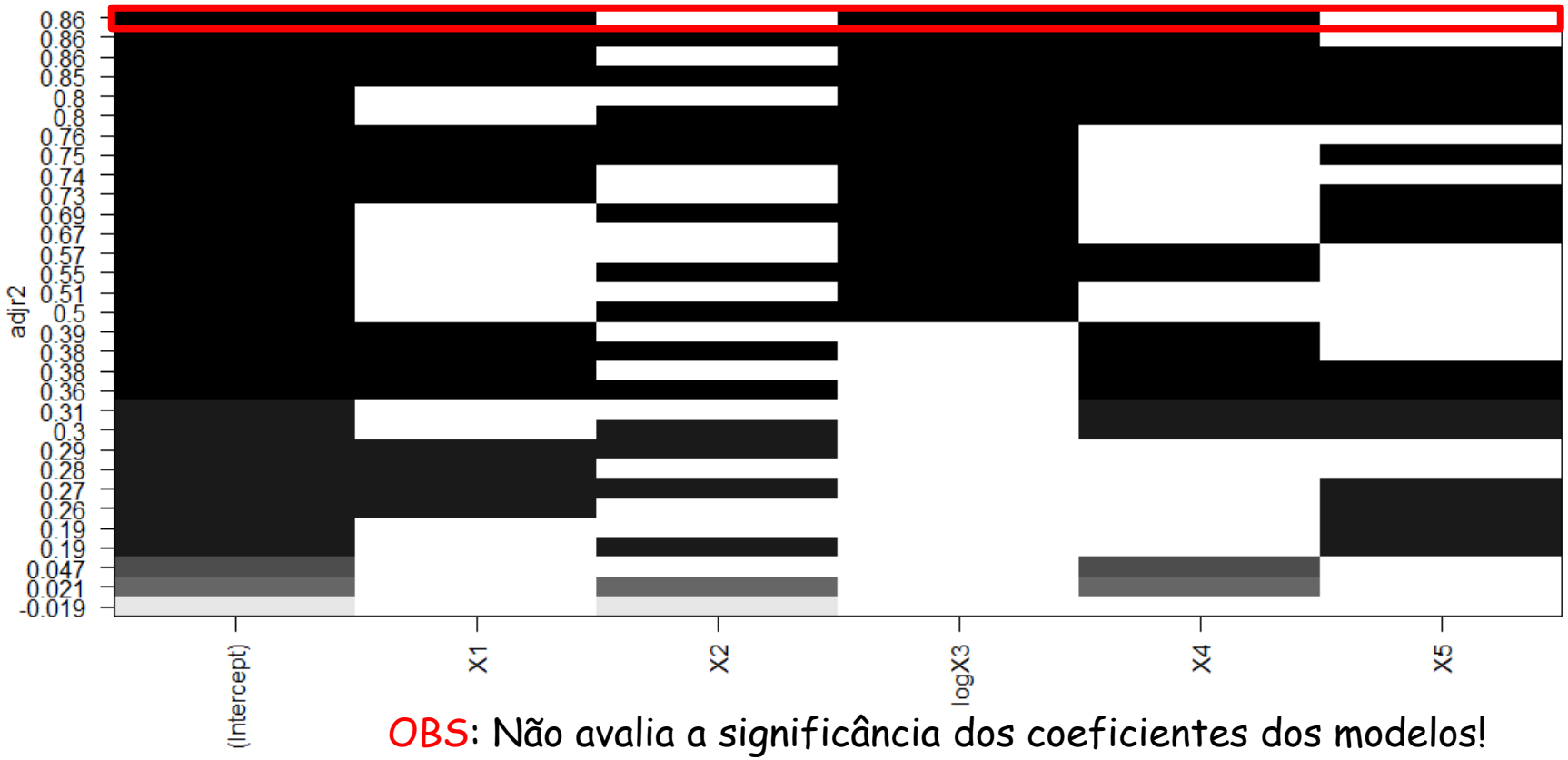


Exemplo

Procura exaustiva...

Melhor Modelo: $Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_3 \, logX_{3,i} + \beta_4 X_{4,i} + \xi_i$

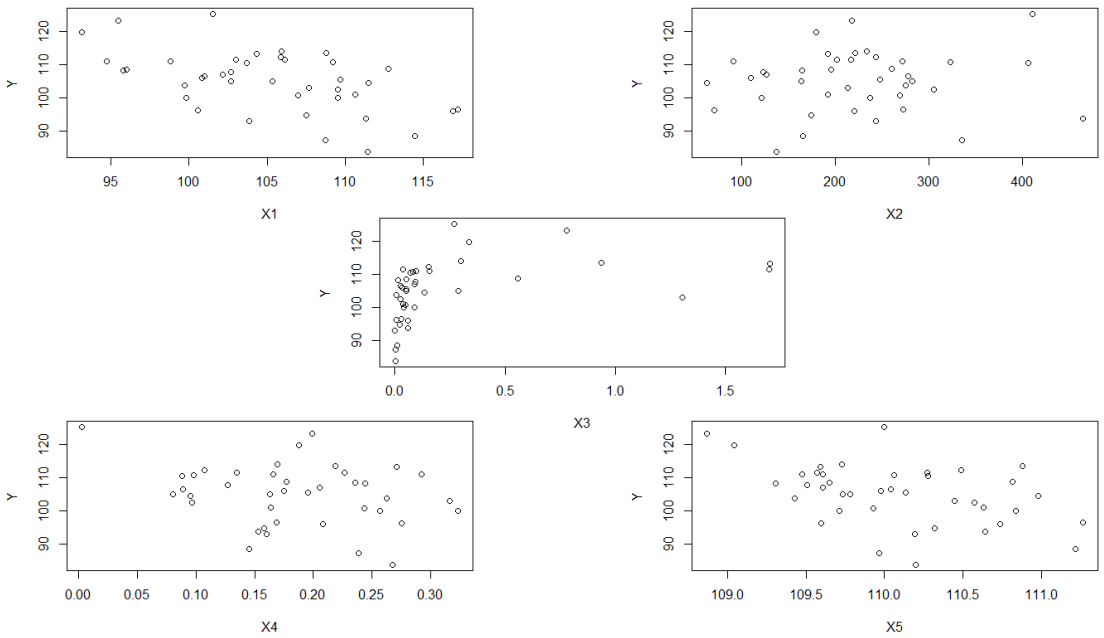
```
library(leaps)
leaps<-regsubsets(Y ~ X1 + X2 + logX3 + X4 + X5,data=dados,nbest=6)
plot(leaps,scale="adjr2")
```



OBS: Não avalia a significância dos coeficientes dos modelos!

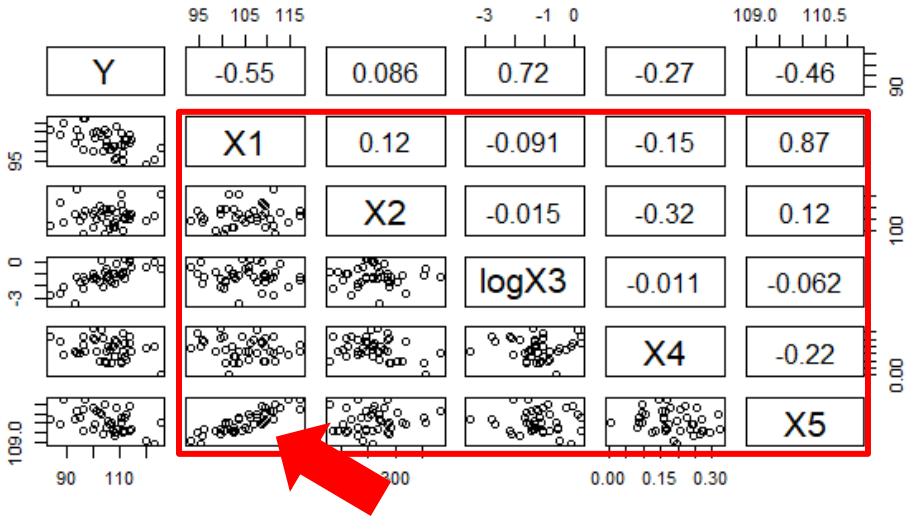
Considerações Finais

- 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.



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.



Considerações Finais

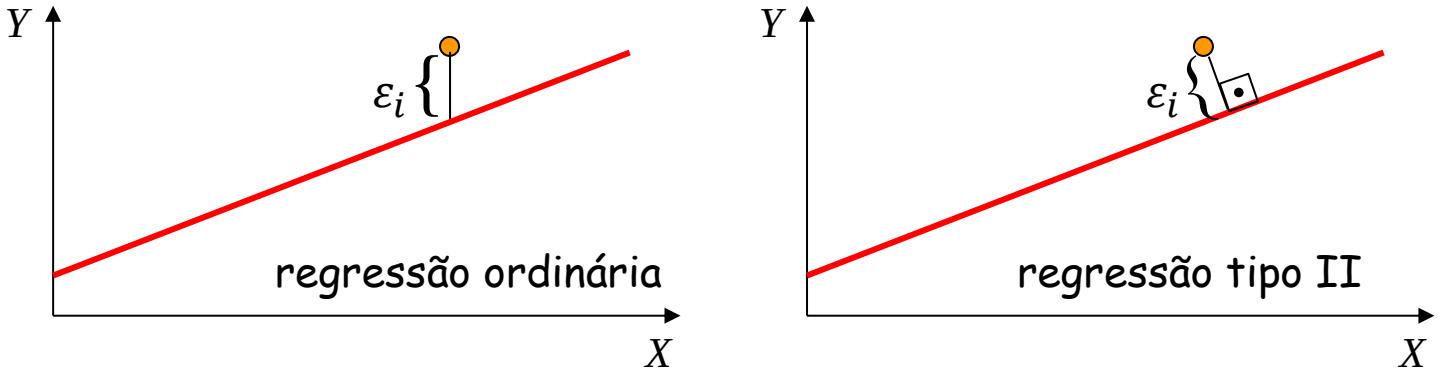
- A análise inicia-se pela avaliação dos diagramas de dispersão de cada variável independente versus dependente, buscando-se anomalias (*outliers*) e/ou falta de linearidade nas relações.
- A análise dos diagramas de dispersão entre variáveis independentes pode ser negligenciada mas pode indicar a presença 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 variância constante 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



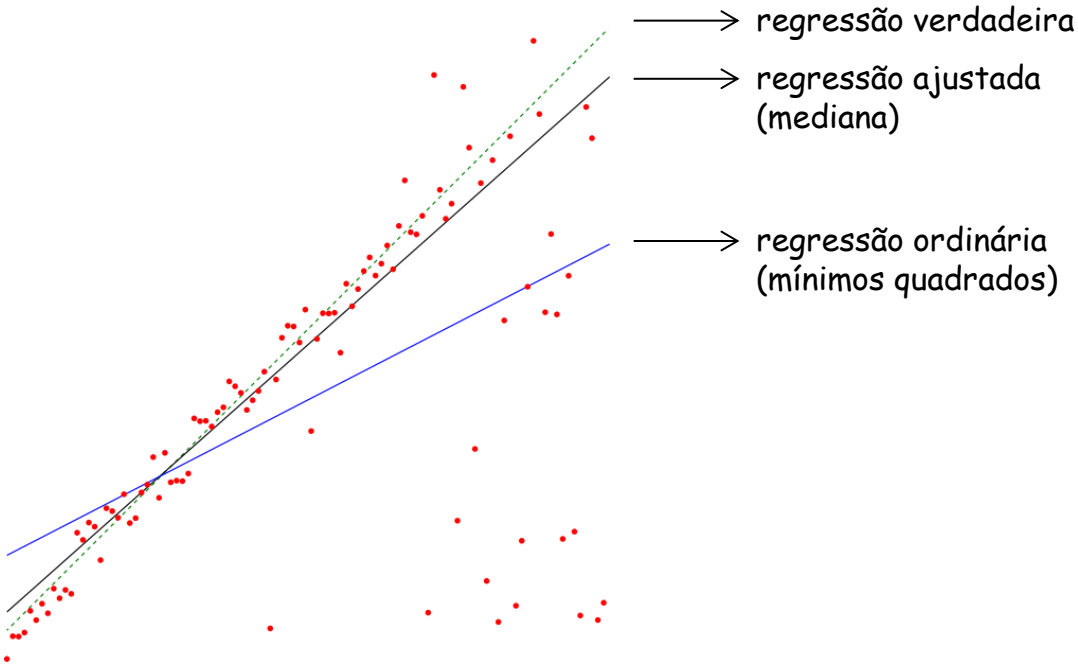
- Regressão não linear

  Nesse caso, a estimação dos coeficientes da regressão é feito 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*

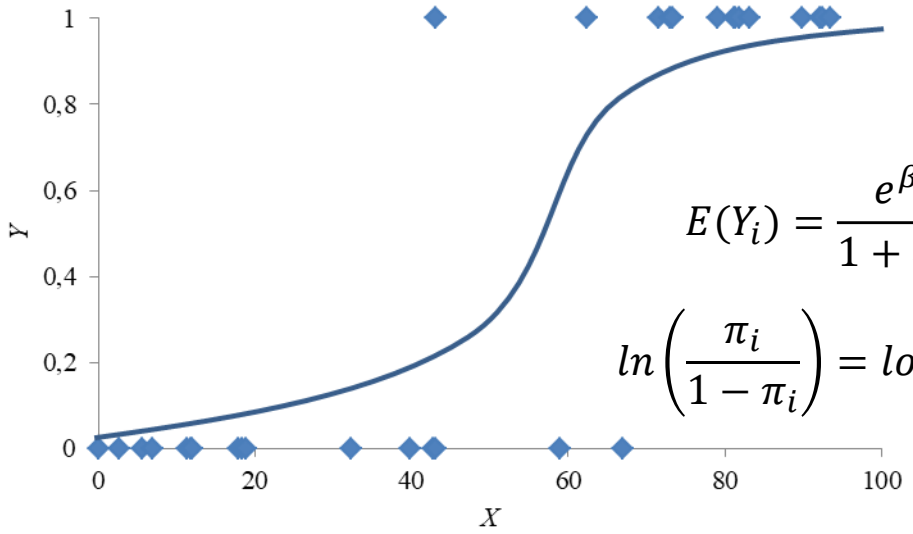


* https://en.wikipedia.org/wiki/Theil%E2%80%93Sen_estimator

Outras abordagens...

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

  Regressão Logística



$$E(Y_i) = \frac{e^{\beta_0+\beta_1 X_i}}{1 + e^{\beta_0+\beta_1 X_i}} = \pi_i$$

$$ln\left(\frac{\pi_i}{1-\pi_i}\right) = logit(\pi_i) = \beta_0 + \beta_1 X_i$$

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

  Regressão Beta

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

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

Outras abordagens...

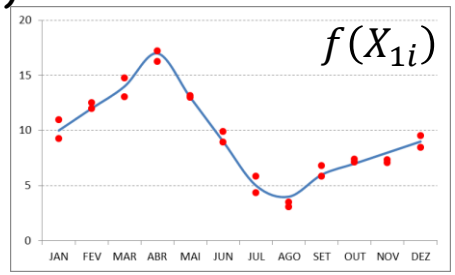
- Modelos Lineares Generalizados

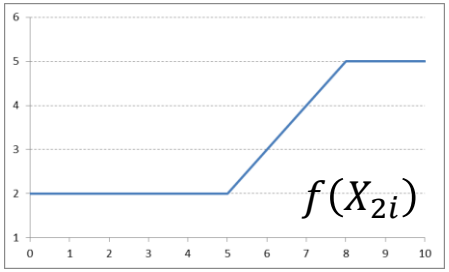
  É uma flexibilização da regressão linear ordinária para permitir que a variável dependente tenha resíduos com distribuição diferente da gaussiana

  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

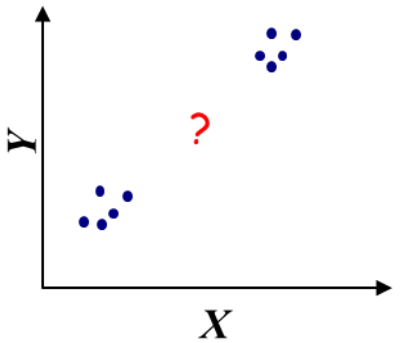
  Spatial Error Models (CAR): atribuem a autocorrelação espacial ao erro

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.