# Técnicas Computacionais em Probabilidade e Estatística I

## Aula VI

Chang Chiann
MAE 5704- IME/USP
1º Sem/2008

# Análise Exploratória

# Duas Variáveis Quantitativas

Lembre-se da construção destes gráficos para estudar relações entre variáveis:

- **Gráficos de Dispersão**
- **Gráficos P-P (Q-Q)**

**Coeficiente de Correlação Robusto**

**Regressão Robusta**

**Suavização**

# Modelo Probabilístico

a) Variáveis quantitativas discretas

(x,y): caracterizada por uma função de probabilidade

i) $P(x,y)=P(X=x,Y=y)$ : dist. Conjunta

ii) $P(X=x)=\sum_y P(X=x,Y=y)$: dist. Marginal de X;

iii) $P(Y=y)=\sum xP(X=x,Y=y)$: dist. Marginal de Y.

X e Y independentes $\rightarrow P(x,y)=P(x)*P(y)$

## Propriedades:

i) $P(x,y) \geq 0$;

ii) $\sum_x \sum_y P(x,y) = 1$;

iii) $P(a \leq X \leq b,\ c \leq Y \leq d) = \sum_x \sum_y P(x,y)$.

## Transformação:

$Z = h(x,y)$

$E(Z) = \sum_i \sum_j h(x_i, y_j) P(X = x_i, Y = y_j)$

Distribuições condicionais:

$P(Y = y/X = x) = P(X = x, Y = y)/P(X = x)$

se $P(X = x)>0$.

Média da distribuição condicional:

$E(Y/X=x) = \sum_y y\ P(Y=y/X=x)$

## b) Variáveis quantitativas contínuas

Uma função de densidade de probabilidade f(x,y), tal que:

i) $f(x,y) \geq 0$, para todos par(x,y);

ii) $\iint f(x,y)dxdy=1$;

iii) $P(a \leq X \leq b, c \leq Y \leq d)= \int_a^b \int_c^d f(x,y)dxdy$

X e Y são independentes $\rightarrow f(x,y)=f(x)f(y)$

Densidades marginais:

$f_x(x) = \int f(x,y)dy$

$f_y(y) = \int f(x,y)dx$

Distribuição condicional:

$f_{y/x}(y/x)=f(x,y)/f_x(x)$, $f_x(x)>0$.

Média condicional:

$E(y/X=x)= \int y f_{y/x}(y/x)dy$: função de x, curva de regressão de y sobre x.

# Vizualizando Dados Bivariados

**DADOS**

| Obs | X1 | Y1 | Y2 | Y3 | X2 | Y4 |
|---|---|---|---|---|---|---|
| 1 | 10 | 8.04 | 9.14 | 7.46 | 8 | 6.58 |
| 2 | 8 | 6.95 | 8.14 | 6.77 | 8 | 5.76 |
| 3 | 13 | 7.58 | 8.74 | 12.74 | 8 | 7.71 |
| 4 | 9 | 8.81 | 8.77 | 7.11 | 8 | 8.84 |
| 5 | 11 | 8.33 | 9.26 | 7.81 | 8 | 8.47 |
| 6 | 14 | 9.96 | 8.10 | 8.84 | 8 | 7.04 |
| 7 | 6 | 7.24 | 6.13 | 6.08 | 8 | 5.25 |
| 8 | 4 | 4.26 | 3.10 | 5.39 | 19 | 12.50 |
| 9 | 12 | 10.84 | 9.13 | 8.15 | 8 | 5.56 |
| 10 | 7 | 4.82 | 7.26 | 6.42 | 8 | 7.91 |
| 11 | 5 | 5.68 | 4.74 | 5.73 | 8 | 6.89 |

**Considere os dados apresentados por Frank Anscombe**

**11 observações de**

**6 variáveis**

# Vizualizando Dados Bivariados

**DADOS**

| Obs | X1 | Y1 | Y2 | Y3 | X2 | Y4 |
|---|---|---|---|---|---|---|
| 1 | 10 | 8.04 | 9.14 | 7.46 | 8 | 6.58 |
| 2 | 8 | 6.95 | 8.14 | 6.77 | 8 | 5.76 |
| 3 | 13 | 7.58 | 8.74 | 12.74 | 8 | 7.71 |
| 4 | 9 | 8.81 | 8.77 | 7.11 | 8 | 8.84 |
| 5 | 11 | 8.33 | 9.26 | 7.81 | 8 | 8.47 |
| 6 | 14 | 9.96 | 8.10 | 8.84 | 8 | 7.04 |
| 7 | 6 | 7.24 | 6.13 | 6.08 | 8 | 5.25 |
| 8 | 4 | 4.26 | 3.10 | 5.39 | 19 | 12.50 |
| 9 | 12 | 10.84 | 9.13 | 8.15 | 8 | 5.56 |
| 10 | 7 | 4.82 | 7.26 | 6.42 | 8 | 7.91 |
| 11 | 5 | 5.68 | 4.74 | 5.73 | 8 | 6.89 |

**Considere os dados apresentados por Frank Anscombe**

**CORRELAÇÕES (Pearson)**

Correlation of X1 and Y1 = 0.816

Correlation of X1 and Y2 = 0.816

Correlation of X1 and Y3 = 0.816

Correlation of X2 and Y4 = 0.817

$$\hat{Y}_j = 3 + 0.5\, X_j$$

**Mesmos valores de correlação**

**Mesma equação de ajuste**

**?**

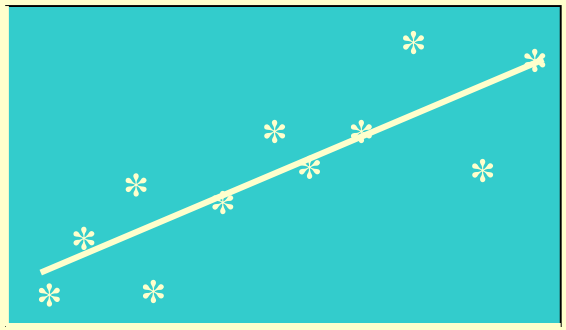
# Vizualizando Dados

**Arquivo F. Anscombe**

$$\hat{Y}_j = 3 + 0.5\, X_j$$

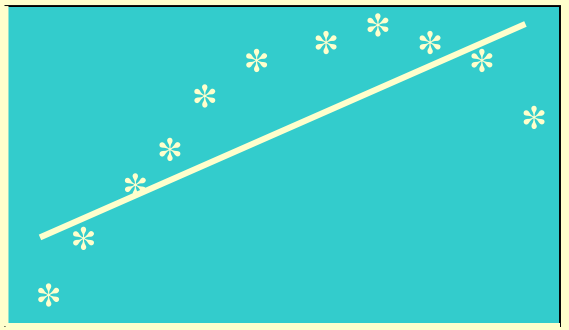
Tendência não linear

Y1



X1

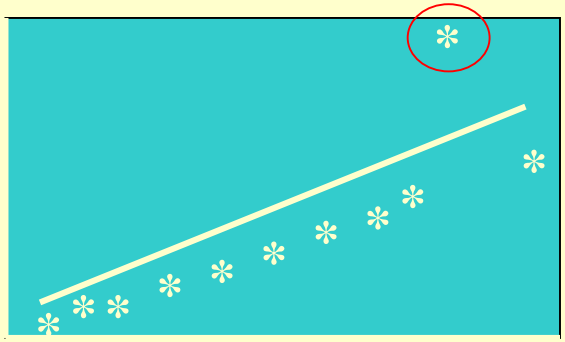
Y2



X1

Obs atípica

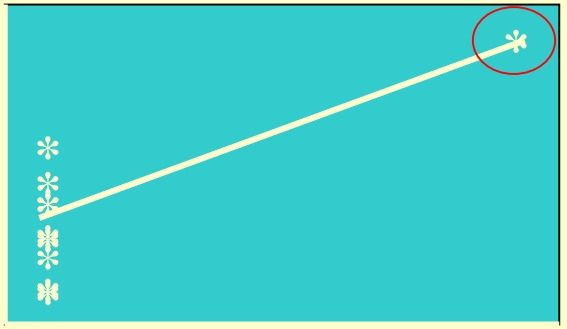
Y3



X1

Obs atípica

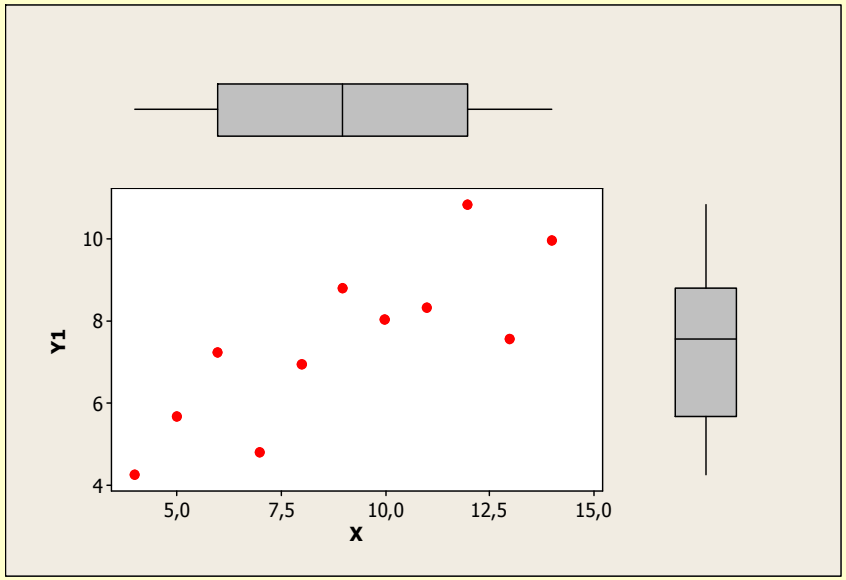
Y4

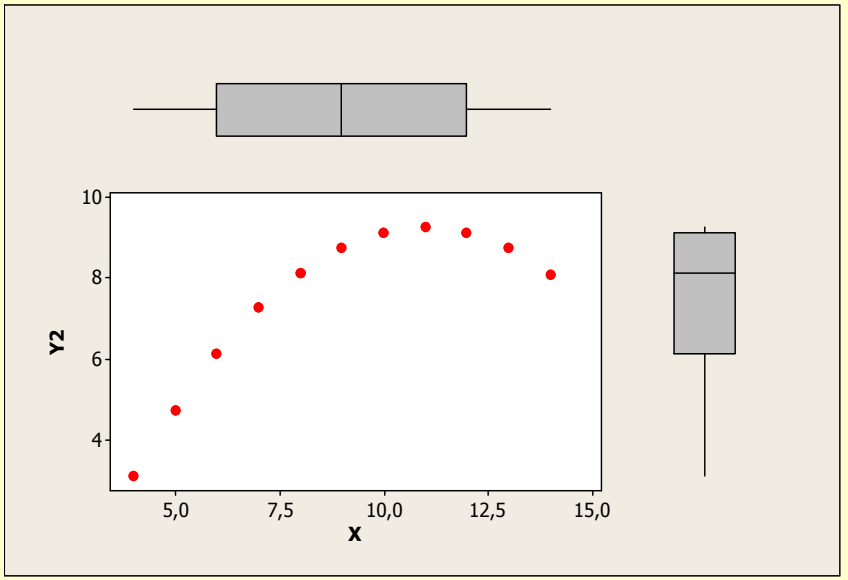


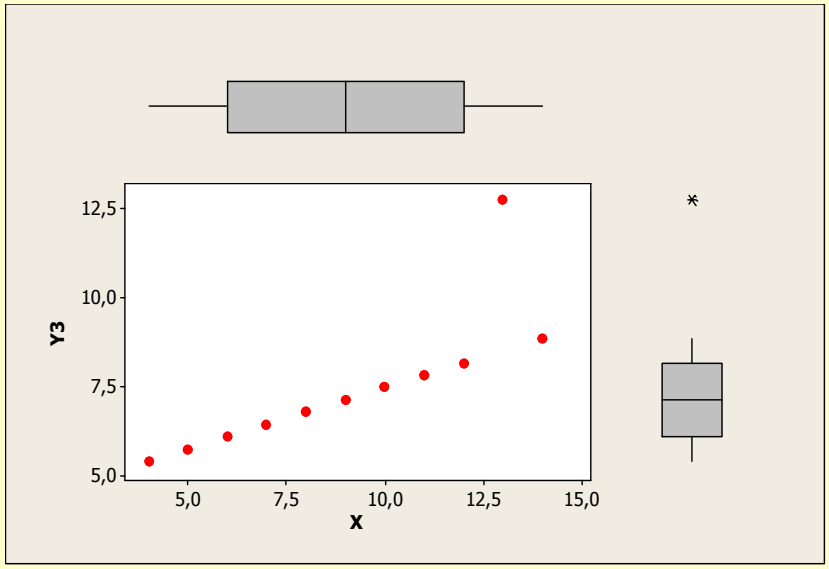
X4

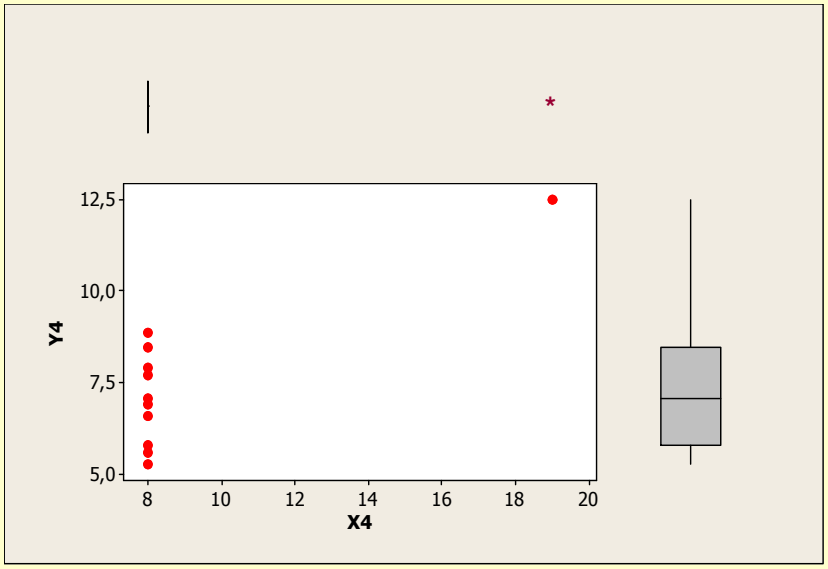
# Vizualizando Dados

## Arquivo F. Anscombe



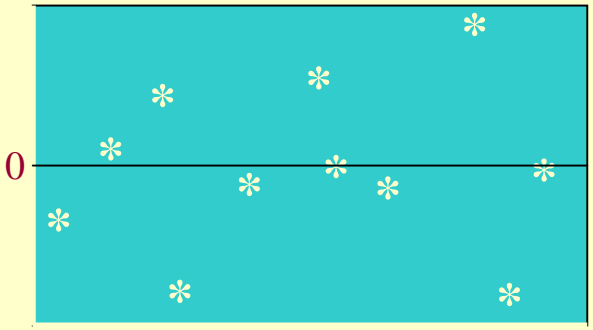



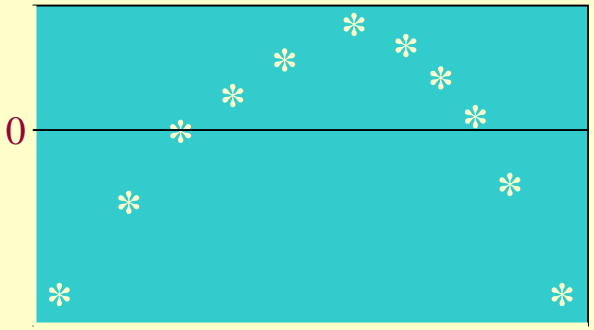



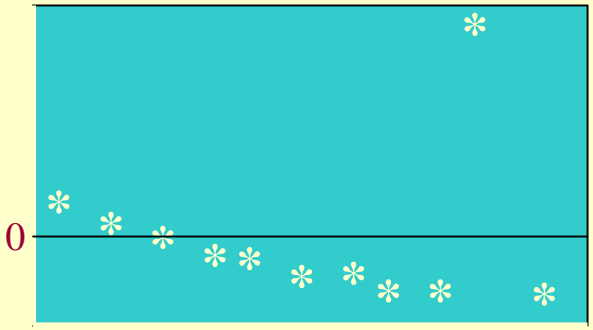



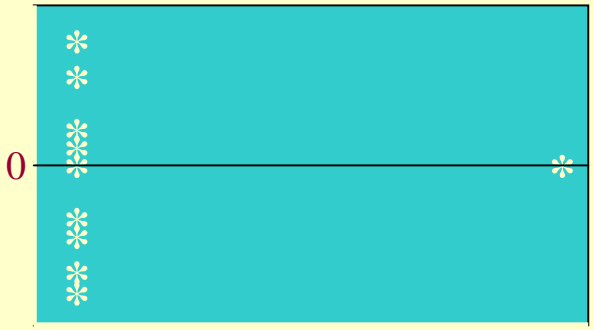
# Vizualizando Resíduos

***R = 66,7% ??***



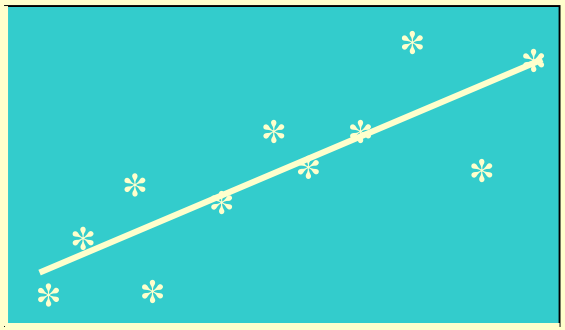




***O que se espera da Dispersão dos resíduos em função dos valores ajustados ??***

$$Y1 = 3 + 0{,}5\,X1$$



| Obs | X1 | Y1 | $\hat{Y}1$ | RESÍDUO |
|---|---|---|---|---|
| 1 | 10 | 8,04 | 8,0010 | 0,03900 |
| 2 | 8 | 6,95 | 7,0008 | -0,05082 |
| 3 | 13 | 7,58 | 9,5013 | -1,92127 |
| 4 | 9 | 8,81 | 7,5009 | 1,30909 |
| 5 | 11 | 8,33 | 8,5011 | -0,17109 |
| 6 | 14 | 9,96 | 10,0014 | -0,04136 |
| 7 | 6 | 7,24 | 6,0006 | 1,23936 |
| 8 | 4 | 4,26 | 5,0005 | -0,74045 |
| 9 | 12 | 10,84 | 9,0012 | 1,83882 |
| 10 | 7 | 4,82 | 6,5007 | -1,68073 |
| 11 | 5 | 5,68 | 5,5005 | 0,17945 |
| | | Total | | -5,41789E-14 |

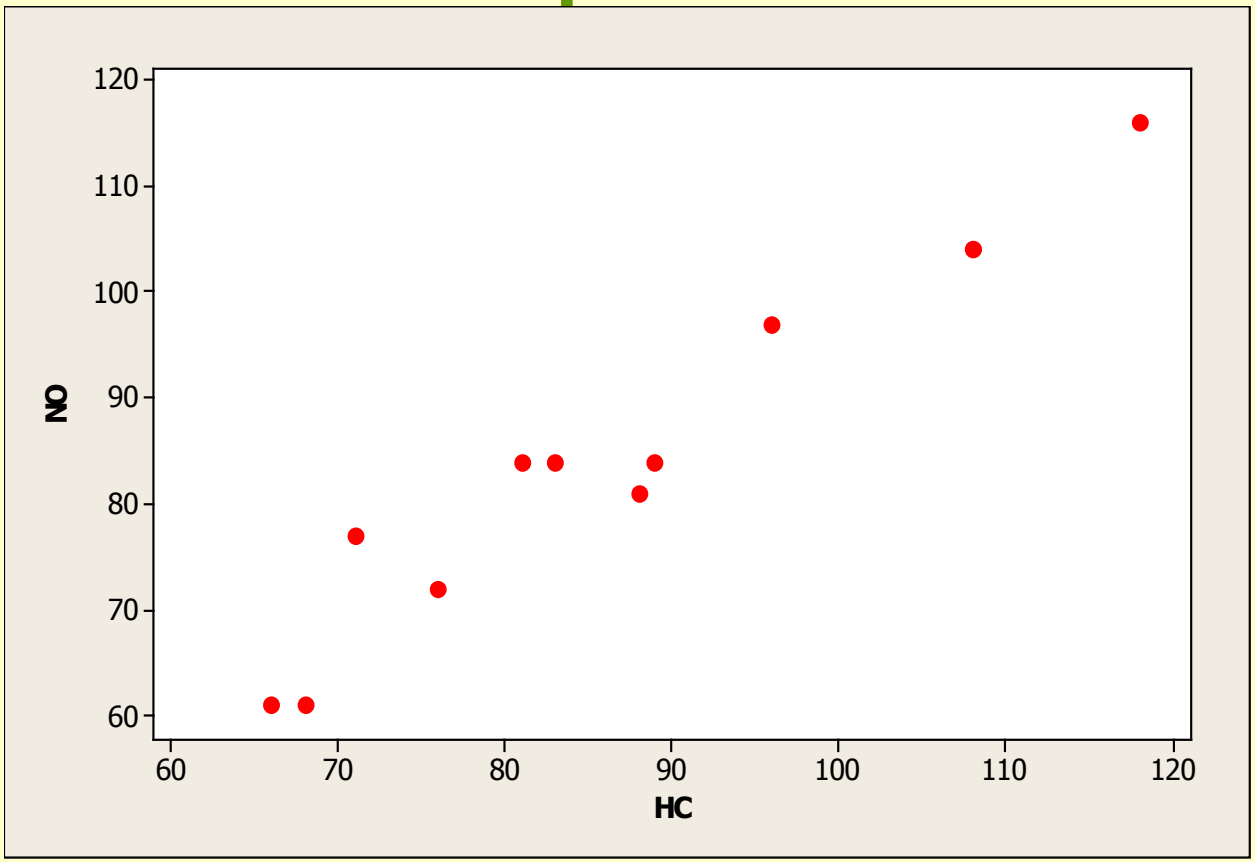
# Análise Exploratória

# Dados Bidimensionais

**Dados de poluentes ambientais**

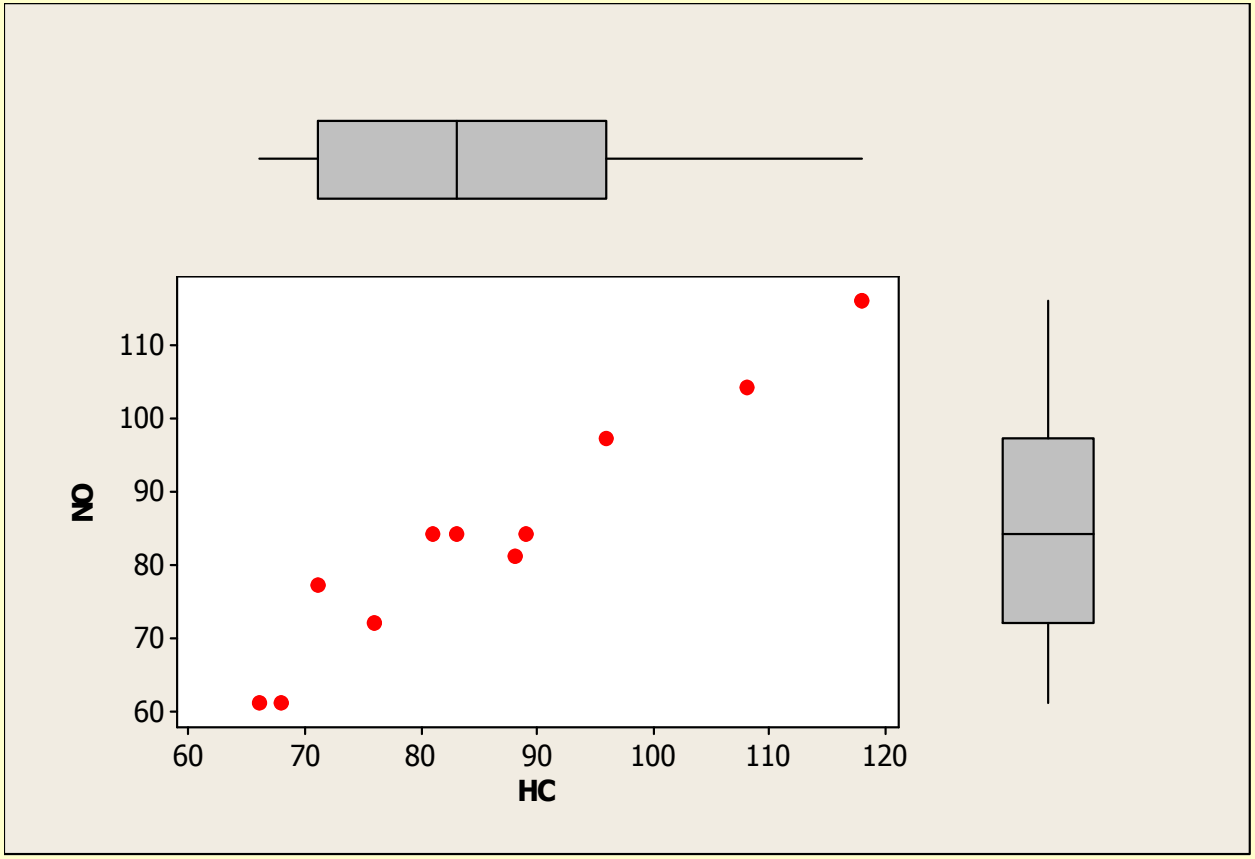
| Day | Nitrogen Oxides | Hydrocarbons |
|---|---|---|
| 1 | 104 | 108 |
| 2 | 116 | 118 |
| 3 | 84 | 89 |
| 4 | 77 | 71 |
| 5 | 61 | 66 |
| 6 | 84 | 83 |
| 7 | 81 | 88 |
| 8 | 72 | 76 |
| 9 | 61 | 68 |
| 10 | 97 | 96 |
| 11 | 84 | 81 |

# Análise Exploratória - Gráfico de Dispersão

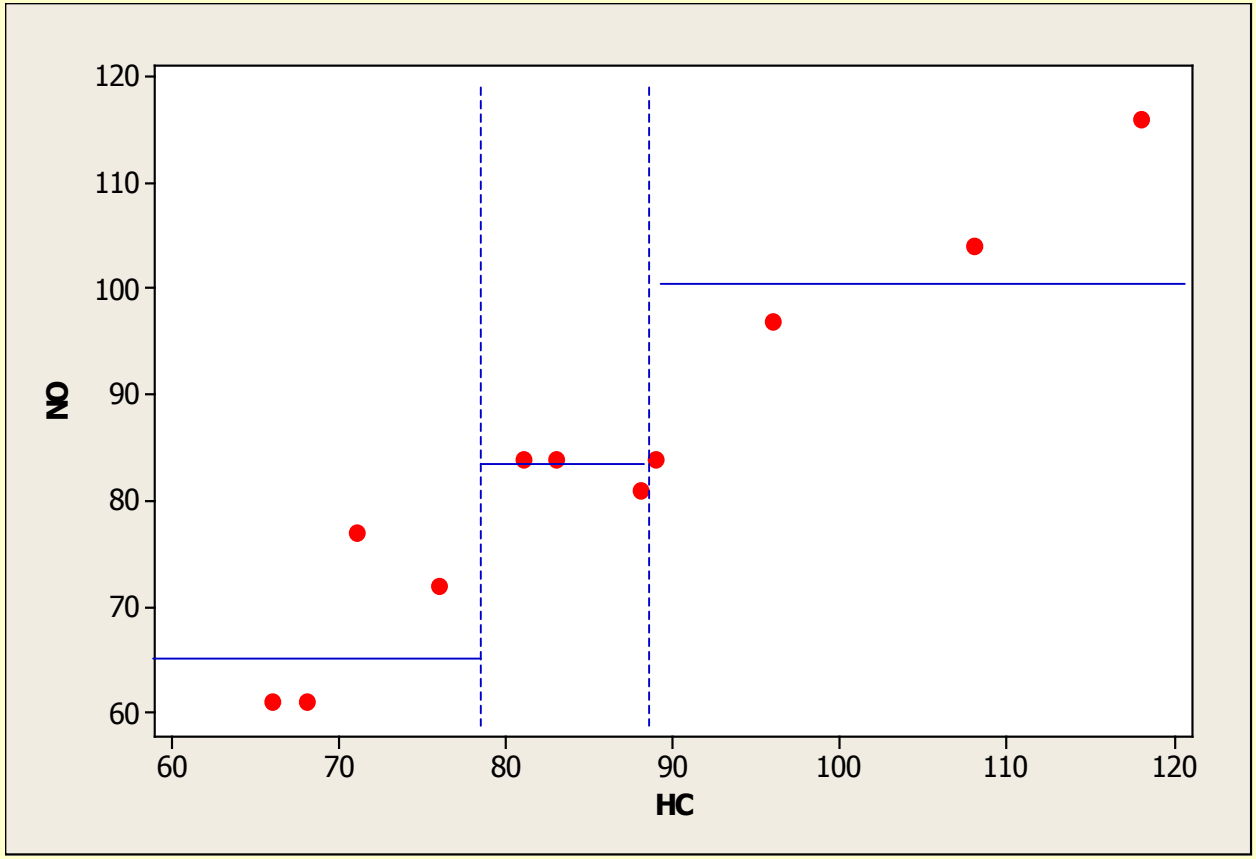



- Gráficos: *antes* que qualquer modelo seja ajustado
- Gráficos: *durante* o processo de ajuste de modelos
- Gráficos: *depois* que o modelo foi ajustado (*Chambers et al., 1992)*

# Vizualizando Dados Bivariados

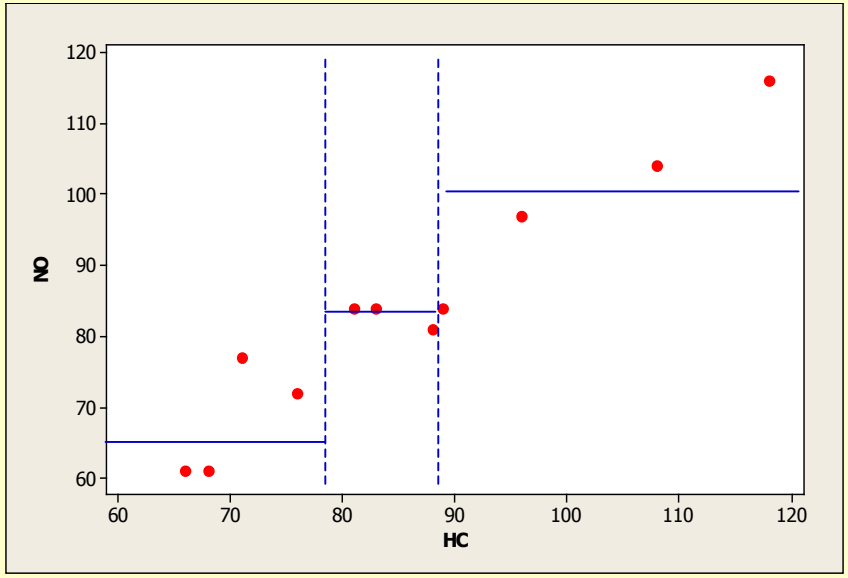


# Análise Exploratória - Gráfico de Dispersão




- Faixas verticais com igual número de obs (tanto quanto possível)
- Medianas (de Y) de cada faixa

# Análise Exploratória - Gráfico de Dispersão

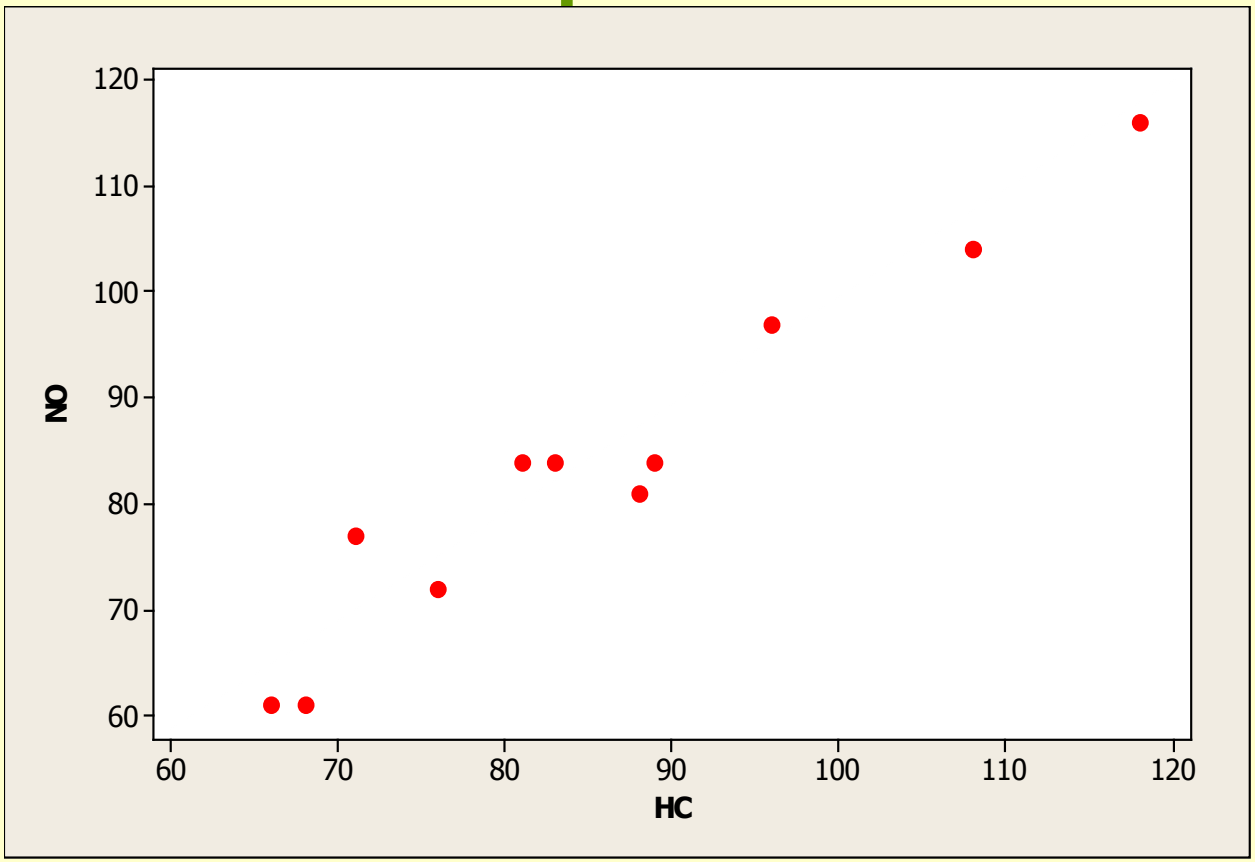



Medidas resumo para o poluente NO ("Y")

| | Faixas do poluente HC ("X") | | |
|---|---|---|---|
| | 66+ 79 | 79+ 89 | 89+ 119 |
| *n* | 4 | 3 | 4 |
| *Mediana* | 66,5 | 84 | 100,5 |
| *Média* | 67,8 | 83 | 100,3 |

- **Categorizando uma das variáveis ⇒ análise exploratória da variável quantitativa em cada nível da variável categorizada**
- **Medidas resumo e gráficos da variável quantitativa em cada nível da variável categorizada (depende de *n* em cada faixa)**

# Análise Exploratória - Gráfico de Dispersão

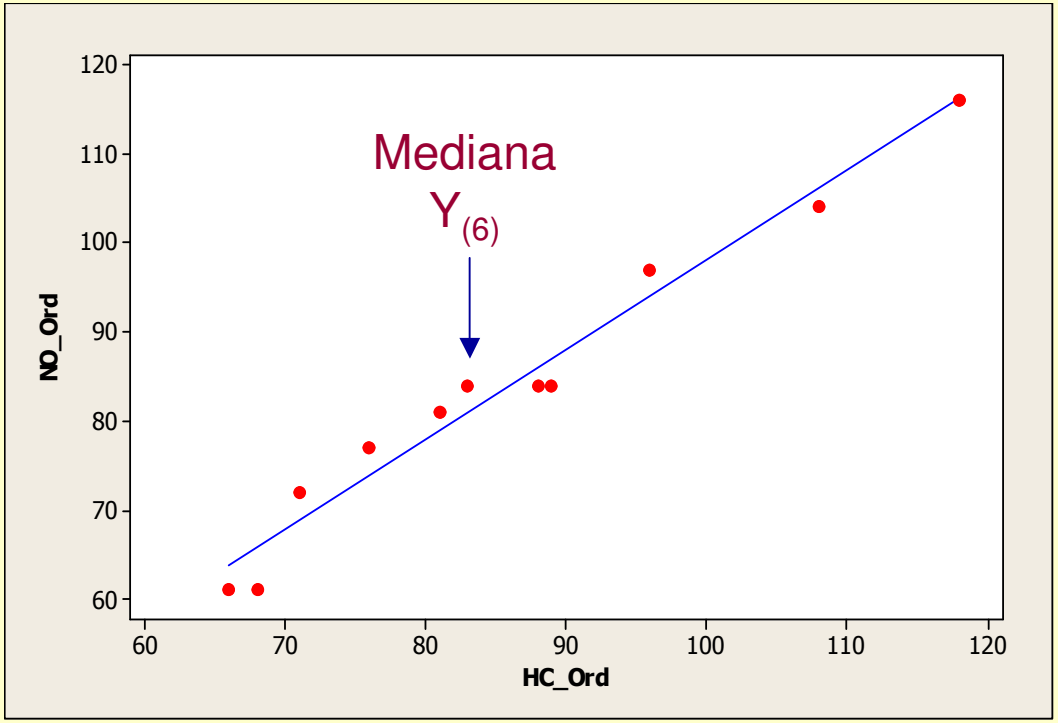



- Gráficos *de dispersão: relação sistemática global entre Y e X ⇒ não permitem comparar diretamente os valores pequenos, medianos e grandes dos dois conjuntos de dados*

⇒ *Gráficos Q-Q (P-P)*

# Gráfico Quantis-Quantis

Dados de Poluentes




*n=11*: pares de obs

⇒Gráfico P-P: $Y_{(j)}$ x $X_{(j)}$

das observações ordenadas

- os menores níveis do poluente HC foram superiores àqueles do poluente NO
- a magnitude dos valores dos dois poluentes, ao longo dos dias, não difere muito

Gráfico de Dispersão: pareamento das observações é pela variável que define a ocorrência dos dados (var. Day, neste caso)

Gráfico P-P: pareamento é pela ordem das observações (percentil/quantil)

P-P plot:

$P_x(q)=P(X\leq q)=F_x(q)$

$P_y(q)=P(Y\leq q)=F_y(q)$

Pares $(P_x(q),P_y(q))$, para qq q real $\rightarrow$ P-P plot: gráfico de probabilidades.

Q-Q plot:

Os pares $(Q_x(p),Q_y(p))$, p/ $0<p<1$ $\rightarrow$ Q-Q plot: gráfico de quantis vs quantis.

$Q_x(p)$: um valor que deixa 100p% das obs. à sua esquerda.

Se as distribuições de X e Y são iguais, então $F_x=F_y$ e as gráficos P-P e Q-Q resultam em retas y=x.

Em geral, os gráficos Q-Q são mais sensíveis a diferenças nas caudas das distribuições, se estas forem aproximadamente simétricas e com a aparência de uma normal.

Considere:

$x_1, x_2, ..., x_n$ e $y_1, y_2, ..., y_m$

a) Caso n=m:

cada par $(x_{(i)},y_{(i)})$ é calculado p/

$p=p_i=(i-0,5)/n$, $i = 1, ..., n$.

Supondo que $y=aX+b$ $\rightarrow$ distribuições são as mesmas, exceto por uma transformação linear.

$$P = P(X \leq Q_x(p)) = P(aX+b \leq aQ_x(p)+b) = P(Y \leq Q_y(p))$$

$$\downarrow$$

$$Q_y(p) = aQ_x(p)+b$$

$$\downarrow$$

$$\text{posição}(y_i) = a * \text{posição}(x_i)+b$$

$$\text{escala}(y_i) = |a| \, \text{escala}(x_i)$$

O gráfico Q-Q mostrará uma reta com inclinação a e intercepto b.



Essa propriedade não vale para gráficos P-P.

## b) Caso $m \neq n$

$x_{(1)}, \ldots, x_{(n)}$ e $y_{(1)}, \ldots, y_{(m)}$

i) Consideramos os valores ordenados $y_{(i)}$ calculados para $p_i=(i-0,5)/m$, i=1, ..., m

ii) Interpolamos um conjunto correspondente de quantis para xi. Queremos um valor j tal que:

$$(j-0,5)/n = (i-0,5)/m \rightarrow j = (i-0,5)n/m+0,5$$

Se j for inteiro:

Colocamos no gráfico Q-Q os pontos$(x_{(j)},y_{(i)})$

Se não:

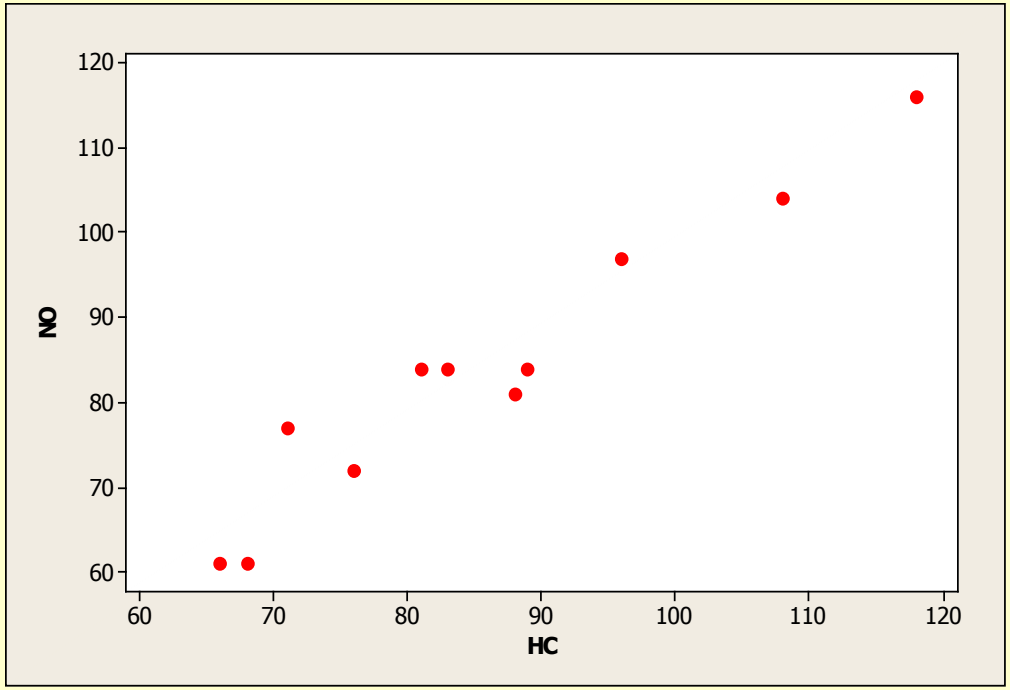
Considere $j=k+r$, com k inteiro e $0<r<1$.

O quantil correspondente a $y_{(i)}$ é

$Q_x((i-0,5)/m)=(1-r)\ x_{(k)}+r\ x_{(k+1)}$

# Análise Exploratória - Gráfico de Dispersão

Dados de Poluentes



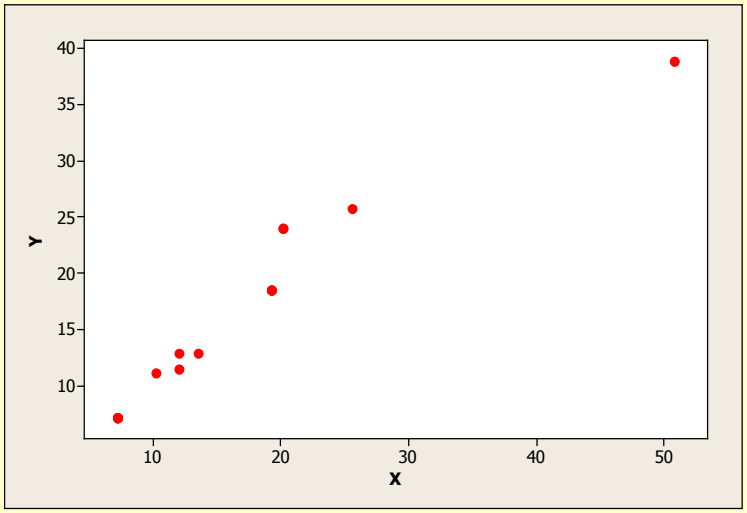

$$r = \frac{\sum_{j=1}^{n} \left(X_j - \bar{X}\right)\left(Y_j - \bar{Y}\right)}{\left[\sum_{j=1}^{n} \left(X_j - \bar{X}\right)^2 \sum_{j=1}^{n} \left(Y_j - \bar{Y}\right)^2\right]^{1/2}} \quad \Rightarrow \quad r = 0{,}967$$

O coeficiente de correlação de Pearson é uma medida não resistente

$\Rightarrow$ Como obter uma medida robusta de relação entre variáveis ?

# Coeficiente de Correlação Robusto

$\Rightarrow \quad r = 0{,}961$




| X | Y |
|---|---|
| 20,2 | 24 |
| 50,8 | 38,8 |
| 12 | 11,5 |
| 25,6 | 25,8 |
| 20,2 | 24 |
| 7,2 | 7,2 |
| 7,2 | 7,2 |
| 7,2 | 7,2 |
| 19,3 | 18,5 |
| 19,3 | 18,5 |
| 19,3 | 18,5 |
| 10,2 | 11,1 |
| 12 | 12,9 |
| 7,2 | 7,2 |
| 13,5 | 12,9 |

- Padronizar as variáveis pela variância aparada $\Rightarrow$ Obter somas e diferenças padronizadas

$$\tilde{X}_j = \frac{X_j}{s_X(\alpha)} \qquad \tilde{Y}_j = \frac{Y_j}{s_Y(\alpha)} \qquad j = 1,\ldots,n$$

$$\Rightarrow \quad r(\alpha) = \frac{s^2_{\tilde{X}+\tilde{Y}}(\alpha) - s^2_{\tilde{X}-\tilde{Y}}(\alpha)}{s^2_{\tilde{X}+\tilde{Y}}(\alpha) + s^2_{\tilde{X}-\tilde{Y}}(\alpha)}$$

$\overline{X}(\alpha = 0{,}05) = 14{,}86 \qquad \overline{Y}(\alpha = 0{,}05) = 15{,}33$

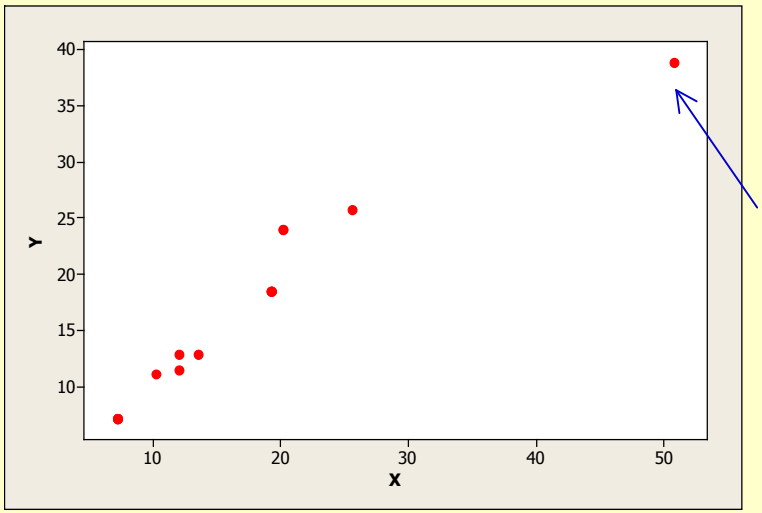
$s_X(\alpha = 0{,}05) = 5{,}87 \qquad s_Y(\alpha = 0{,}05) = 6{,}40$

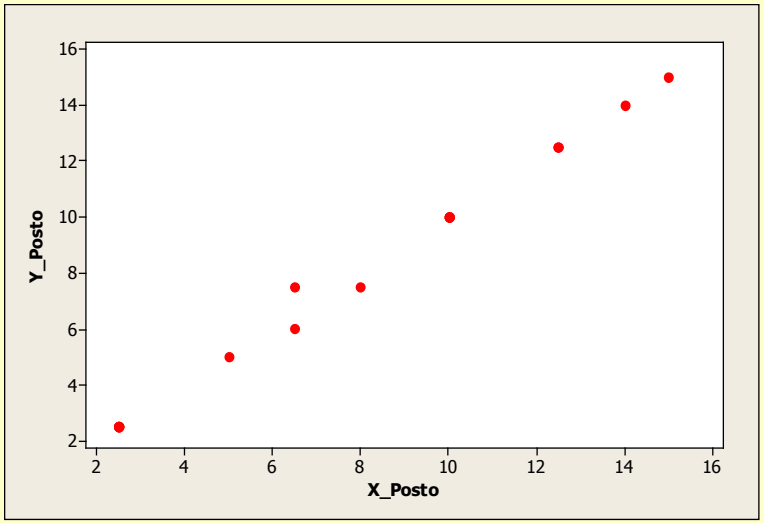
$\Rightarrow \quad r(\alpha = 0{,}05) = 0{,}973$

$\overline{(\tilde{X}+\tilde{Y})}(\alpha) = 4{,}93 \qquad \overline{(\tilde{X}-\tilde{Y})}(\alpha) = 0{,}14$

$s^2_{\tilde{X}+\tilde{Y}}(\alpha) = 3{,}93 \qquad s^2_{\tilde{X}-\tilde{Y}}(\alpha) = 0{,}054$

# Outros Coeficientes de Correlação Robustos




$\Rightarrow \quad r = 0{,}961$

$\Rightarrow \quad r(\alpha = 0{,}05) = 0{,}973$

**Coeficiente de Correlação de Spearman:**

**"atribuição de postos às obs em Y e em X, independentemente"**

$\Rightarrow \quad r_{Posto} = 0{,}997$

| X | Y |
|---|---|
| 20,2 | 24 |
| 50,8 | 38,8 |
| 12 | 11,5 |
| 25,6 | 25,8 |
| 20,2 | 24 |
| 7,2 | 7,2 |
| 7,2 | 7,2 |
| 7,2 | 7,2 |
| 19,3 | 18,5 |
| 19,3 | 18,5 |
| 19,3 | 18,5 |
| 10,2 | 11,1 |
| 12 | 12,9 |
| 7,2 | 7,2 |
| 13,5 | 12,9 |

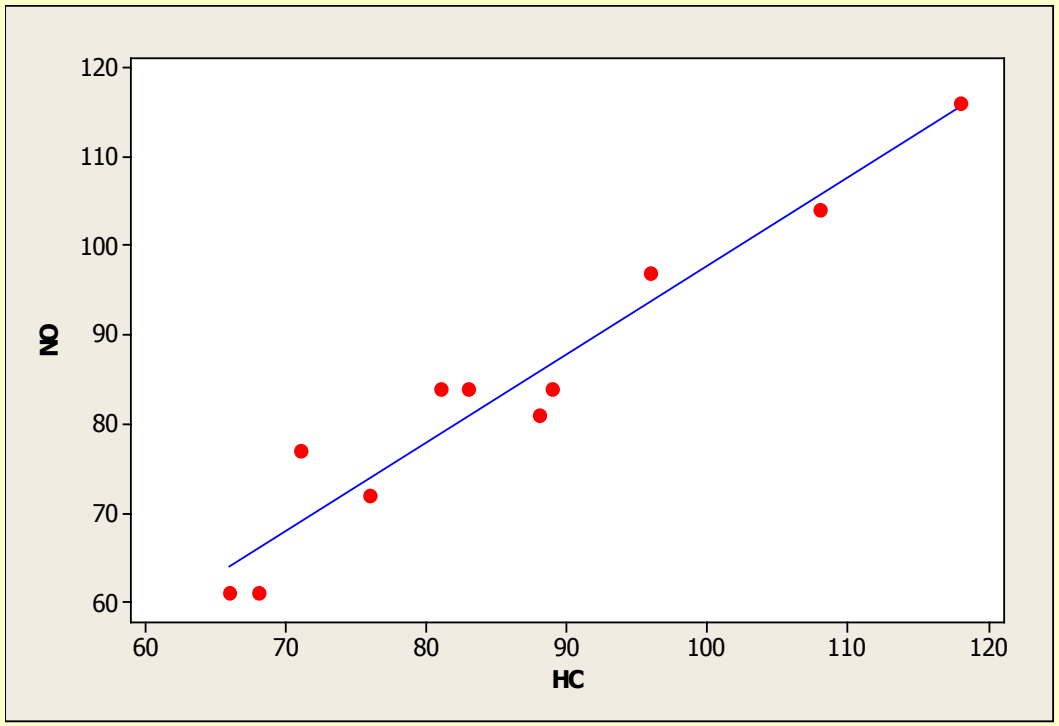
$\Rightarrow$

| X_Posto | Y_Posto |
|---|---|
| 12,5 | 12,5 |
| 15 | 15 |
| 6,5 | 6 |
| 14 | 14 |
| 12,5 | 12,5 |
| 2,5 | 2,5 |
| 2,5 | 2,5 |
| 2,5 | 2,5 |
| 10 | 10 |
| 10 | 10 |
| 10 | 10 |
| 5 | 5 |
| 6,5 | 7,5 |
| 2,5 | 2,5 |
| 8 | 7,5 |

# Análise Exploratória

# Regressão Linear

## Dados de Poluentes




**Propriedades do ajuste por Mínimos Quadrados ?**

# Modelos de Regressão

**Revisando**

$$Y_j = \beta_0 + \beta_1 X_j + \varepsilon_j \qquad j = 1,\ldots,n$$

$$Y_j = \beta_0 + e^{\beta_1 X_j} + \varepsilon_j \qquad j = 1,\ldots,n$$

$\Rightarrow$ Modelo não-linear

$$Y_j = \beta_0 + \sum_{k=1}^{K} \beta_k X_{kj} + \varepsilon_j \qquad j = 1,\ldots,n$$

$\Rightarrow$ Modelo linear aditivo

$$Y_j = \beta_0 + \sum_{k=1}^{K} \beta_k X_{kj} + \sum_k \sum_l \beta_{kl} X_{kj} X_{lj} + \varepsilon_j \qquad j = 1,\ldots,n$$

$$\Rightarrow \qquad Y_{n\times 1} = X_{n\times p}\beta_{p\times 1} + \varepsilon_{n\times 1}$$

Notação matricial: Y é vetor de respostas

X é matriz de variáveis explicativas

$\varepsilon$ é vetor de resíduos; $\varepsilon = ( Y - X\beta$

)

# Modelos de Regressão

Revisando

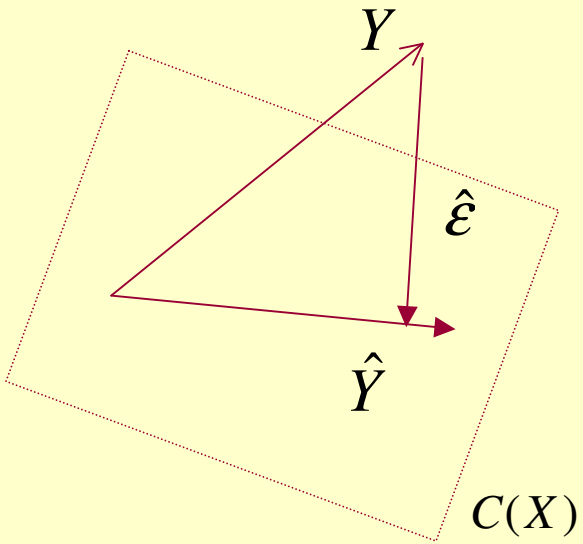
$$\Rightarrow \qquad Y_{n\times 1} = X_{n\times p}\beta_{p\times 1} + \varepsilon_{n\times 1}$$

$$\hat{\beta} \quad ; \quad \min_{\beta=\hat{\beta}} S(\beta) = (Y - X\beta)'(Y - X\beta)$$

$$\Rightarrow \qquad X'X\,\hat{\beta} = X'Y$$

$$\Rightarrow \qquad \hat{\beta} = (X'X)^{-1}X'Y$$

$$\Rightarrow \qquad \hat{Y} = X\,\hat{\beta} = X\,(X'X)^{-1}X'Y = H\,Y$$




Projeção *HY*

Mínimos Quadrados Ponderados $\hat{\beta}_W \quad ; \quad \min_{\beta=\hat{\beta}} S(\beta) = (Y - X\beta)'W(Y - X\beta)$

# Modelos de Regressão

**Revisando**

$$Y_{n\times 1} = X_{n\times p}\beta_{p\times 1} + \varepsilon_{n\times 1} \quad \Rightarrow \quad \hat{\beta} = (X'X)^{-1}X'Y$$ é estimador de M.Q.Ordinários

$$\Rightarrow \quad \hat{\beta}_1 = \frac{\sum_{j=1}^{n}(x_j - \bar{x})(y_j - \bar{y})}{\sum_{j=1}^{n}(x_j - \bar{x})^2} = \frac{\sum_{j=1}^{n}(x_j - \bar{x})y_j}{\sum_{j=1}^{n}(x_j - \bar{x})^2} = \sum_{j=1}^{n} w_j y_j \qquad \Rightarrow \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$$

- **Teorema de Gauss-Markov:**

  EMQ têm variância mínima na classe dos estimadores não viesados e que sejam funções lineares das observações

# Modelos de Regressão


$$Y_{n\times 1} = X_{n\times p}\beta_{p\times 1} + \varepsilon_{n\times 1} \quad \Rightarrow \quad \hat{\beta} = (X'X)^{-1}X'Y$$ é estimador de MQO

- **Supondo que a variável explicativa é fixa e que os erros são *iid* $(0;\sigma^2)$:**

$$\Rightarrow \quad E(\hat{\beta}_1) = \beta_{1,} \qquad E(\hat{\beta}_0) = \beta_0,$$

$$\Rightarrow \quad Var(\hat{\beta}_0) = \frac{\sigma^2 \sum x_j^2}{n\sum (x_j - \bar{x})^2}, \qquad Var(\hat{\beta}_1) = \frac{\sigma^2}{\sum (x_j - \bar{x})^2}, \qquad \hat{\sigma}^2 = Var(\hat{\varepsilon}) = \frac{S(\hat{\beta}_0, \hat{\beta}_1)}{n-2}$$

$$\Rightarrow \quad \text{se } \varepsilon_j \sim N: \quad y_j \sim N(\beta_0 + \beta_1 x_j; \sigma^2) \text{ e } \hat{\beta} \text{ são EMVS}$$

$$\Rightarrow \quad t_{\hat{\beta}_0} = \frac{\hat{\beta}_0 - \beta_0}{\hat{\sigma}} \sqrt{\frac{n\sum (x_j - \bar{x})^2}{\sum x_j^2}} \sim t_{n-2} \qquad t_{\hat{\beta}_1} = \frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}} \sqrt{\sum (x_j - \bar{x})^2} \sim t_{n-2}$$

# Modelos de Regressão

Revisando

$$Y_j = \bar{Y} \quad + \quad \left(Y_j - \hat{Y}_j\right) \quad + \quad \left(\hat{Y}_j - \bar{Y}\right)$$

$$\underbrace{\sum_{j=1}^{n}\left(Y_j - \bar{Y}\right)^2}_{SQTotal} = \underbrace{\sum_{j=1}^{n}\left(Y_j - \hat{Y}_j\right)^2}_{SQRes} + \underbrace{\sum_{j=1}^{n}\left(\hat{Y}_j - \bar{Y}\right)^2}_{SQMod = \hat{\beta}_1^2 \sum_{j=1}^{n}(x_j - \bar{x})^2}$$

$$\Rightarrow \qquad R^2 = \frac{QMMod}{QMTotal}$$

porcentagem da variabilidade de y explicada pelo modelo

$$\Rightarrow \qquad F = \frac{QMMod}{QM\,\mathrm{Re}\,s} = \frac{\sum(\hat{y}_j - \bar{y})^2 / 1}{\sum(y_j - \hat{y}_j)^2 / (n-2)} \sim F_{1;(n-2)}$$

# Modelos de Regressão

# Análises Exploratórias


- Análises de Resíduos

$\Rightarrow$ Histogramas e box-plots dos resíduos

$\Rightarrow$ Quantis dos resíduos contra quantis da normal

} Simetria, Normalidade

$\Rightarrow$ $\hat{\varepsilon}_j \times \text{ordem(obs)}$; $\hat{\varepsilon}_j \times x_j$; $\hat{\varepsilon}_j \times \hat{y}_j$; $y_j \times \hat{y}_j$

erros independentes

erros independentes de X

$\varepsilon_j = \varepsilon(x_j)$?

homocedasticidade

tendências não modeladas

falta de ajuste

$\varepsilon = y - \hat{y}$

$= (y - \bar{y}) + (\bar{y} - \hat{y})$

# Modelos de Regressão

# Análises Exploratórias


- Análises de Resíduos

  $\Rightarrow$ Histogramas e box-plots dos resíduos

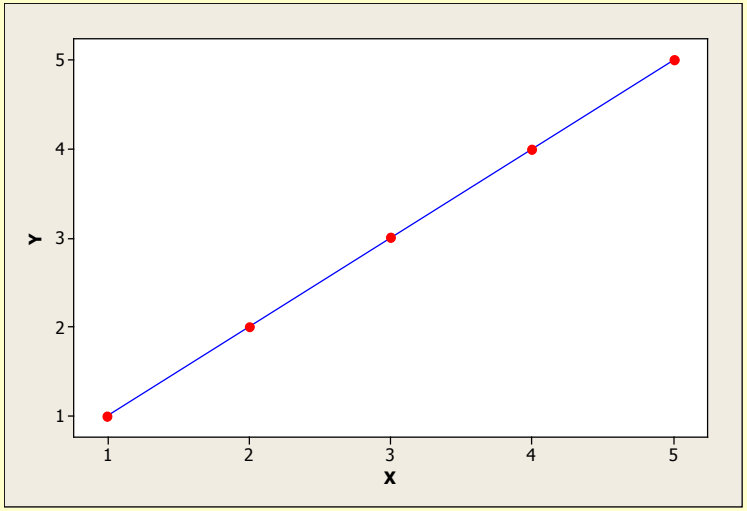
  $\Rightarrow$ Quantis dos resíduos contra quantis da normal

  $\Rightarrow$ $\hat{\varepsilon}_j \times x_j$; $\hat{\varepsilon}_j \times$ ordem(obs); $\hat{\varepsilon}_j \times \hat{y}_j$; $y_j \times \hat{y}_j$

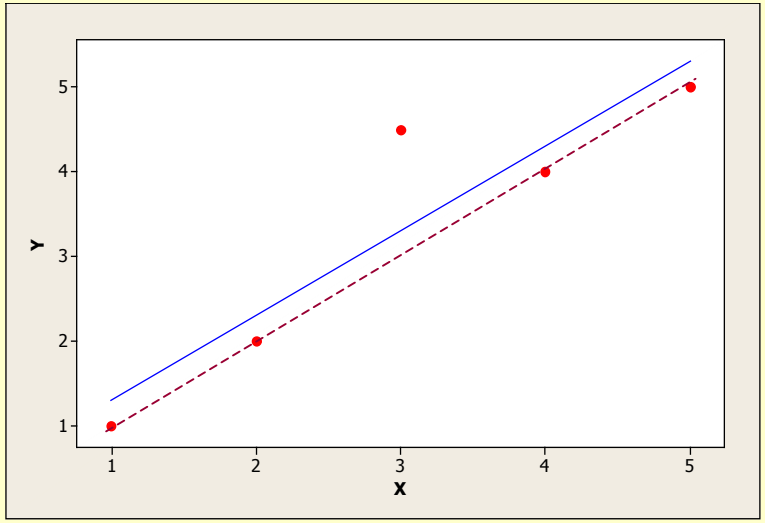
- Identificação de Pontos Aberrantes
- Identificação de Pontos de Alavanca
- Identificação de Pontos de Influência
- ...

# Pontos Discrepantes (Atípicos)

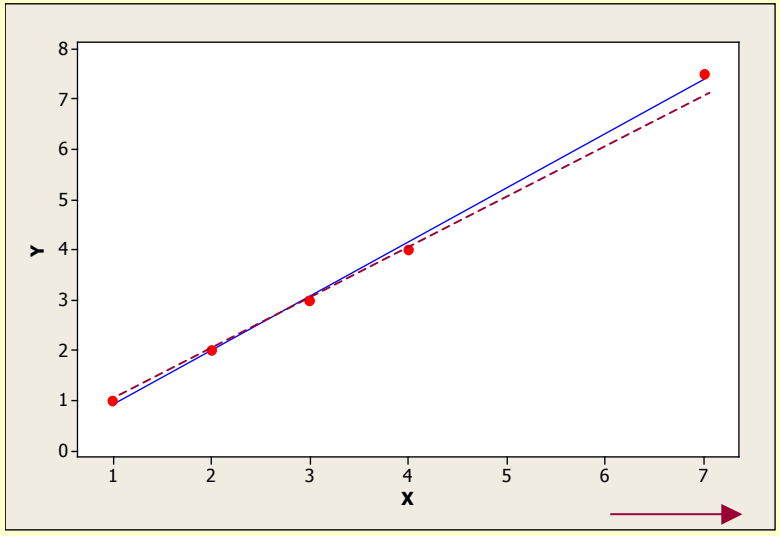
## Ajuste perfeito



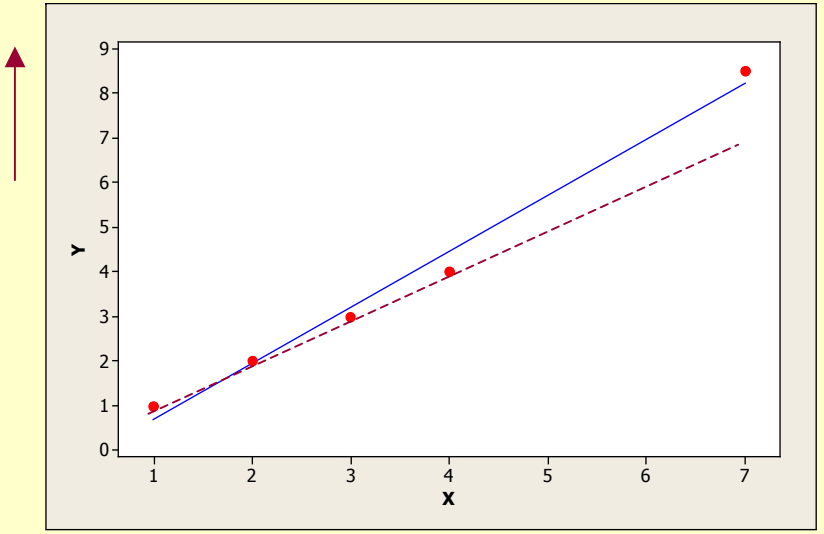

## Ponto aberrante




## Ponto de Alavanca




## Ponto Influente (e de Alavanca)



# Modelos de Regressão


$$\Rightarrow \qquad Y_{n\times 1} = X_{n\times p}\beta_{p\times 1} + \varepsilon_{n\times 1} \qquad \Rightarrow \qquad \hat{Y} = X\,\hat{\beta} = X\left(X'X\right)^{-1}X'Y = H\,Y$$

- Identificação de pontos de alavanca (alto *leverage*):

$$\hat{y}_j = \sum_{j=1}^{n} x_j'\left(X'X\right)^{-1}x_j\,y_j \qquad \Rightarrow \qquad \hat{y}_j = h_{jj}\,y_j + \left(1-h_{jj}\right)X_j'\,\hat{\beta}_{(j)}$$

alavanca do valor ajustado $\left(h_{jj} > {}^{2p}/_{n}\right)$

- Identificação de pontos aberrantes:

$$t_j^* = \frac{\hat{\varepsilon}_j}{s_{(j)}\left(1-h_{jj}\right)^{1/2}} \sim t_{n-p-1}$$

resíduo studentizado (deletado)

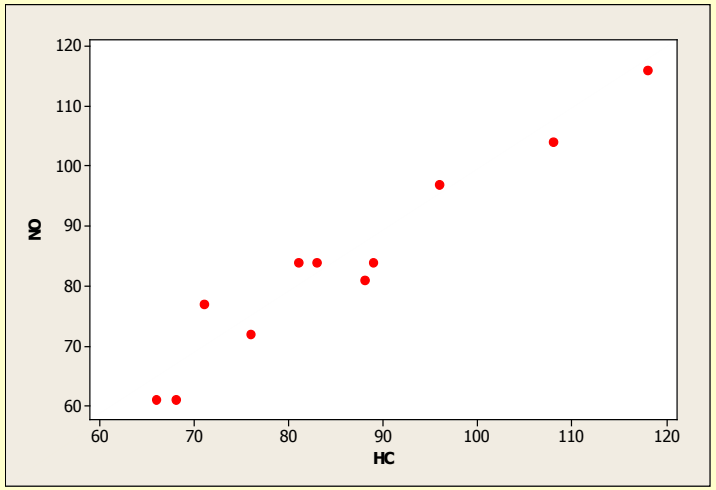
- Identificação de pontos influentes (Cook):

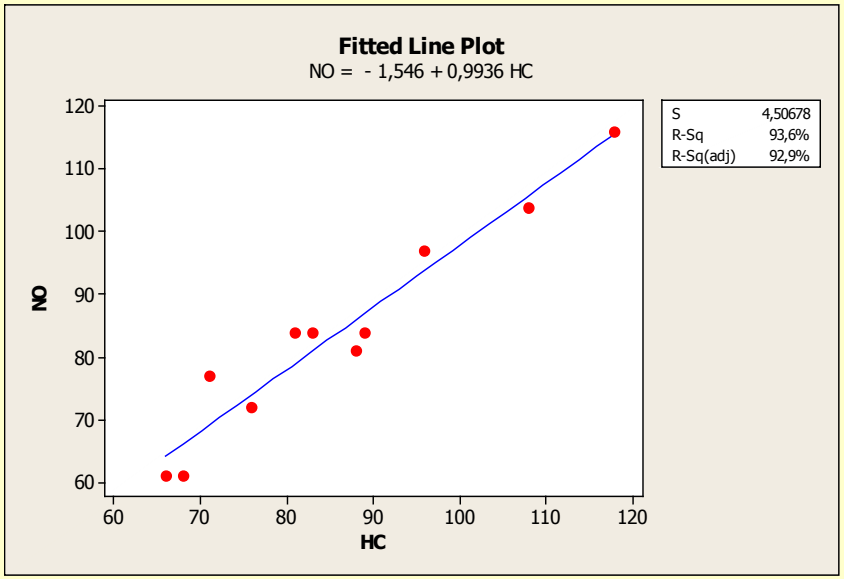
$$\uparrow t_j^2 \text{ e/ou } \uparrow h_{jj} \Rightarrow \uparrow D_j$$

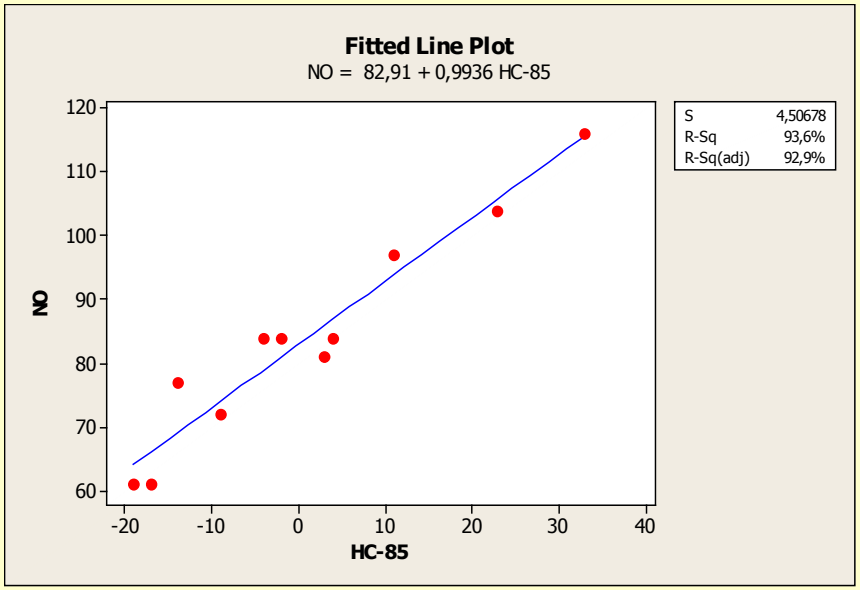
$$D_j = \frac{\left(\hat{\beta}-\hat{\beta}_{(j)}\right)'X'X\left(\hat{\beta}-\hat{\beta}_{(j)}\right)}{p\,s^2} \quad > \quad F_{p,(n-p-1)}(1-\alpha)$$

# Análise Exploratória - Regressão Linear

## Dados de Poluentes

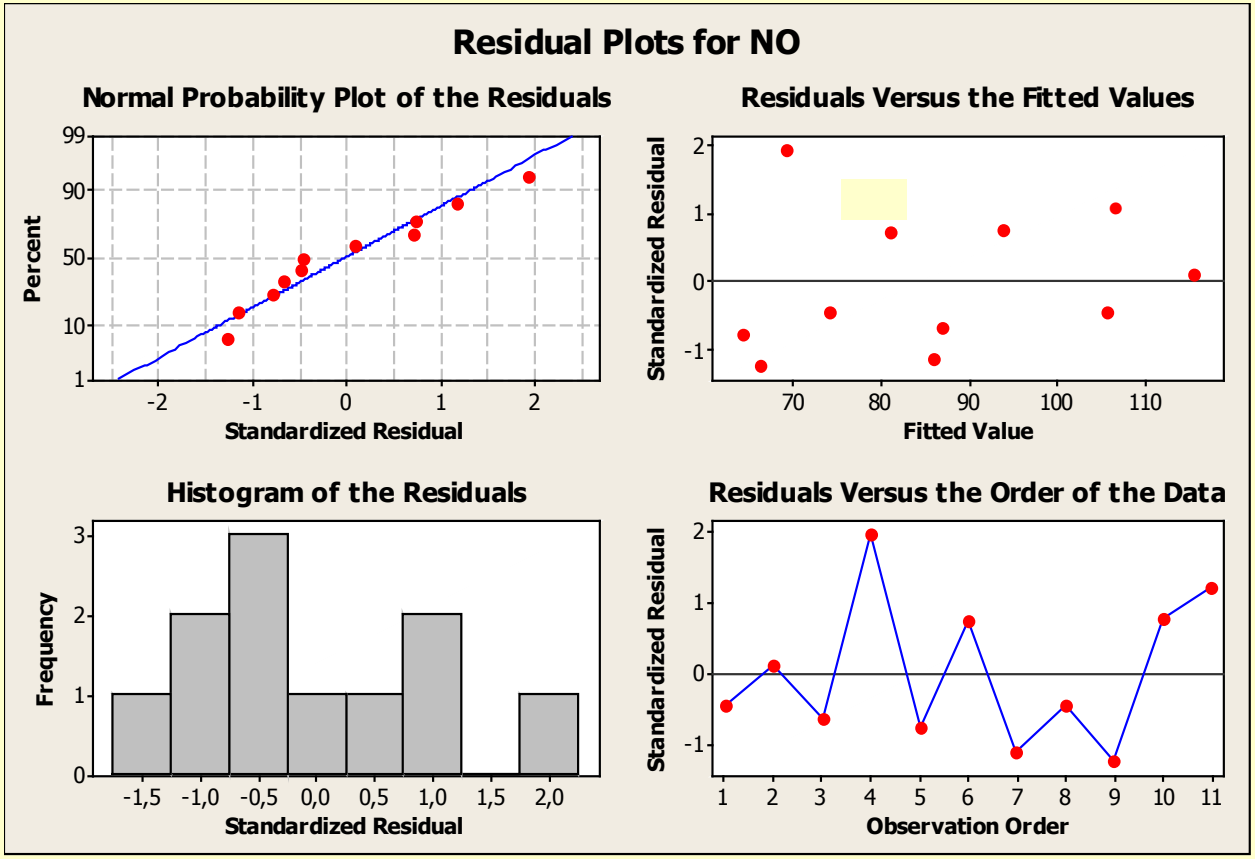


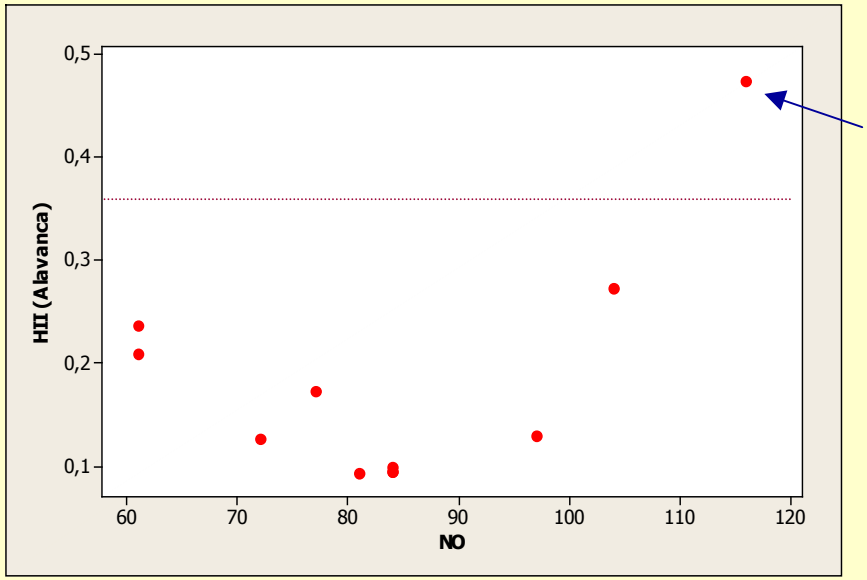
# Análise Exploratória - Regressão Linear

## Análise de Resíduos

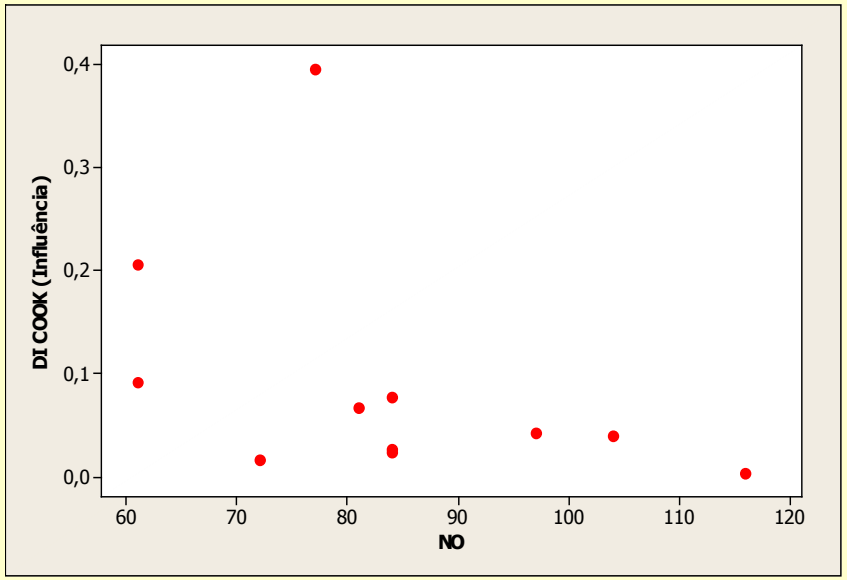


# Análise Exploratória - Regressão Linear Simples

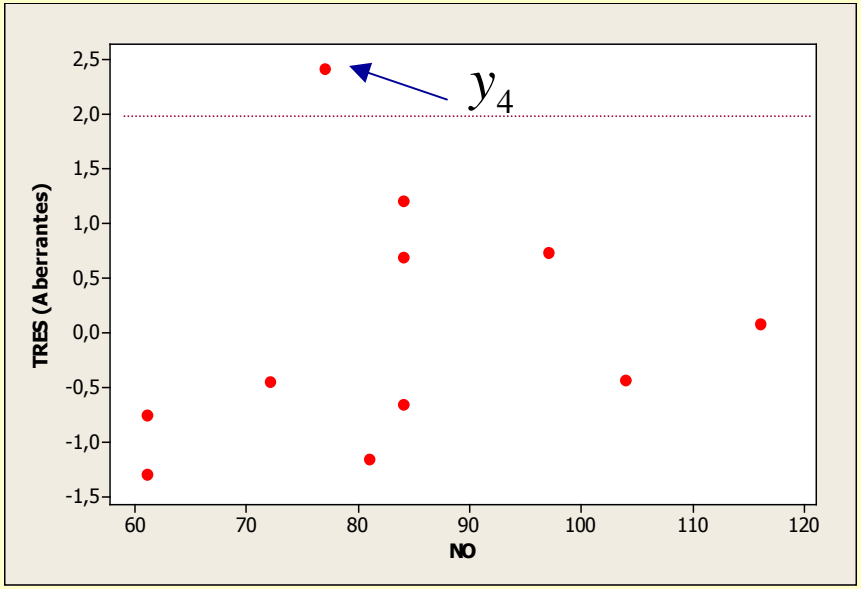
## Dados de Poluentes




$y_2$

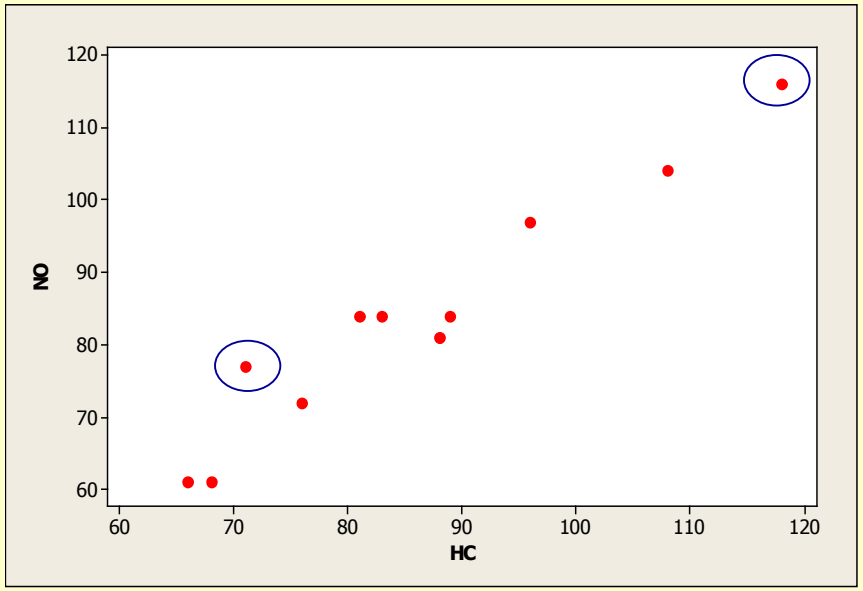



$F_{2,8}(0.5) = 0.757$




$y_4$

## Dados de popluentes ambientais




| Day | Nitrogen Oxides | Hydrocarbons |
|---|---|---|
| 1 | 104 | 108 |
| 2 | 116 | 118 |
| 3 | 84 | 89 |
| 4 | 77 | 71 |
| 5 | 61 | 66 |
| 6 | 84 | 83 |
| 7 | 81 | 88 |
| 8 | 72 | 76 |
| 9 | 61 | 68 |
| 10 | 97 | 96 |
| 11 | 84 | 81 |

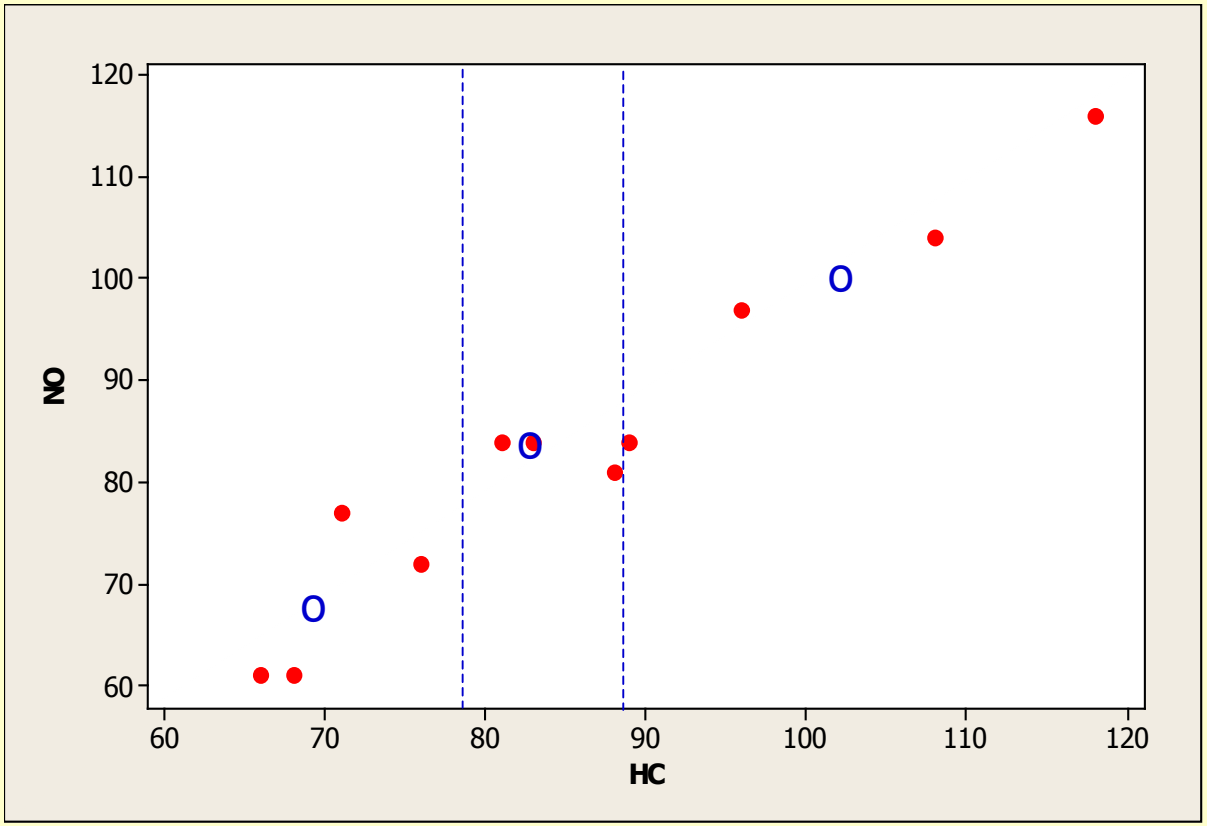
Alavanca

Aberrante

# Análise Exploratória

## Regressão Linear Simples Robusta

Dados de Poluentes



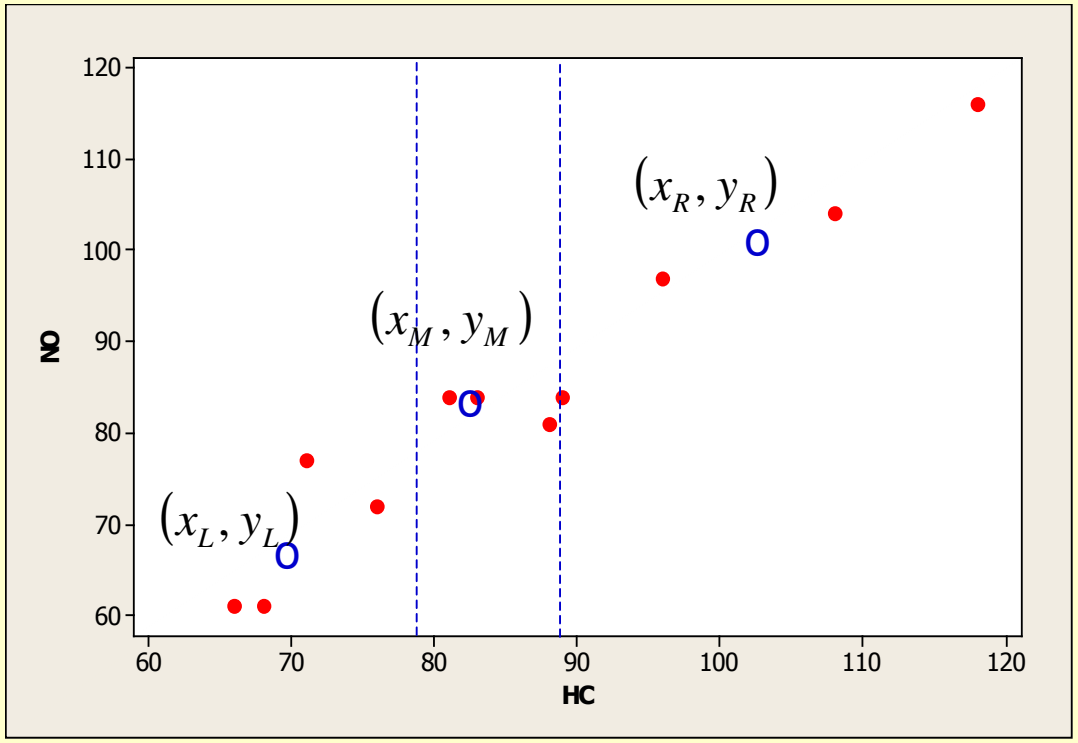

Método dos 3 Grupos

**Construção de Linhas Resistentes**

- Dividir os n pontos em Três grupos (por x)
- Calcular pontos “medianos” (em Y) dentro de cada grupo

# Análise Exploratória - Linha Resistente

## Dados de Poluentes




E.M.Q:

$$\hat{\beta}_1 = \hat{\rho}\frac{s_Y}{s_X} \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

## Pontos Medianos

$$(x_R, y_R) = (69.5; 66.5)$$
$$(x_M, y_M) = (83; 84)$$
$$(x_L, y_L) = (102; 100.5)$$

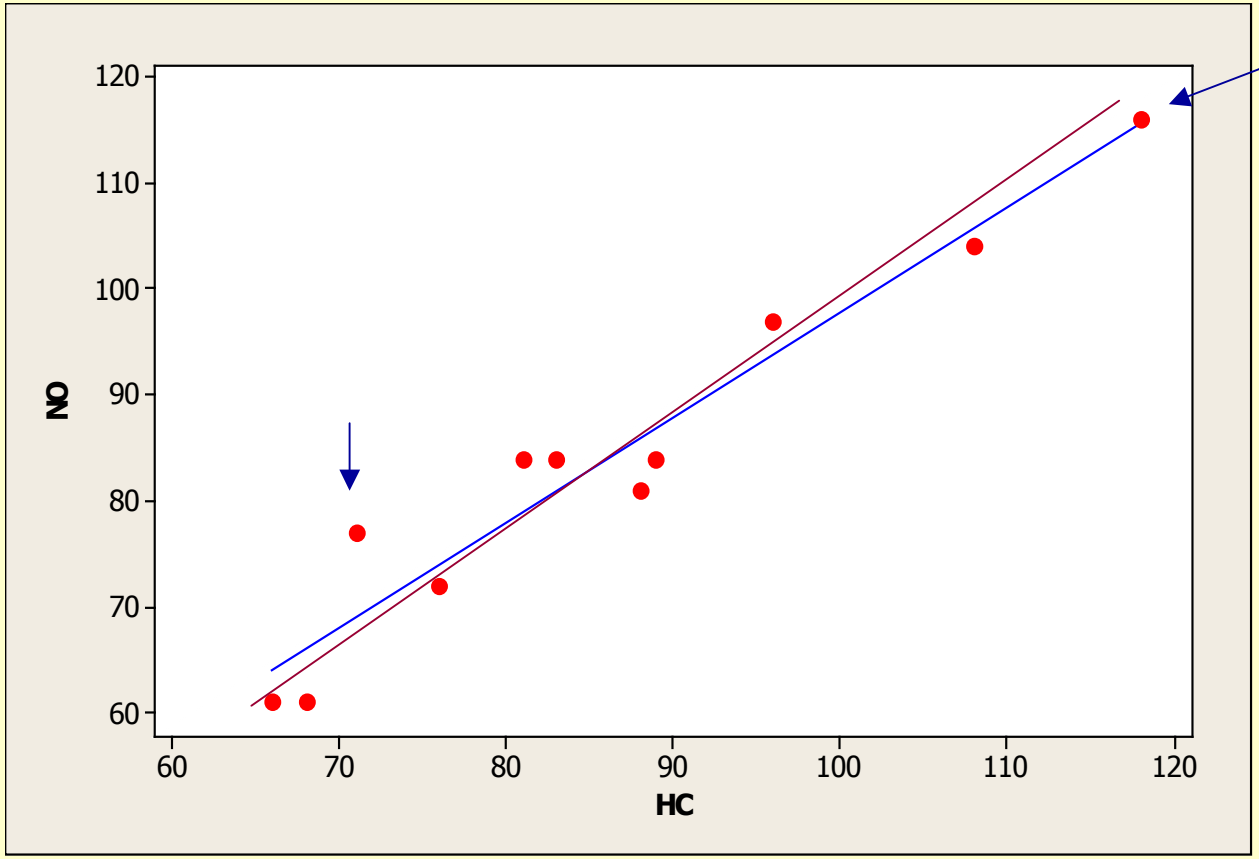
$$\hat{y}_R = a^0 + b^0 x \qquad \Rightarrow \quad b^0 = 1.046 \qquad a^0 = -5.06$$

$$\Rightarrow b^0 = \frac{y_R - y_L}{x_R - x_L}$$

$$\Rightarrow a^0 = \frac{1}{3}\left[(y_L - b^0 x_L) + (y_M - b^0 x_M) + (y_R - b^0 x_R)\right]$$

$a_L$ $a_M$ $a_R$

# Modelos de Regressão




— **Reta de MQ:** $\hat{y} = -1.546 + 0.9936\,x$

— **Linha Resistente (Método dos 3 Grupos):** $\hat{y}_R = -5.06 + 1.046\,x$

**Comparar os resíduos. O que é esperado?**

# Análise Exploratória - Linha Resistente

# Método dos 3 Grupos

Pontos Medianos $\quad \hat{y} = a + bx$

$(x_R, y_R)$ $\quad \Rightarrow b^0 = \dfrac{y_R - y_L}{x_R - x_L}$

$(x_M, y_M)$

$(x_L, y_L)$ $\quad \Rightarrow a^0 = \dfrac{1}{3}\left[(y_L - b^0 x_L) + (y_M - b^0 x_M) + (y_R - b^0 x_R)\right]$

Interpretação de a $\Rightarrow$ Ajuste em torno de um valor central $x = \tilde{x}$

$$\hat{y} = \tilde{a}^0 + b^0(x - \tilde{x})$$

$$\Rightarrow b^0 = \frac{y_R - y_L}{x_R - x_L}$$

$$\Rightarrow \tilde{a}^0 = \frac{1}{3}\left[[y_L - b^0(x_L - \tilde{x})] + [y_M - b^0(x_M - \tilde{x}) + [y_R - b^0(x_R - \tilde{x})]\right]$$

# Análise Exploratória

## Melhorando o Ajuste da Linha Resistente

**<u>Propriedade dos Resíduos</u>: é esperado que os resíduos não contenham nenhuma informação que possa ser sumarizada por meio de uma função linear dos valores X**

$$\hat{\varepsilon}_j = y_j - \left(a^0 + b^0\ x_j\right)$$

$$\Rightarrow \quad \text{Considere os pontos: } \left(x_j, \hat{\varepsilon}_j\right) \quad j = 1,\ldots,n$$

$$\text{Obtenha a Linha de Resistência: } \hat{\varepsilon}_j \times x_j \quad \Rightarrow \quad \delta_b^1 \quad \delta_a^1$$

$$\text{Coeficientes Atualizados: } b^1 = b^0 + \delta_b^1 \qquad a^1 = a^0 + \delta_a^1$$

$$\Rightarrow \quad \hat{y}_j^1 = a^1 + b^1 x_j \quad \Rightarrow \quad \hat{\varepsilon}_j^1 = y_j - \hat{y}_j^1$$

$$\Rightarrow \quad \text{Considere os novos pontos: } \left(x_j, \hat{\varepsilon}_j^1\right) \quad j = 1,\ldots,n$$

$$b^2 = b^1 + \delta_b^2 \qquad a^2 = a^1 + \delta_a^2 \qquad \ldots \qquad \delta_b^I \to 0$$

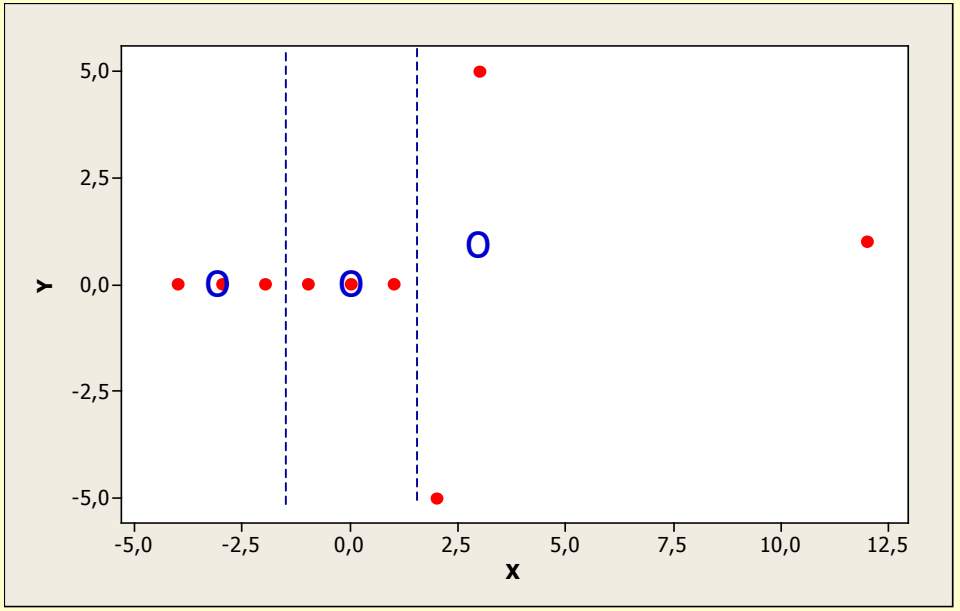
# Análise Exploratória - Linha Resistente

Dados “Patológicos”

(Hoaglin et al., 1983)

| X | Y |
|---|---|
| -4 | 0 |
| -3 | 0 |
| -2 | 0 |
| -1 | 0 |
| 0 | 0 |
| 1 | 0 |
| 2 | -5 |
| 3 | 5 |
| 12 | 1 |

⇒




| Iteração | Linha Ajustada |
|---|---|
| 1 | 0.167X+0.333 |
| 2 | -0.083X-0.167 |
| 3 | 0.0292X+0.583 |
| 4 | -0.217X-0.542 |
| 5 | 0.573X+1.15 |
| 6 | -0.693X-1.39 |
| 7 | 0.833X+1.67 |
| 8 | -0.694X-1.39 |
| 9 | 0.833X+1.67 |
| 10 | -0.694X-1.39 |

Em algumas situações extremas pode ocorrer problemas com o processo de convergência ⇒ prposta de correções no processo iterativo nas situações com pontos de alavanca

# Suavização

Considere $(x_i,y_i)$, i=1,2,...,n

Objetivo: visualizar alguma "tendência" nos dados.

O primeiro passo é fazer o diagrama de dispersão das variaveis X e Y.

Nas técnicas de suavização, substituiremos o valor y do par(x,y) por um valor suavizado.

## Procedimentos:

- Médias móveis
- Medianas móveis
- Lowess

## Caso ideal:

temos vários valores de Y, para um dado valor de X. O valor suavizado poderia ser, por exemplo, a média condicional de Y, dado esse valor de X.

# Suavização – Médias Móveis

$$\hat{y}_t = \sum_{j=-m}^{m} c_j \, y_{t+j}, t = m+1, \cdots, n-m$$

com:

$$\sum_{j=-m}^{m} c_j = 1.$$

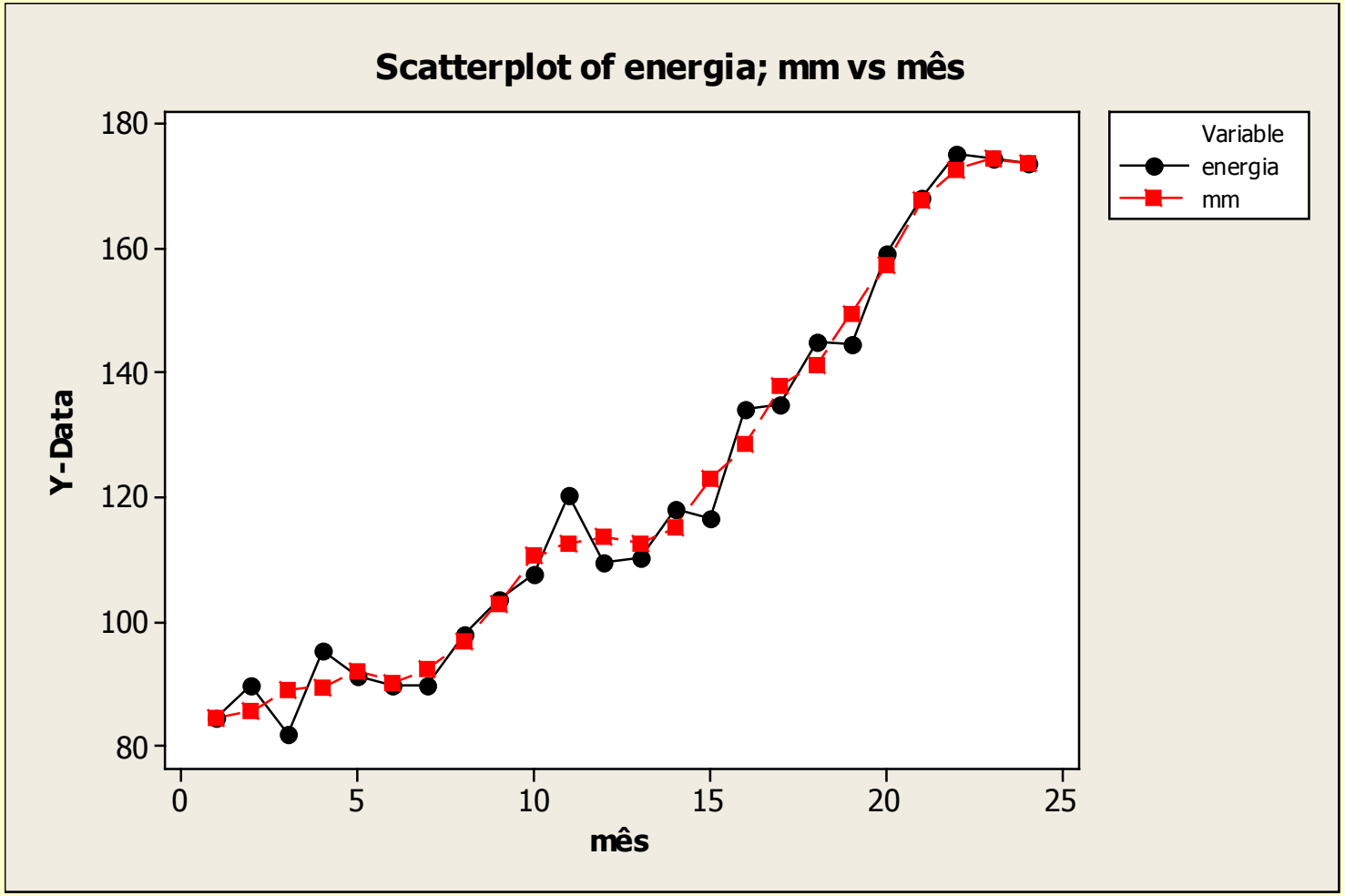
Caso mais simples: $c_j = 1/(2m+1)$

Técnica: $(x_t, y_t)$ é substituido por $(x_t, \hat{y}_t)$

Obs: perdemos m pares no inicio e m pares no final.

Consumo de Energia Elétrica no Espírito Santo, 01/1977-12/1978

| mês | energia | M.M.(m=1) | Med.M(m=1) | mês | energia | M.M.(m=1) | Med.M(m=1) |
|---|---|---|---|---|---|---|---|
| 1 | 84,60 | | | 13 | 110,30 | 112,67 | 110,30 |
| 2 | 89,90 | 85,47 | 84,60 | 14 | 118,10 | 114,97 | 116,50 |
| 3 | 81,90 | 89,07 | 89,90 | 15 | 116,50 | 122,93 | 118,10 |
| 4 | 95,40 | 89,50 | 91,20 | 16 | 134,20 | 128,47 | 134,20 |
| 5 | 91,20 | 92,13 | 91,20 | 17 | 134,70 | 137,90 | 134,70 |
| 6 | 89,80 | 90,23 | 89,80 | 18 | 144,80 | 141,30 | 144,40 |
| 7 | 89,70 | 92,47 | 89,80 | 19 | 144,40 | 149,47 | 144,80 |
| 8 | 97,90 | 97,00 | 97,90 | 20 | 159,20 | 157,27 | 159,20 |
| 9 | 103,40 | 102,97 | 103,40 | 21 | 168,20 | 167,53 | 168,20 |
| 10 | 107,60 | 110,47 | 107,60 | 22 | 175,20 | 172,63 | 174,50 |
| 11 | 120,40 | 112,53 | 109,60 | 23 | 174,50 | 174,47 | 174,50 |
| 12 | 109,60 | 113,43 | 110,30 | 24 | 173,70 | | |

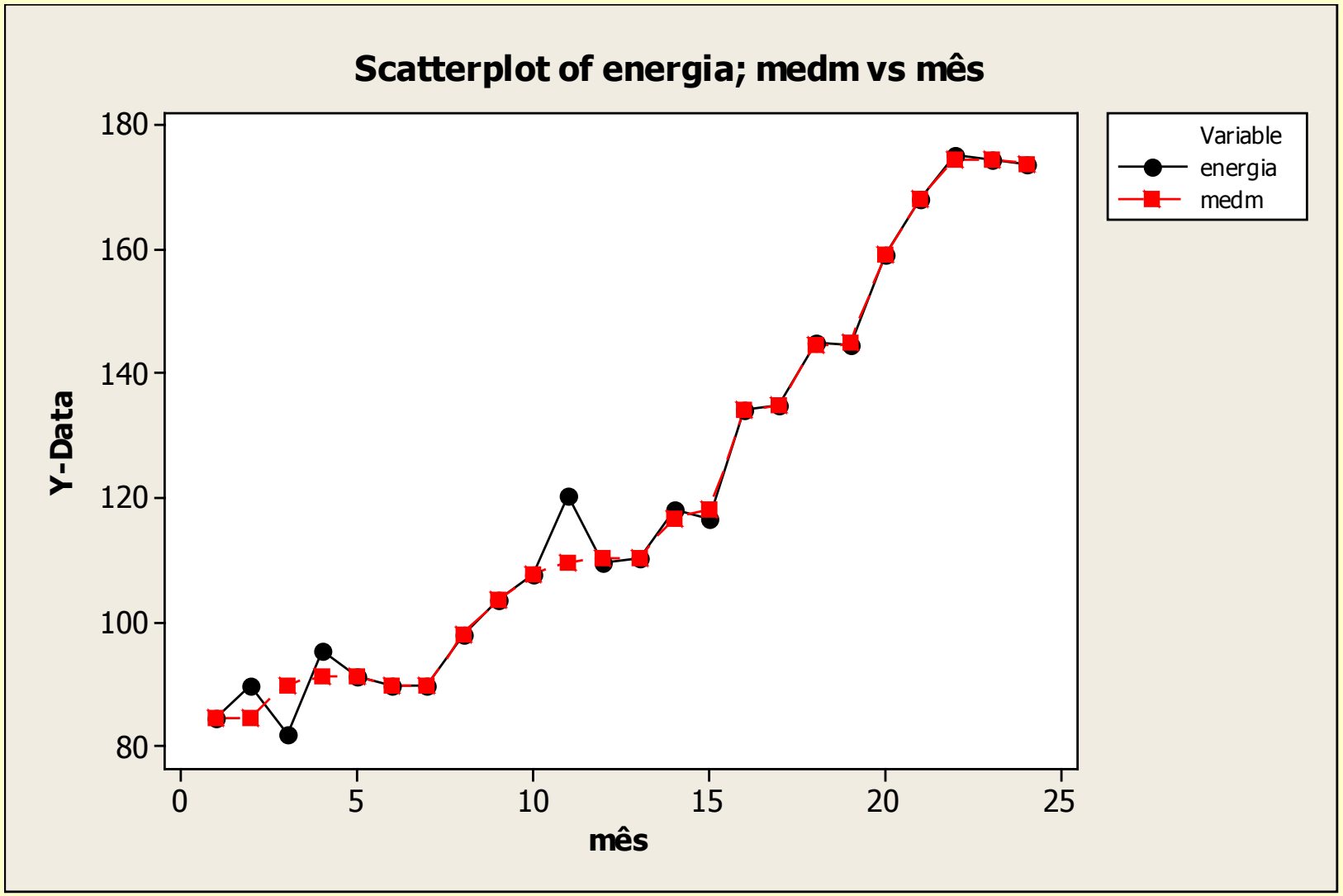

Scatterplot of energia; mm vs mês
Y-Data
180
160
140
120
100
80
0
5
10
15
20
25
mês
Variable
energia
mm

# Suavização: Medianas Móveis

$$\tilde{y}_t = mediana(y_{t-m}, y_{t-m+1}, \cdots, y_{t+m})$$

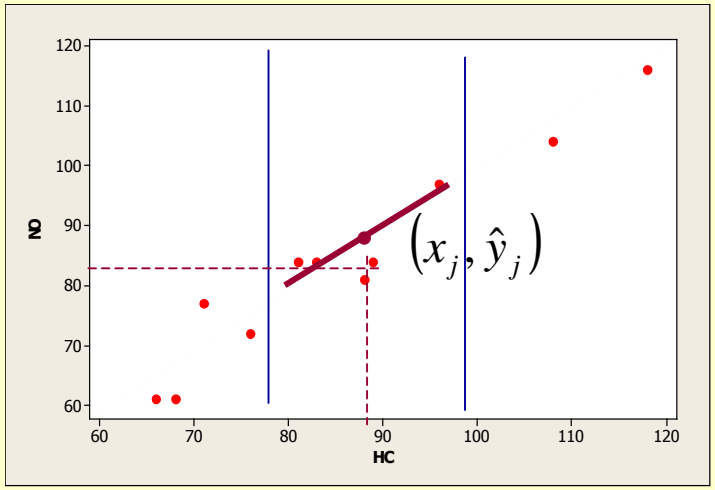
Técnica: $(x_t, y_t)$ é substituído por $(x_t, \tilde{y}_t)$

VANTAGEM: MEDIANA É UMA MEDIDA RESISTENTE A VALORES DISCREPANTES


Scatterplot of energia; medm vs mês
Y-Data
180
160
140
120
100
80
0
5
10
15
20
25
mês
Variable
energia
medm

# Suavização - Lowess

## Lowess: Locally weighted regression scatter plot smoothing




Para obter $(x_j, \hat{y}_j)$ :

- Abrir uma faixa vertical centrada em ($x_j$,$y_j$), contendo *q=[fn]* pontos (0<f<1). Quando maior o valor de f, mais suave será o ajustamento. S-Plus

$$1/3 \le f \le 2/3$$

- Definir pesos para os pontos vizinhos de $(x_j, y_j)$

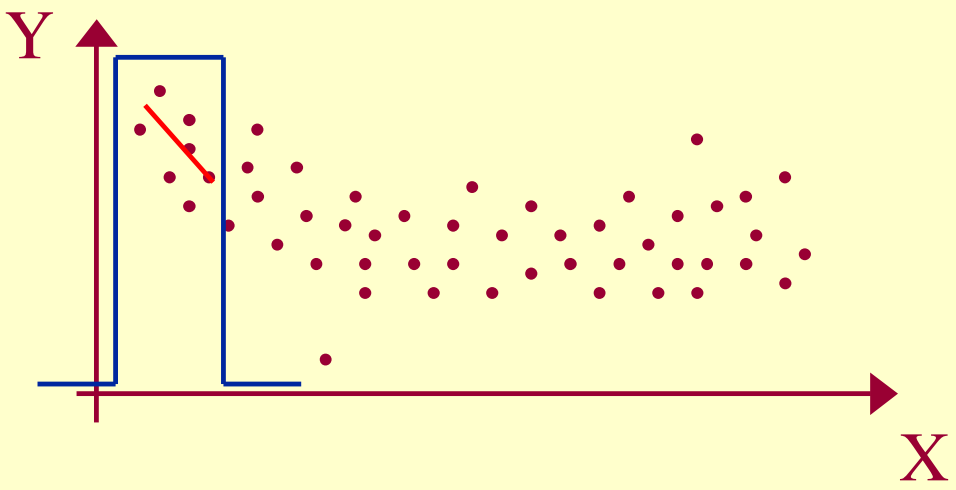
$$h(u) = \begin{cases} (1-|u|^3)^3 & \text{se } |u|<1 \\ 0 & \text{cc} \end{cases} \Rightarrow \text{o peso atribuído a } (x_k, y_k) \text{ é } h(x_k) = h\left(\frac{x_j - x_k}{d_j}\right)$$

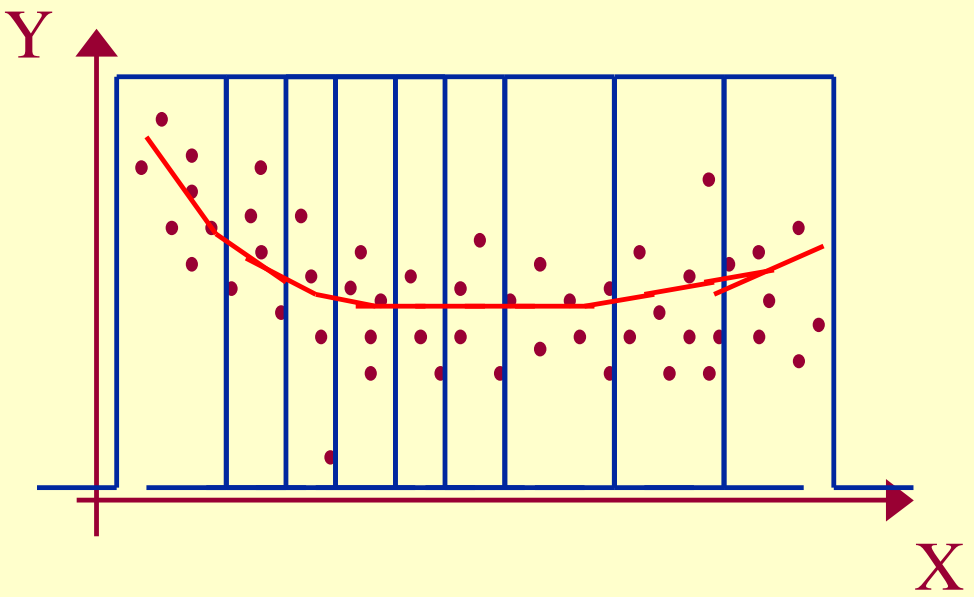
distância *ao vizinho mais afastado*

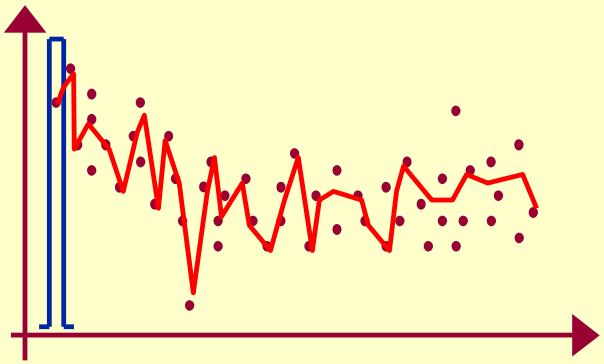
- Ajustamos uma reta aos *q* pontos (M.Q.P.)

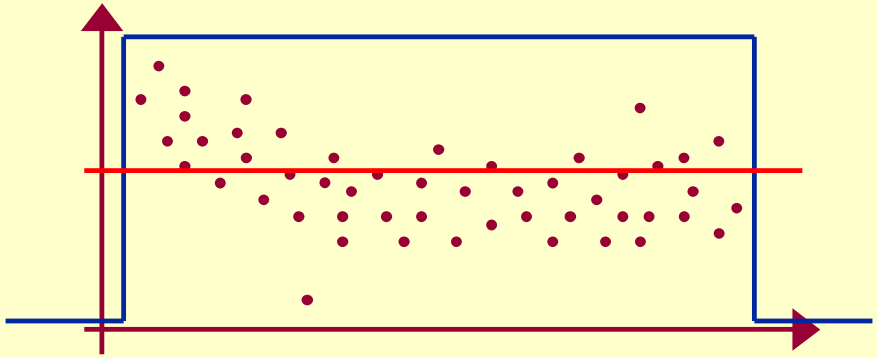
$$\hat{y}_j = \hat{\alpha} + \hat{\beta}\, x_j; \qquad \sum_{k=1}^{n} h(x_k)\,(y_k - \alpha - \beta\, x_k)^2$$

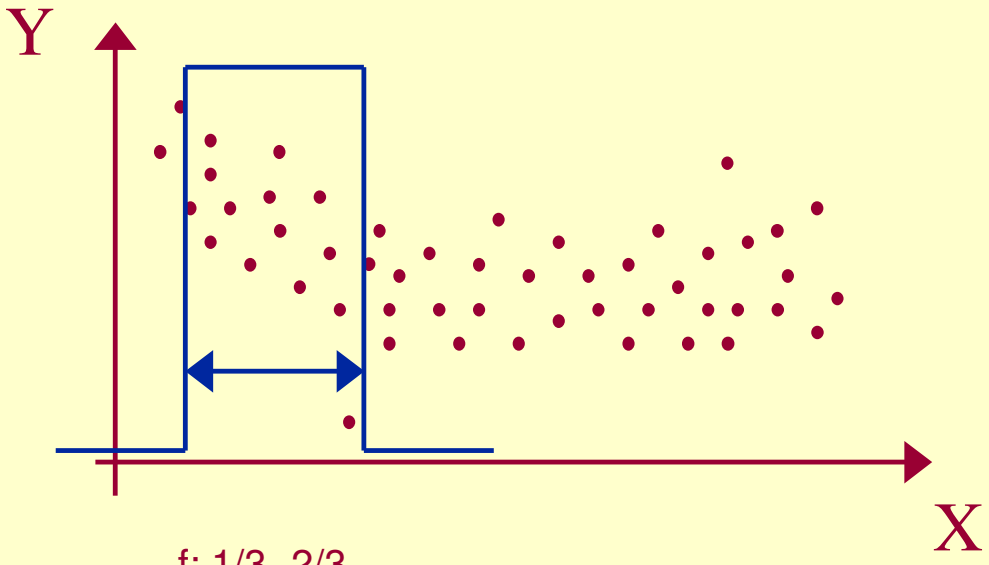
↑ resíduo:↓ peso


Y
X


Y
X

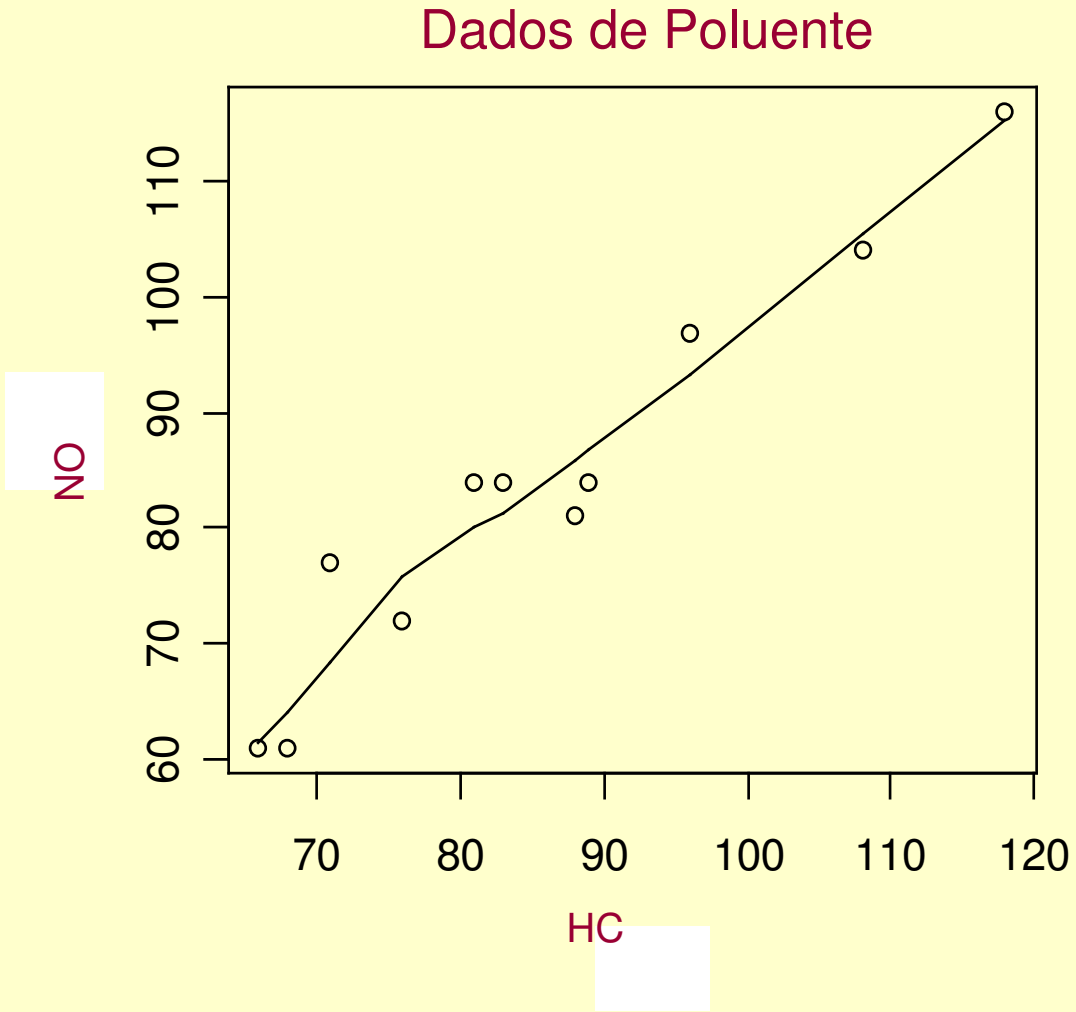







f: 1/3 -2/3

# Suavização - Lowess




> plot (x,y)

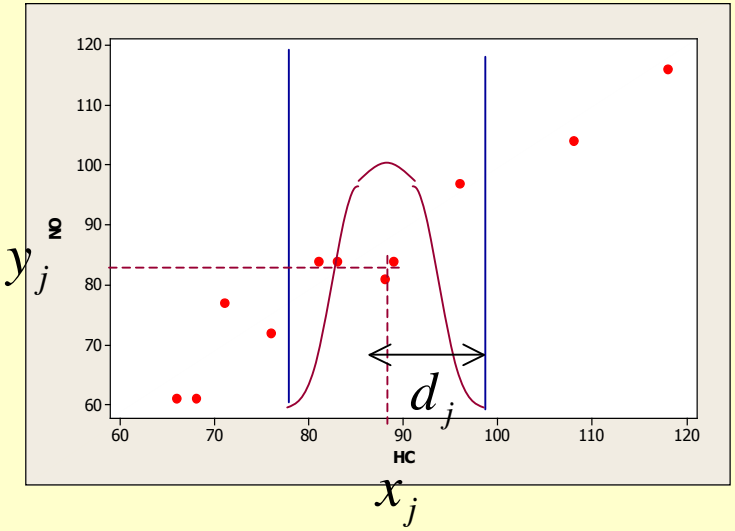
> lines(lowess(x,y,f=2/3))

# Suavização – Lowess Robusto

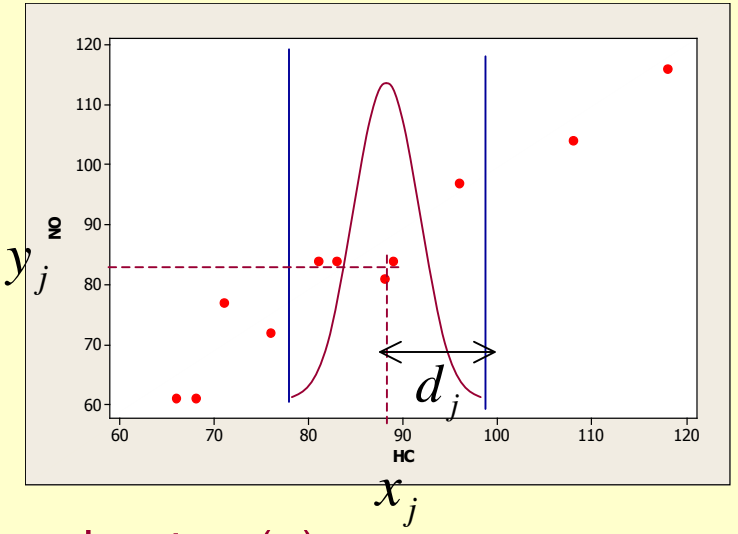
**Na presença de valores atípicos ⇒ Lowess Robusto (Iterativo)**

Atribuição de pesos robustos às observações $(x_j, y_j)$

h(u): Tri-cúbica




g(u): Bi-quadrática




- Lowess (h) ⇒ Lowess robusto: pesos robustos (g)

$$\hat{\varepsilon}_j = y_j - \hat{y}_j \qquad Md: \text{mediana dos valores} |\hat{\varepsilon}_j|$$

$$g(u) = \begin{cases} (1-|u|^2)^2 & \text{se } |u|<1 \\ 0 & \text{cc} \end{cases}$$

⇒ o peso atribuído a $(x_k, y_k)$ é $g(x_k) = g\left(\frac{\hat{\varepsilon}_k}{6Md}\right)$

$$\hat{y}_j = \hat{\alpha} + \hat{\beta}\, x_j; \qquad \sum_{k=1}^{n} g\, h(x_k)\,(y_k - \alpha - \beta\, x_k)^2$$

## Procedimentos:

1. Calculamos os resíduos: $\hat{\varepsilon}_j = y_j - \hat{y}_j$

Um gráfico de dispersão $\hat{\varepsilon}_j \times x_j$ mostrará os valores discrepantes.

2. Definimos novos pesos $g(x_k)$. Assim,

a) Se $\hat{\varepsilon}_j \prec\prec 6m \Rightarrow g(x_j) \approx 1$

b) Se $\hat{\varepsilon}_j \approx 6m \Rightarrow g(x_j) \cong 0$

Para dados normais $6m \approx 4\sigma \rightarrow$ raramente teremos pesos pequenos.

3. Ajustamos uma nova reta aos q pontos, atribuindo a $(x_k,y_k)$ o peso $h_j(x_k)g(x_k)$. Assim, se $(x_k,y_k)$ for discrepante, o resíduo será grande e o peso final será pequeno. O procedimento deve ser repetido duas ou mais vezes.

# Uma v.a. qualitativa e outra quantitativa

Exemplo: sexo e salário

Técnica:

histograma, boxplot, ramo-e-folha, da variável salário para cada sexo e verificar como esta varia segundo os dois atributos: masculino e feminino.