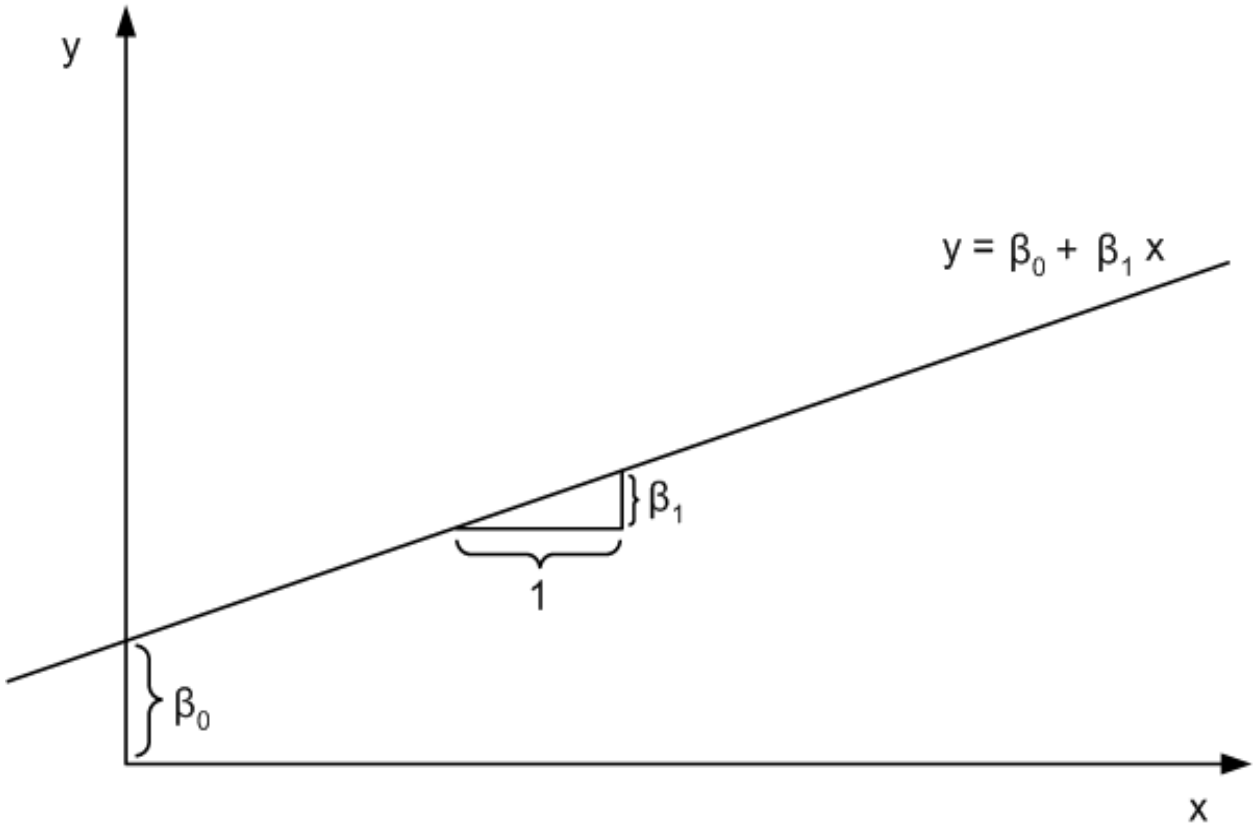
REGRESSÃO



Suposições do Modelo

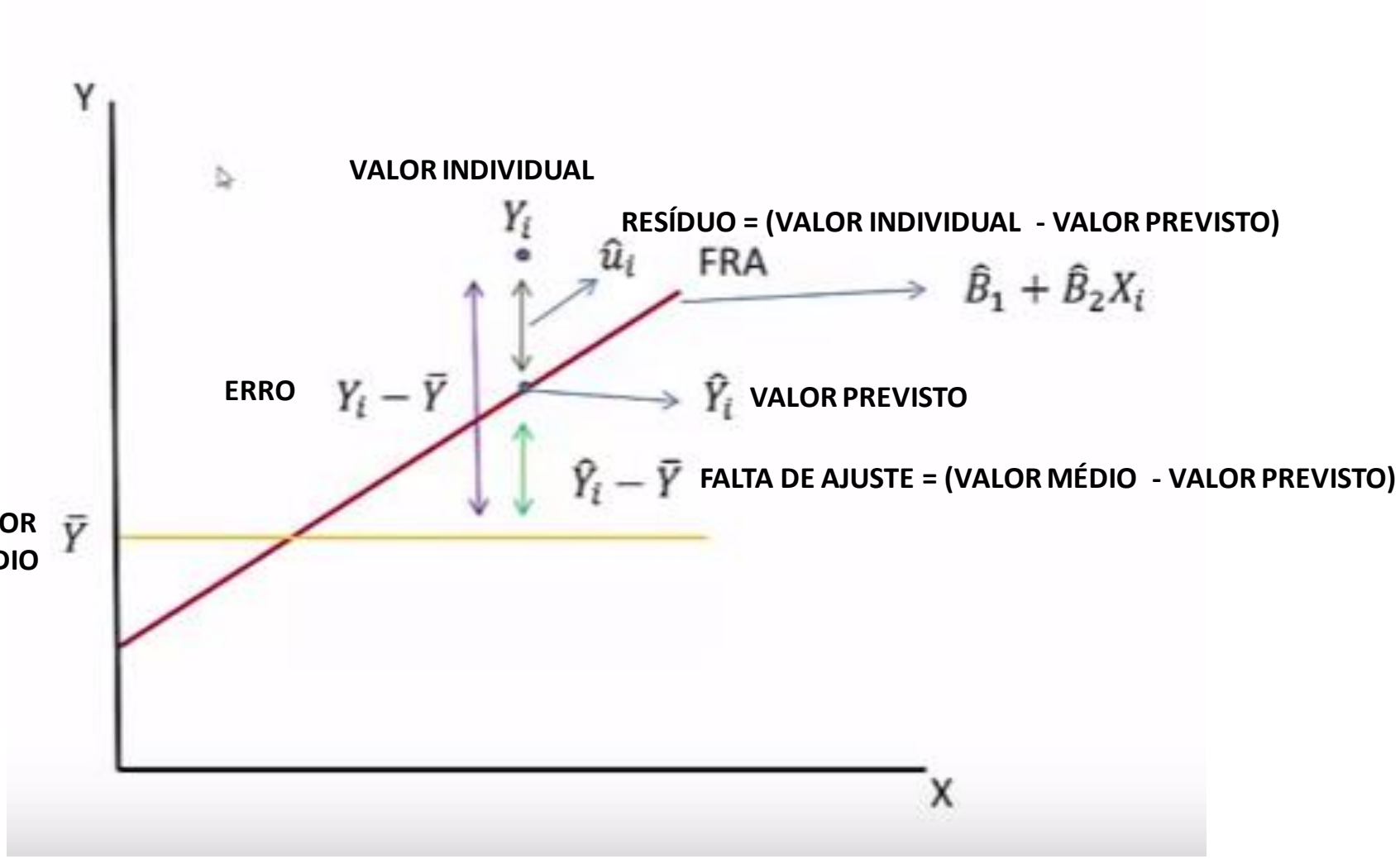
Modelo de Regressão Linear Múltipla (MRLM)

- O erro tem média zero e variância σ^2, desconhecida
- Os erros são não correlacionados
- Os erros têm distribuição normal
- As variáveis regressoras $x_1, x_2, \ldots, x_p$ assumem valores fixos

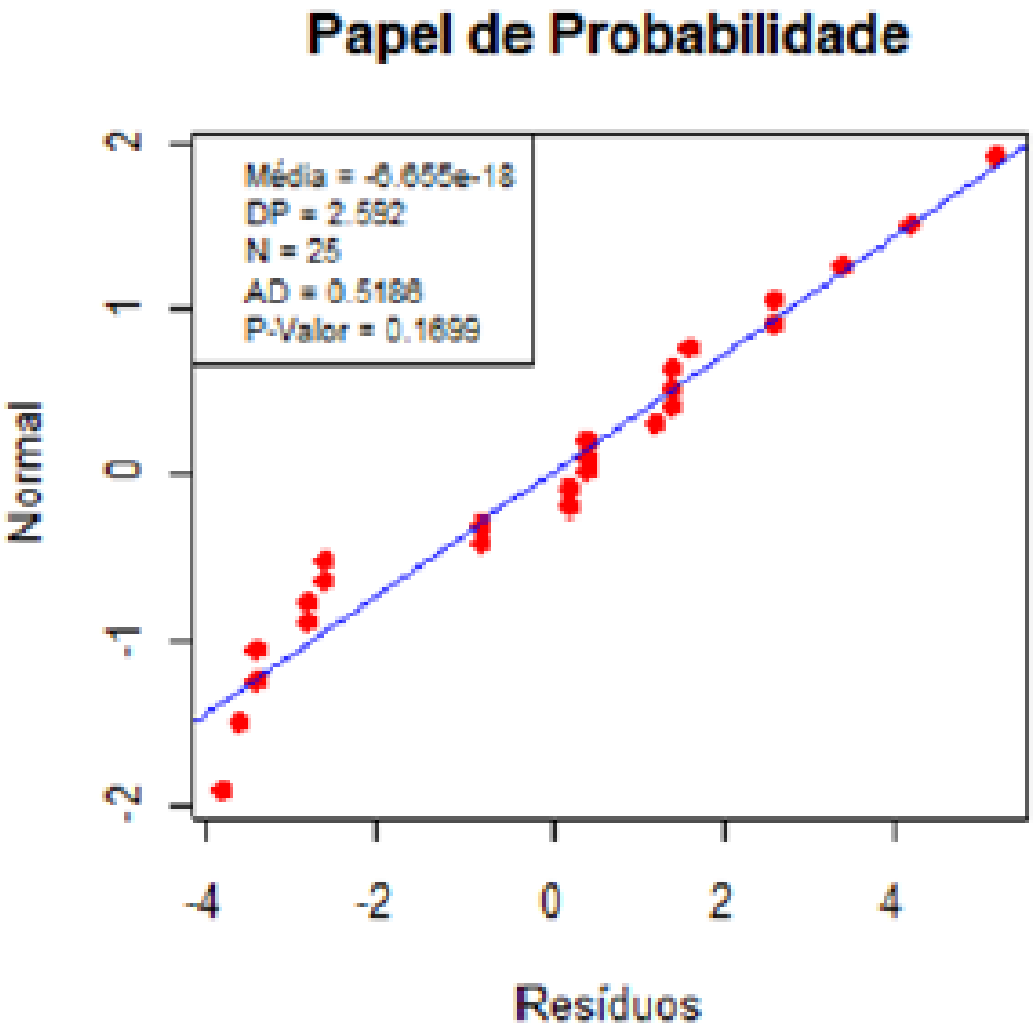
Se as suposições do MRLM se verificam, então a variável Y tem distribuição normal com variância σ^2 e média

$$E(Y|x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

RESÍDUOS FALTA AJUSTE ERRO GRÁFICO EXPLICATIVO



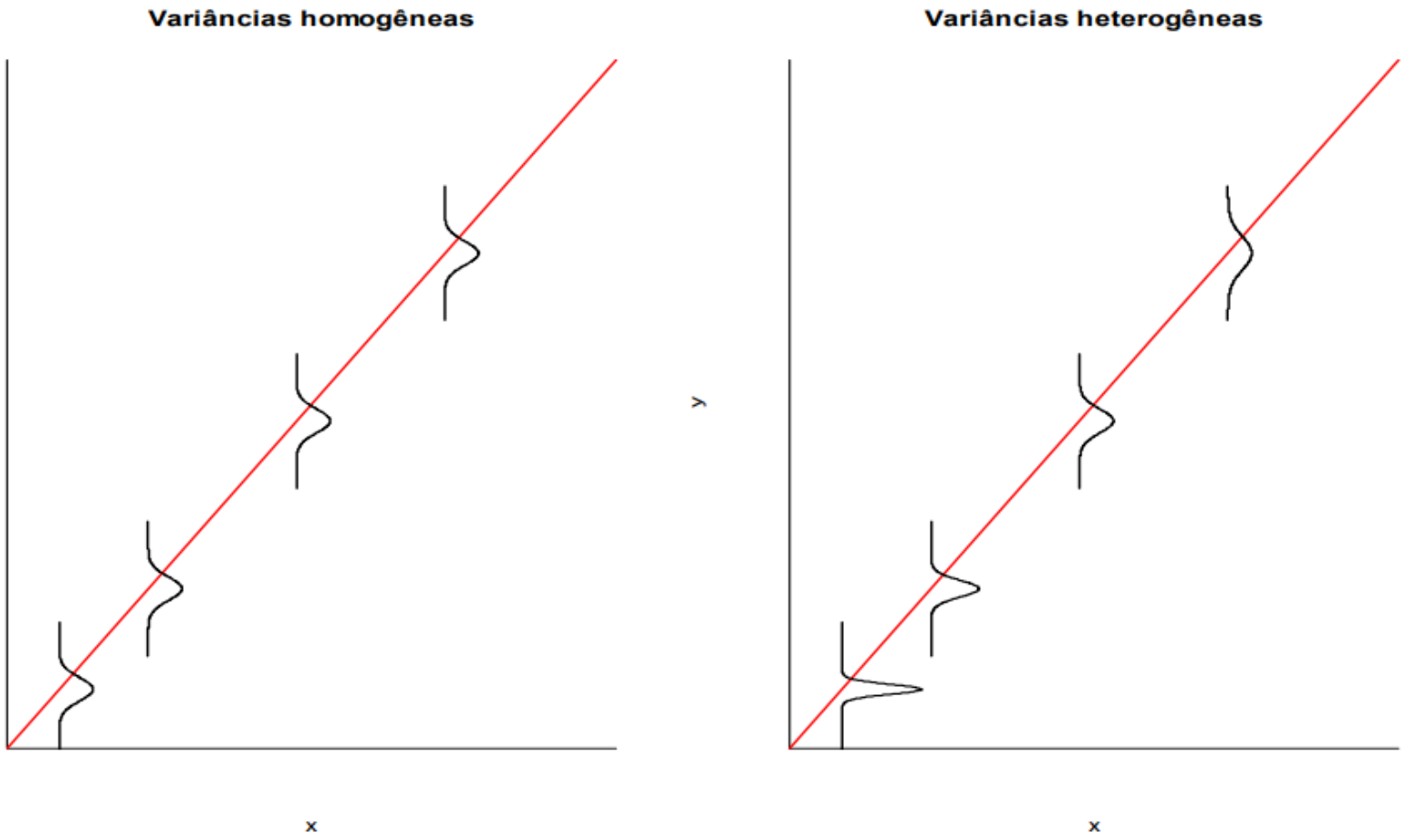
RESÍDUOS



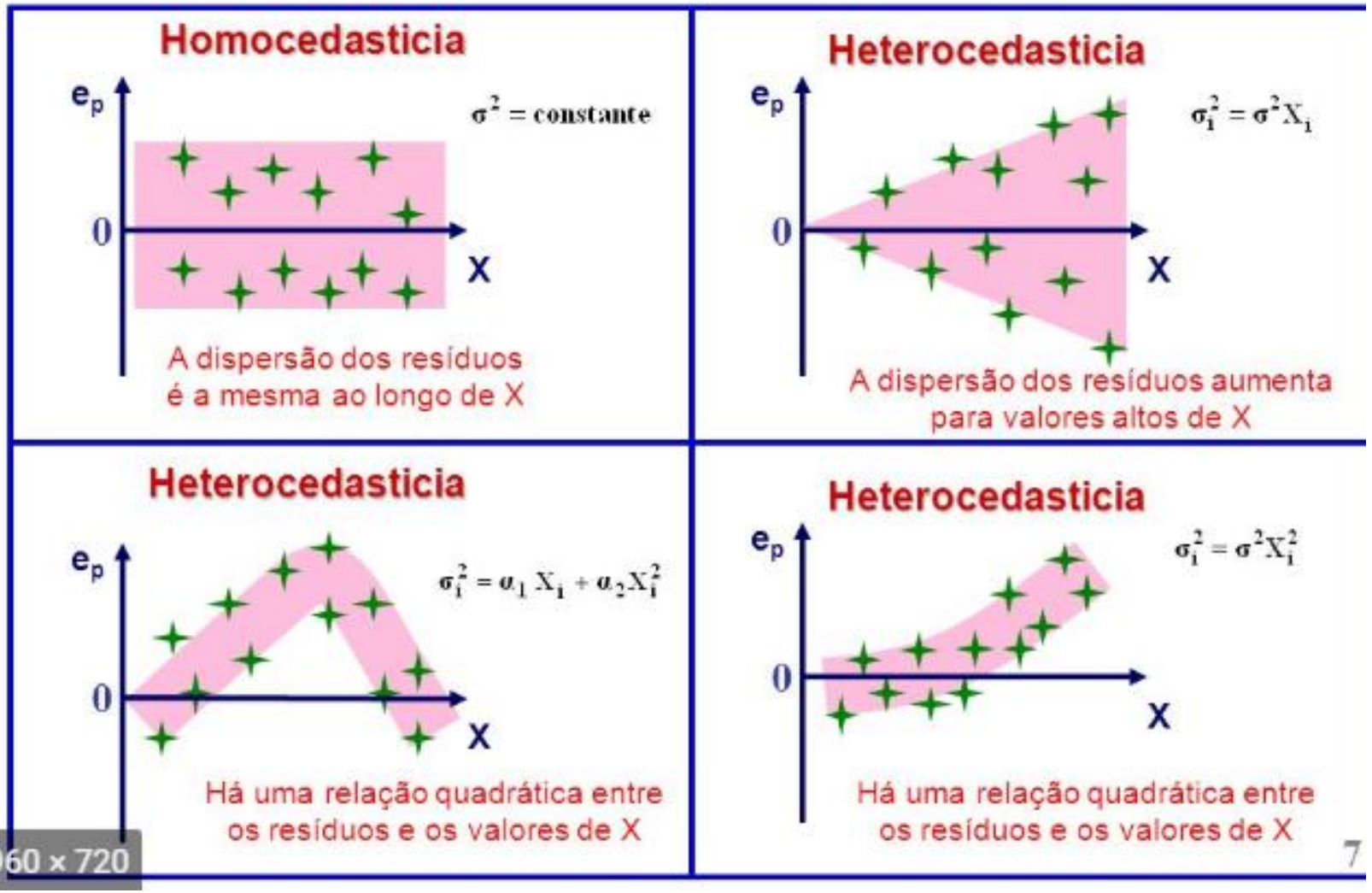
H_0= resíduos seguem normalidade
P > 0,05

H_1 = resíduos não tem normalidade
P < 0,05

HOMOCEDASTICIDADE



HOMOCEDASTICIDADE



960 × 720

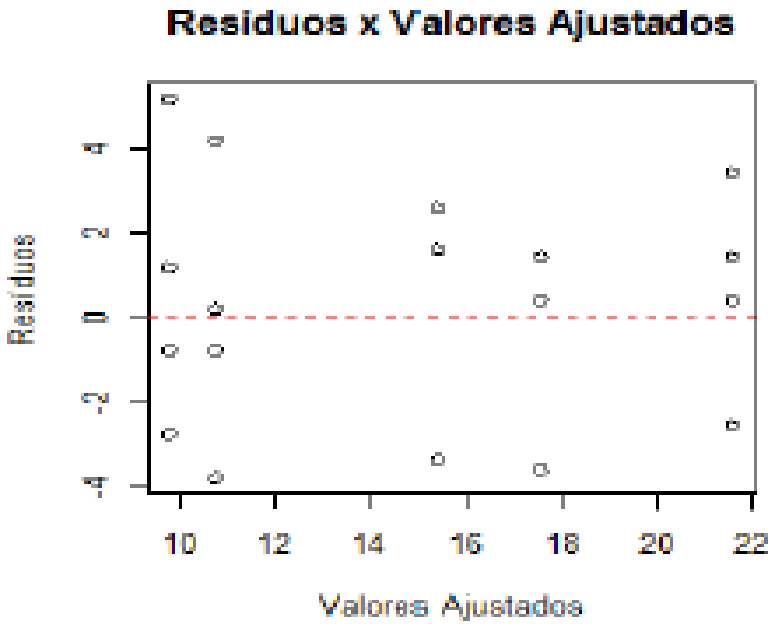
TESTE HOMOCEDASTICIDADE

$$\begin{cases} H_0 : \sigma_1^2 = \sigma_2^2 = ... = \sigma_k^2 \\ H_1 : \text{ pelo menos um dos } \sigma_i^2\text{'s diferente}, \quad i = 1, \ldots, k. \end{cases}$$

Ao longo desta seção, vamos apresentar os diversos testes propostos na literatura.

Teste de Breusch-Pagan – AMOSTRAS PEQUENAS

Teste de Goldfeld-Quandt – AMOSTRAS GRANDES



Diagnóstico de Independência

Para verificar se os resíduos são independentes, podemos utilizar técnicas gráficas e o testes de Durbin-Watson.

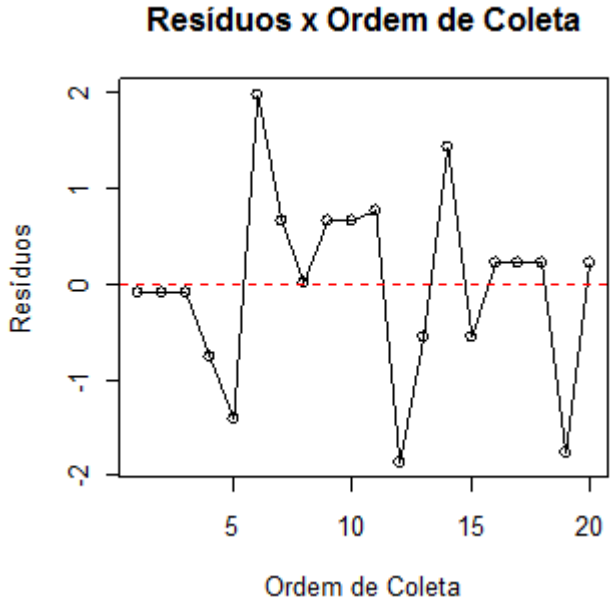
A seguir apresentamos algumas técnicas utilizadas:

Teste de Durbin-Watson → é utilizado para detectar a presença de auto correlação (dependência) nos resíduos de uma análise de regressão.

Testamos a presença de autocorrelação por meio das hipóteses:
p = coeficiente de correlação

$H_0: \rho = 0$ para $p > 0{,}05$

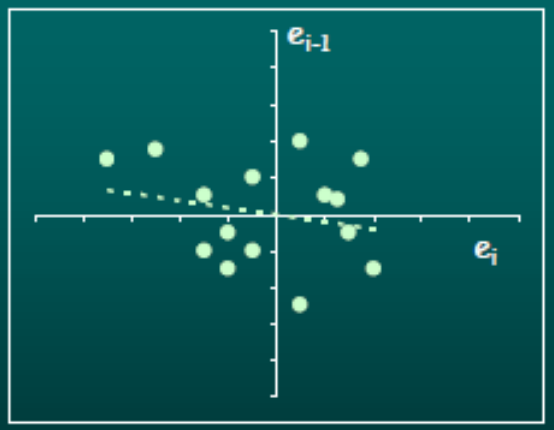
$H_1: \rho \neq 0$ para $p < 0{,}05$



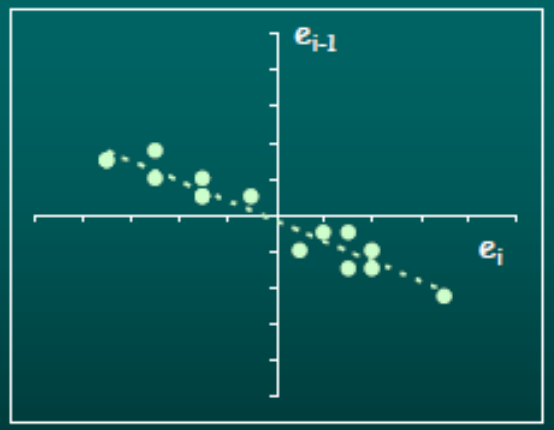
Técnica gráfica → Resíduos vs. Ordem de Coleta

Diagnóstico de Independência

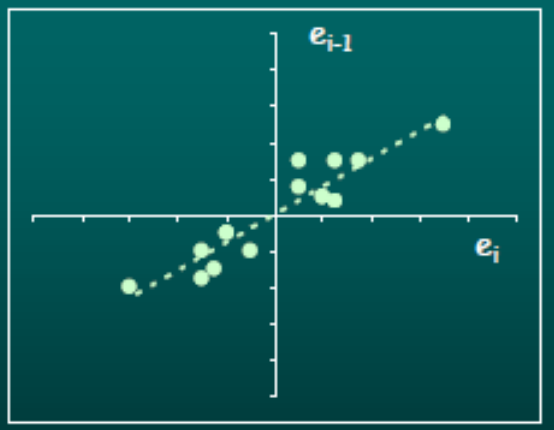
- Teste de Durbin-Watson



Independência



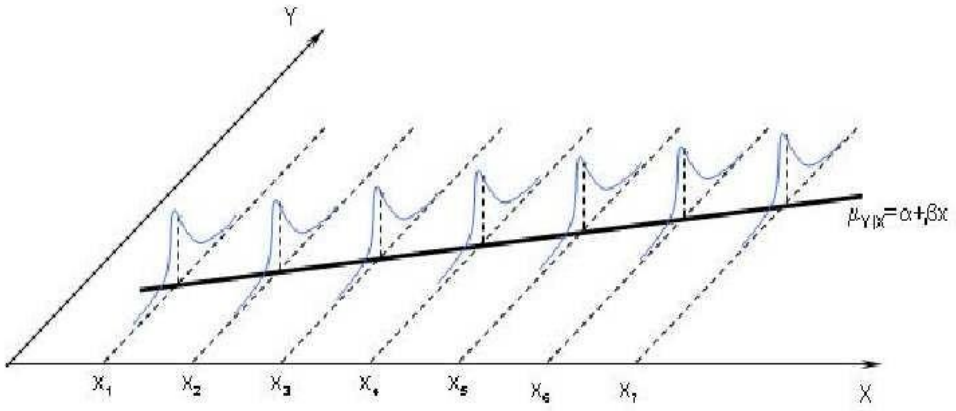
Autocorrelação negativa



Autocorrelação positiva

Teste da Falta de Ajuste (Lack of Fit)

Após o ajuste, é importante verificar se o modelo linear é adequado

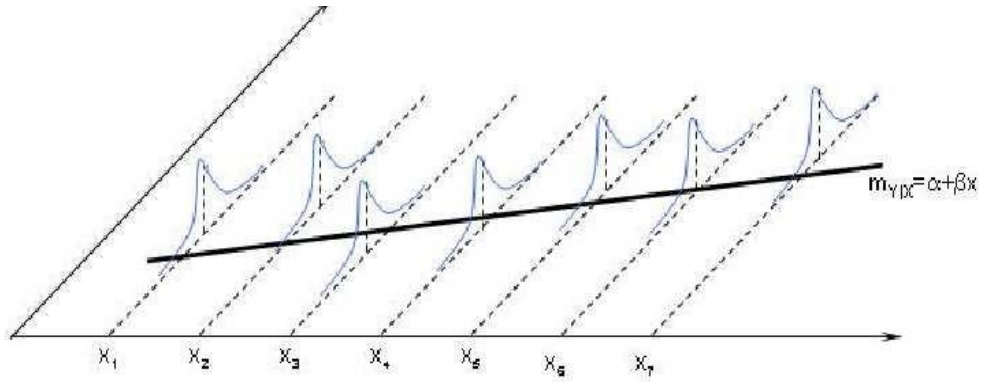


Reta de regressão perfeitamente ajustada sem Falta de Ajuste

$$\begin{cases} H_0 : E(Y_i) = \beta_0 + \beta_1\ x_i \ \text{ modelo linear adequado} \\ H_1 : E(Y_i) \neq \beta_0 + \beta_1\ x_i \ \text{ modelo linear inadequado} \end{cases}$$

p > 0,05

p < 0,05



Reta de regressão com Falta de Ajuste

Diagnóstico de Outliers

Outlier é uma observação extrema, ou seja, é um ponto com comportamento diferente dos demais. Além de diagnosticar heteroscedasticidade, o gráfico de resíduos versus valores ajustados também auxilia na detecção de pontos atípicos.

A detecção de pontos atípicos tem por finalidade identificar:

- Outliers com relação a X
- Outliers com relação a Y
- Observações influentes

A seguir apresentamos algumas técnicas utilizadas:

Outliers em Y

- Resíduos Padronizados
- Resíduos Studentizados

Outliers em X

- Matriz chapéu H
- h_{ii} é o i-ésimo elemento da diagonal principal da matriz H e também é chamado de leverage da observação i.

OUTLIERS EM “Y”
PADRONIZAÇÃO DE RESÍDUOS

MEDE QUANTO O DP ESTÁ AFASTADO

$$z_i = \frac{e_i}{S_e}$$

Resíduo original

Desvio padrão dos resíduos

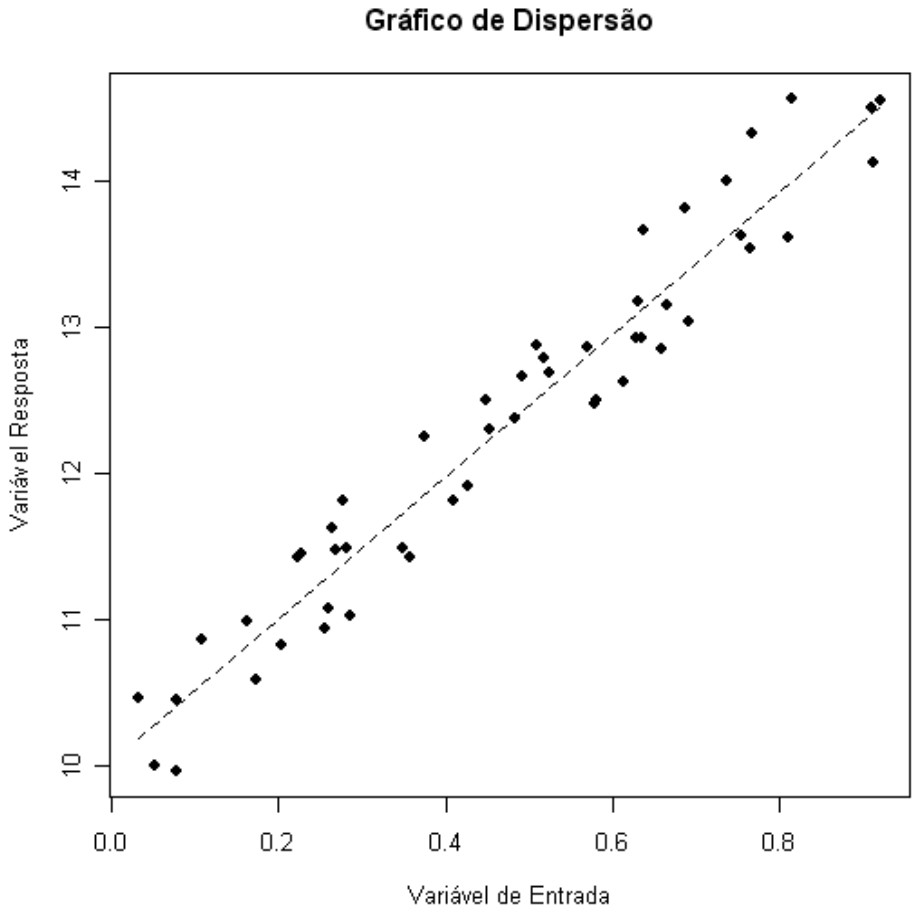
O resíduo padronizado é o quociente entre o resíduo e a estimativa do seu desvio padrão

Esses resíduos têm média zero e variância aproximadamente igual a um. A maioria dos resíduos padronizados deve estar no intervalo $-3 \leq di \leq 3$

$$r_i = \frac{e_i}{\sqrt{QME(1-h_{ii})}},$$

Modelo Estatístico

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + E$$



$H_0: \beta_j = 0$

$p > 0.05$

$H_1: \beta_j \neq 0$

$P < 0,05$

RESÍDUOS STUDENTIZADOS

Traz a mesma informação
dos resíduos padronizados.

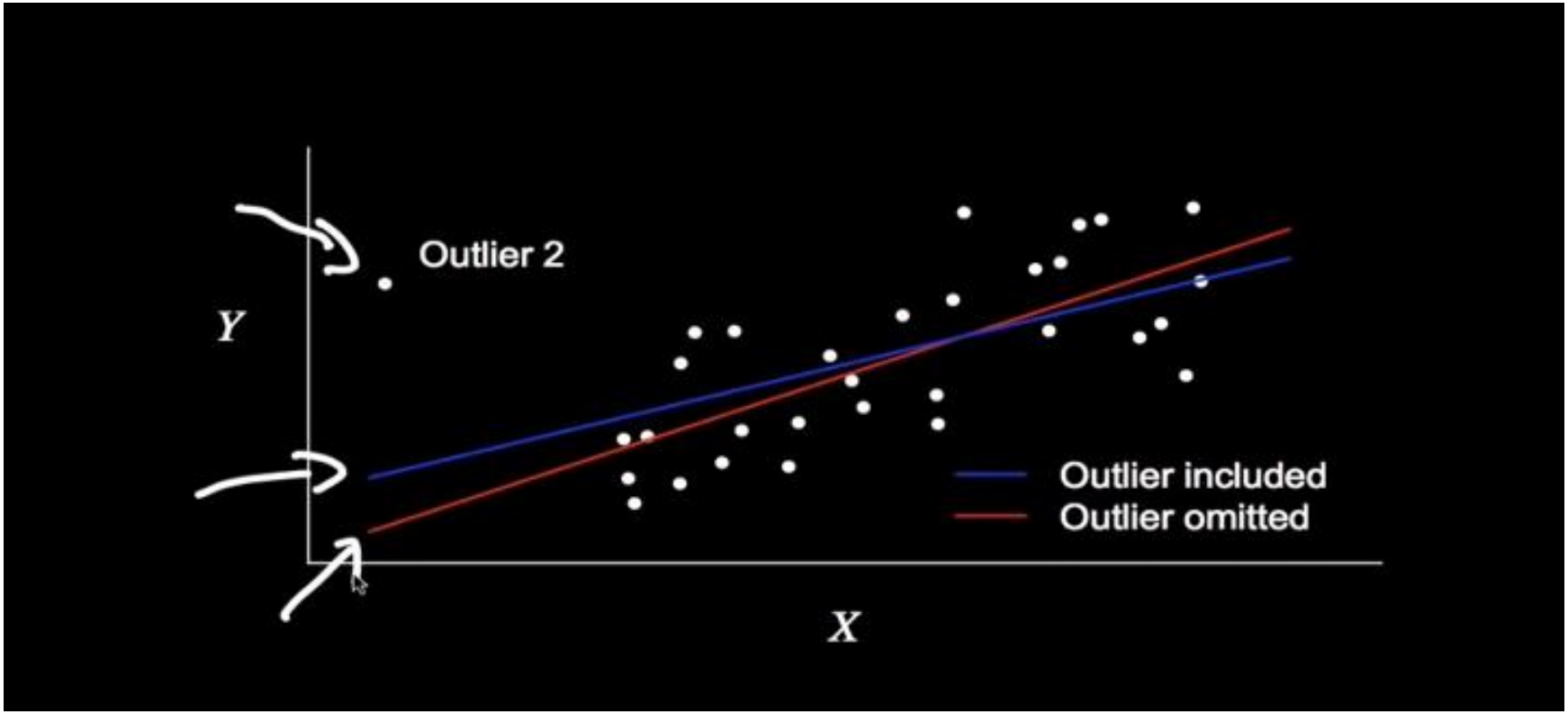
OUTLIERS EM “x”

- Leverage: Leverage (Hi) mede a distância entre o valor x de uma observação e a média dos valores de x para todas as observações em um conjunto de dados. Use para identificar observações com valores de preditores atípicos em comparação aos outros dados.

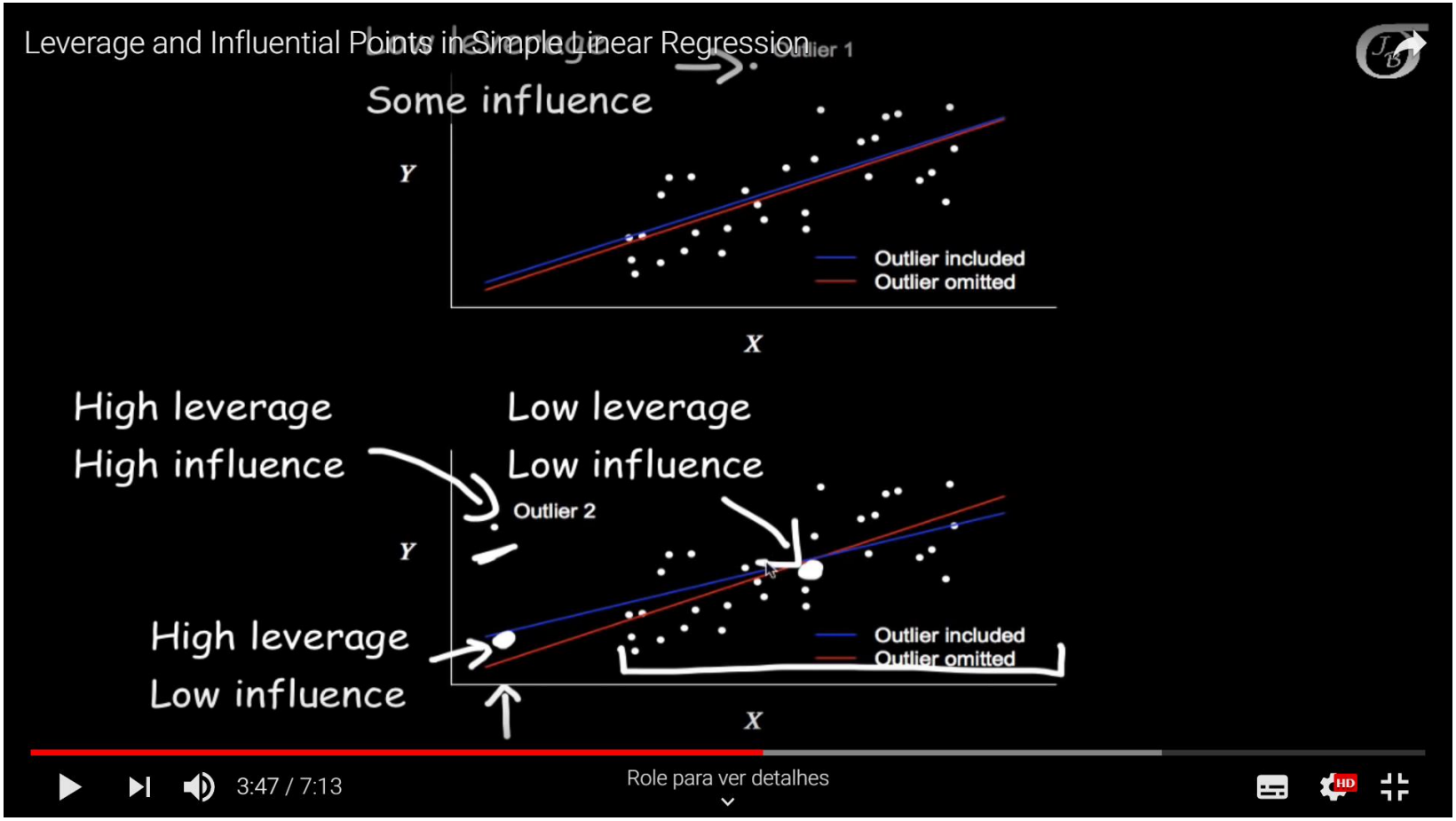
$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum(x - \bar{x})^2}$$

- Limite é dado pela equação: 2 (p+1)/n
- onde p é o número de termos do modelo (incluindo a constante) e n é o número de observações.

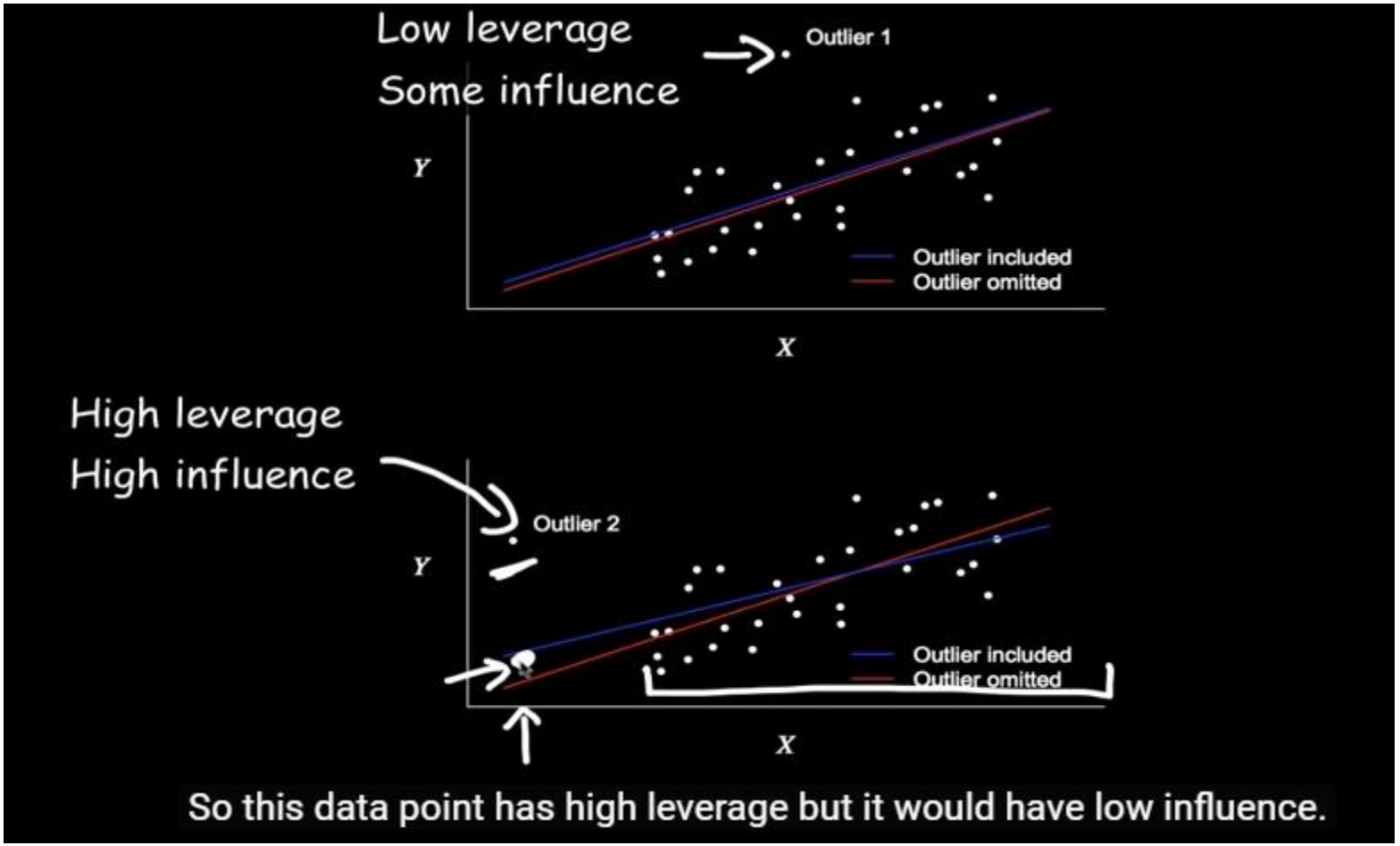
LEVERAGE/PONTOS INFLUENTES



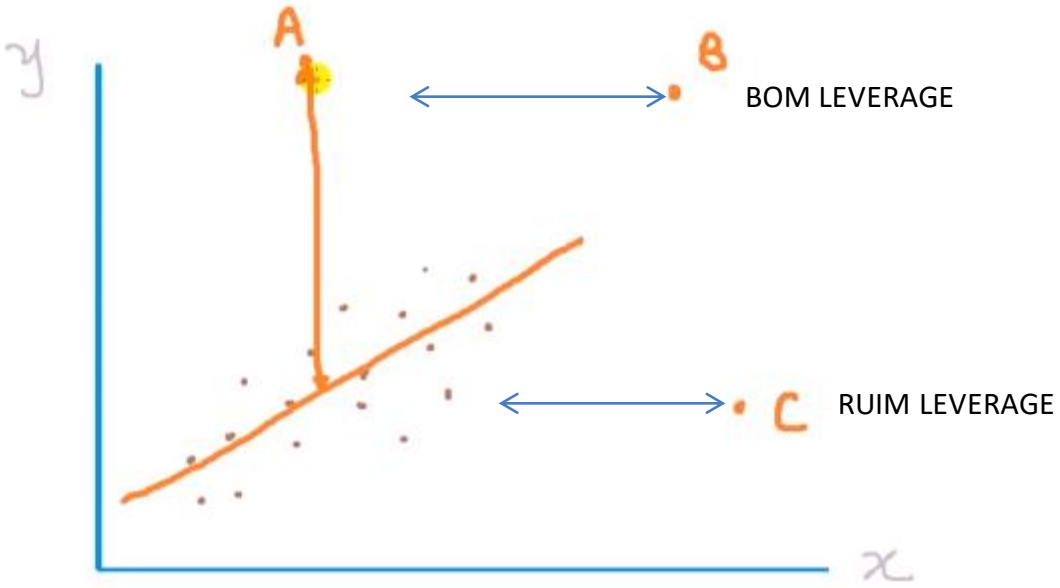
LEVERAGE/PONTOS INFLUENTES



LEVERAGE/PONTOS INFLUENTES



LEVERAGE/PONTOS INFLUENTES



Pontos Influentes

- **Se retirássemos determinados casos, teríamos coeficientes de regressão diferentes???**

- **Objetivo da análise:** determinar se o modelo de regressão é estável para toda a amostra ou se ele pode estar sendo influenciado somente por poucos casos (atípicos).

Distância de COOK

A distância de Cook mede a influência da observação i sobre todos n valores ajustados .

$$D_i = \frac{e_i^2}{(p+1)QME} \frac{h_{ii}}{(1-h_{ii})^2}.$$

Basicamente é como o modelo se comportaria excluindo determinada observação

DFFITS

A medida DFFITS mede a influência que a observação i tem sobre seu próprio valor ajustado. Neste caso, medimos a influência da exclusão da i-ésima observação no seu valor previsto ou ajustado.

$$DFFITS_{(i)} = \frac{\hat{Y}_i - \hat{Y}_{i(i)}}{\sqrt{QME_{(i)}h_{ii}}},$$

A diferença dos valores preditos de Yi com e sem a observação i (se i está entre parênteses, significa que excluímos observação i), expressa em unidades de desvios padrão dos valores preditos de Yi.

DFBETA

A medida DFBETA mede a influência da observação i sobre o coeficiente de Xi. Esta medida de influência é definida por :

$$DFBETA_{j(i)} = \frac{\hat{\beta}_j - \hat{\beta}_{j(i)}}{\sqrt{QME_i c_{jj}}}, \qquad j = 0, 1, ..., p,$$