Insper

Ciência dos Dados

Modelo de regressão linear

Análise de regressão

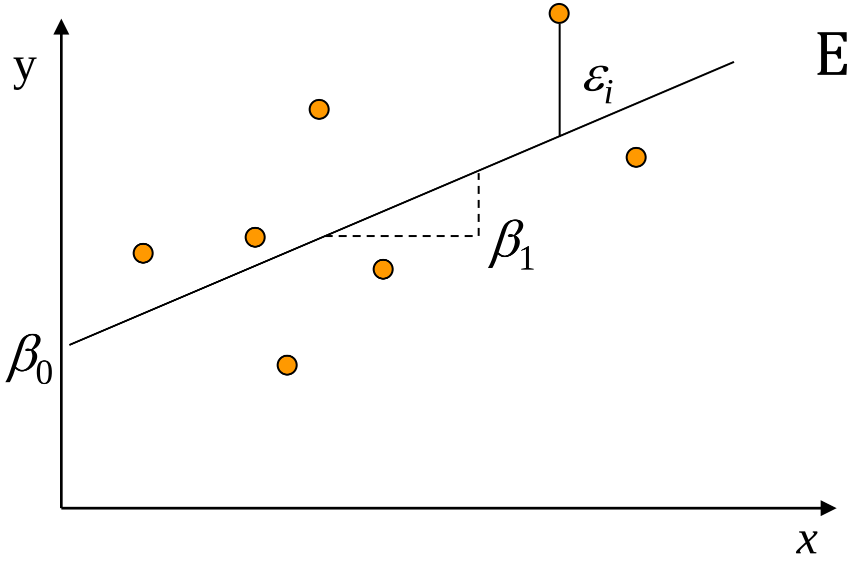
Objetivo: Explicar como uma ou mais variáveis se comportam em função de outra.

Variável dependente (resposta) - y: variável de interesse, cujo comportamento se deseja explicar.

Variável independente (explicativa) - x: variável ou variáveis que são utilizadas para explicar a variável dependente.

Modelo de regressão: equação (reta) que associa **y** e um ou vários **x**.

Modelo de Regressão Linear Simples



$$E(Y|x) = \beta_0 + \beta_1 x$$

Inclinação populacional

Intercepto populacional

Erro Aleatório

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

Variável Dependente

Variável Independente

Método dos Mínimos Quadrados

Os valores populacionais de β_0 e β_1 são desconhecidos.

Para estimá-los, é necessário minimizar o resíduo que é dado pela diferença entre o valor verdadeiro de y e seu valor estimado $\hat{y}$, ou seja,

$$\hat{\varepsilon}_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i.$$

O método utilizado na estimação desses parâmetros é o método dos mínimos quadrados.

Logo, o método dos mínimos quadrados requer que consideremos a soma dos n resíduos quadrados, denotado por SQRes:

$$\mathrm{SQRes} = \sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2 = \sum_{i=1}^{n}\left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i\right)^2$$

Insper

Suposições do modelo linear simples

- Os **erros têm distribuição normal** com média e variância constante, ou seja,

$$\varepsilon_i \sim N(0,\sigma^2).$$

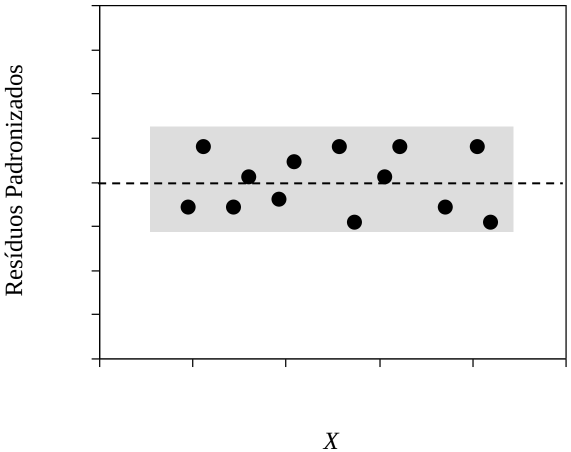
- Os **erros são independentes** entre si, ou seja,

$$\text{Corr}(\varepsilon_i , \varepsilon_j)=0$$

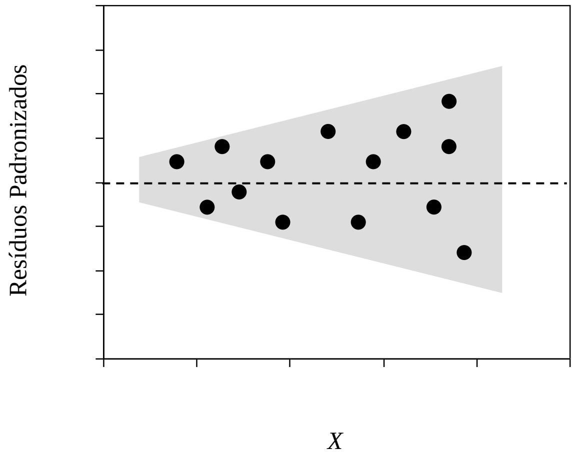
- **Modelo é linear nos parâmetros**.
- **Homocedasticidade**: $\text{Var}(\varepsilon_i) = \sigma^2$ para qualquer $i = 1, \ldots, n$.

Análise de Resíduos

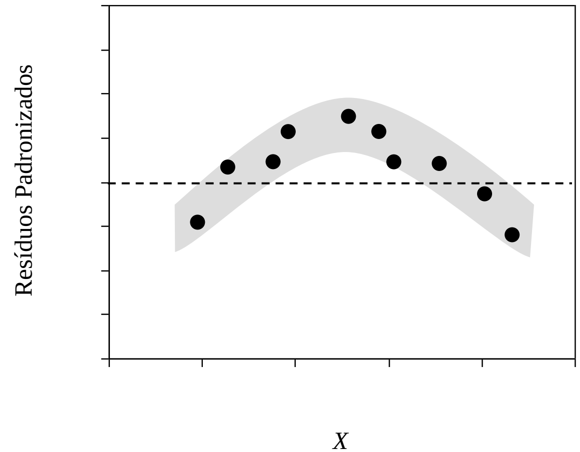
“ideal”



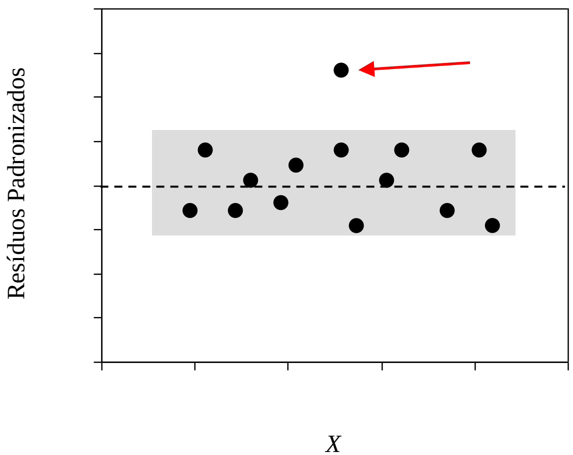
σ^2 não constante



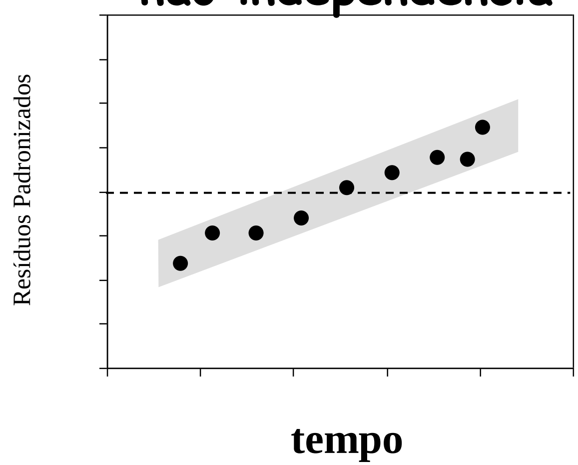
não linearidade



“outlier”



não independência



Inferência em Análise de Regressão

Usualmente, uma das hipóteses em análise de regressão é avaliar a significância da regressão.

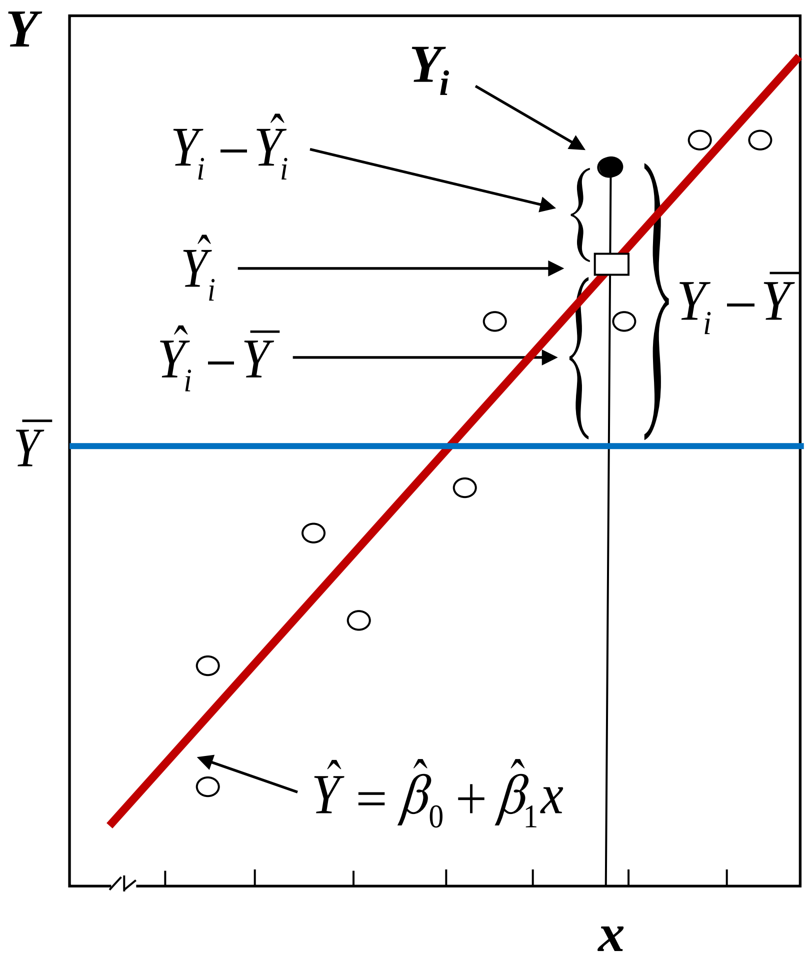
Ou seja,

H_0: $\beta_1 = 0$ → não há relação entre x e Y

H_1: $\beta_1 \neq 0$ → há relação entre x e Y

Para realizar esse teste de hipóteses, será necessário atribuir distribuição aos erros ε_i, além de outras suposições ao modelo.

Qualidade do ajuste



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

$$\text{SQT} = \text{SQReg} + \text{SQRes}$$

$$R^2 = \frac{\text{SQReg}}{\text{SQT}} = \frac{\text{SQT - SQRes}}{\text{SQT}} = 1 - \frac{\text{SQRes}}{\text{SQT}}$$

Coeficiente de determinação

$$0 \le R^2 \le 1$$

Interpretação do Coeficiente de determinação: mede a fração da variação total de Y explicada pela regressão.

ATENÇÃO: Associação não é causalidade

Suponha que encontremos alta correlação entre duas variáveis A e B. Podem existir diversas explicações do porque elas variam conjuntamente, incluindo:

– Mudanças em outras variáveis causam mudanças tanto em A quanto em B.

– Mudanças em A causam mudanças em B.

– Mudanças em B causam mudanças em A.

– A relação observada é somente uma coincidência (**correlação espúria**). **CUIDADO!!**

Fonte: http://leg.ufpr.br/~silvia/CE003/node77.html

Um particular problema

arquivo ipynb

Atividade 1 com contexto de regressão linear simples

Renda per capita (usado PIB per capita como proxy de renda per capita) tem alguma relação com a emissão de CO_2 produzido por um país?

Um particular problema
arquivo ipynb

Atividade 2 com contexto de regressão múltipla