Aula de hoje

Um passo para cenário mais realista com encontrado na prática

“Modelo de Regressão SIMPLES”

1º: uso do notebook Aula26_Atividade

Ciência dos dados

Sobre Inferência

Insper

Inferência

Inferência Estatística consiste em fazer afirmações (tomar decisões) sobre características populacionais utilizando evidências fornecidas por uma amostra extraída dessa população.

Resumo Cast no Spotify
T2#043 Weapons of math destruction

α: Probabilidade de erro do tipo I associado à decisão

Fixar um valor para α permite buscar uma regra de decisão (construção de uma região crítica) que aponte quais resultados amostrais levam a rejeição de H_0, ou seja, levam a concluir pelo descrito na hipótese alternativa H_A.

Usualmente, esses valores são fixados em 1%, 5% ou 10% e é chamado de **nível de significância**.

Valor-p

O valor- p é a probabilidade de obter um resultado **igual ao da amostra ou mais extremo**, sob H_0 verdadeira.

Note que se o teste for bicaudal a definição de *mais extremo* vai nos fazer considerar valores simétricos nas duas pontas.

Regra geral de decisão ➔

se valor-p < α, então rejeita-se a hipótese nula.

Ciência dos dados

Modelo de regressão SIMPLES

Motivação

"O acidente espacial *Challenger*, ocorrido em janeiro de 1986, foi o resultado da falha em O-rings usados para selar juntas no motor do foguete. Essa falha ocorreu por causa de temperaturas extremamente baixas do ambiente na hora do lançamento.

Antes do lançamento, havia dados sobe a ocorrência de falha no O-ring e sobre a temperatura correspondente para os 24 lançamentos anteriores." (Montgomery e Runger, 2018, Capítulo 11).

Entretanto, a <u>**exposição inadequada dos dados**</u>, não permitiu que fosse tomada a decisão correta sobre fazer ou não o lançamento *Challenger*.

Fonte: https://www.vice.com/pt_br/article/ae7xka/a-historia-por-tras-da-explosao-do-onibus-espacial-challenger

Motivação

Gráfico apresentado à NASA antes do lançamento do *Challenger*

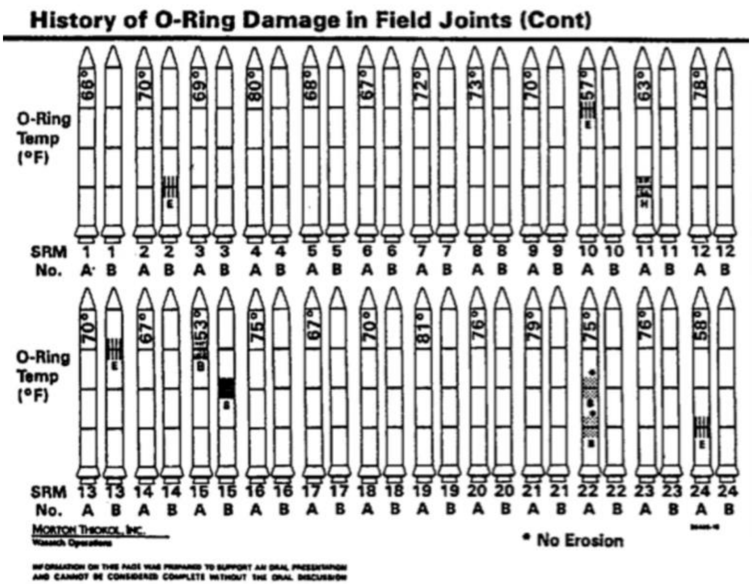
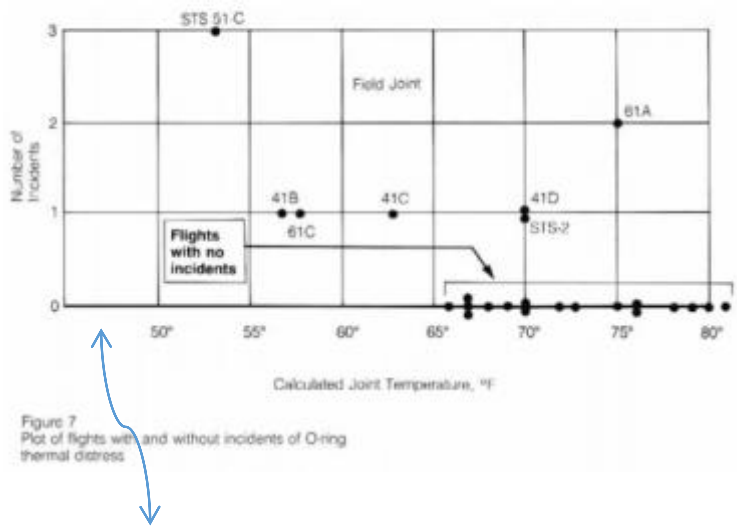


Gráfico desenvolvido por outra Comissão



26º-29º F: temperatura ambiente no dia anterior ao lançamento do *Challenger*

Fonte: https://www.vice.com/pt_br/article/ae7xka/a-historia-por-tras-da-explosao-do-onibus-espacial-challenger

Ferramentas Estatísticas

	Ferramentas estatísticas
Duas variáveis qualitativas	Tabela cruzadas (com uso de *normalize* adequado ao problema); Gráficos de barras (empilhados ou *stacked*); entre outras
Duas variáveis quantitativas	Medidas de associação; Gráfico de dispersão; entre outras
Uma variável de cada	Medidas-resumo da variável quantitativa segmentando por rótulo da variável qualitativa; Histograma (ou boxplot) da variável quantitativa segmentando por rótulo da variável qualitativa; entre outras

Modelo de regressão

Objetivo: Explicar como uma variável se comporta em função de outras.

Variável dependente ou resposta ou *target*: variável aleatória de interesse, cujo comportamento se deseja explicar.

Variáveis independentes ou preditoras ou explicativas ou *features*: variáveis que são utilizadas para explicar o comportamento da variável dependente.

Modelo de regressão

Identifique a variável explicativa e a resposta nos estudos a seguir:

- Anos de estudos e salário de chefe de famílias brasileiras;
- Gasto em supermercado e salário de chefe de famílias brasileiras;
- Nota na prova e tempo de estudo de alunos da escola AAA;
- Preço de um imóvel e metros quadrados de área construída.

Aula de hoje:

- Renda vs CO2 *dataset*:
  - compreender estimação dos parâmetros via fórmulas e gráficos para análise de resíduos
  - compreender resultados do **`statsmodels.OLS`** no caso de um modelo linear simples

Problema - GAPMinder

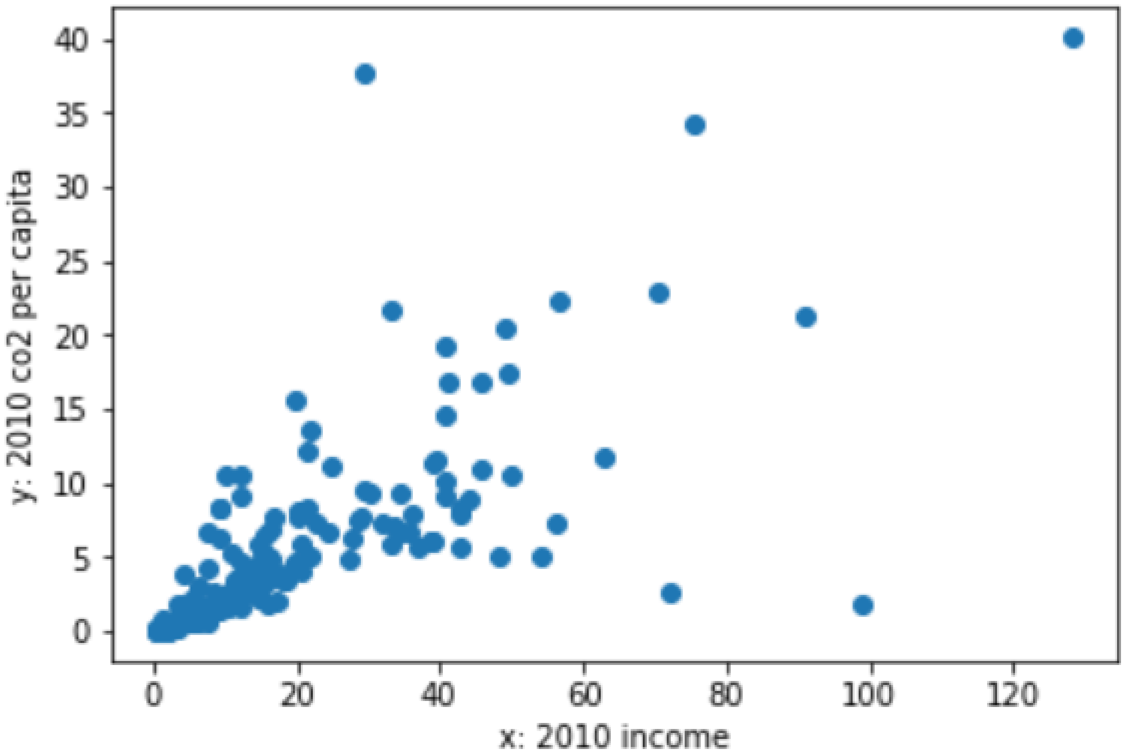
Problema: Considere que o objetivo aqui seja **explicar/prever** a emissão de gás carbono (CO_2) per capita de um país em função da renda (PIB) per capita.

$\boldsymbol{y}$**: variável dependente (resposta)** – emissão de CO_2.

$\boldsymbol{x}$**: variável independente (explicativa) –** renda (PIB) per capita.

Modelo de regressão linear simples: modelo que associa $\boldsymbol{y}$ em função de uma variável explicativa $\boldsymbol{x}$.

Problema - GAPMinder



Problema - GAPMinder

Modelo geral:

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

Problema:

$$E(\text{emissão CO2}|Renda) = \beta_0 + \beta_1 Renda$$

Significado dos termos do modelo geral:

Intercepto populacional

Inclinação populacional

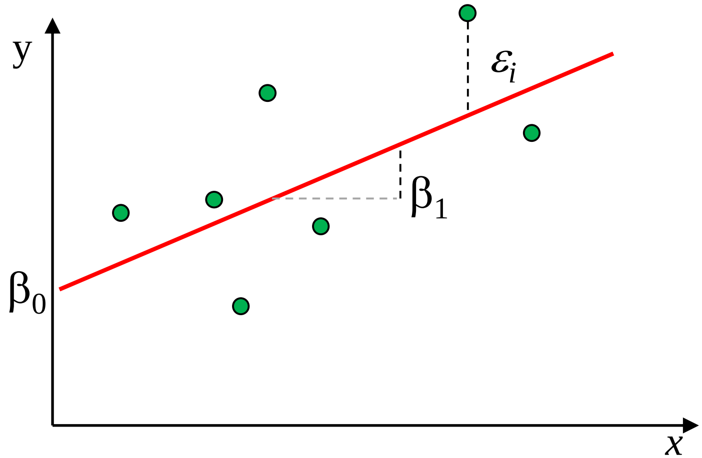
$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

Variável Dependente

Variável Independente

Erro Aleatório

Modelo de regressão linear simples



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

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,

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

O método utilizado na estimação desses parâmetros é o método dos mínimos quadrados. (se interessar veja a demonstração dos betas)

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

$$SQRes = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)^2$$

Resultados encontrados pelo MMQ:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})\,(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1\overline{x}$$

A mesma estratégia pode ser estendida a regressões com mais variáveis exemplicativas (próxima aula).

Inferência...

Teste de hipóteses

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

$H_0: \beta_1 = 0$ ➔ H_0: não há relação entre x e Y
$H_1: \beta_1 \neq 0$ ➔ H_1: 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.

Suposições do modelo

- 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, $\text{Corr}(\varepsilon_i, \varepsilon_j)=0$, para qualquer $i \neq j$.

- **O modelo é linear nos parâmetros**.

- **Homocedasticidade**: $\text{Var}(\varepsilon_i) = \sigma^2$ para qualquer $i = 1, \dots, n$.

Suposições do modelo

Como verificar essas suposições?

- Os **erros têm distribuição normal** com média e variância constante, ou seja, $\varepsilon_i \sim N(0, \sigma^2)$.

Teste Omnibus e Teste Jarque-Bera

- Os **erros são independentes** entre si, ou seja, $Corr(\varepsilon_i, \varepsilon_j) = 0$, para qualquer $i \neq j$.

Teste Durbin-Watsom

- O modelo é linear nos parâmetros.

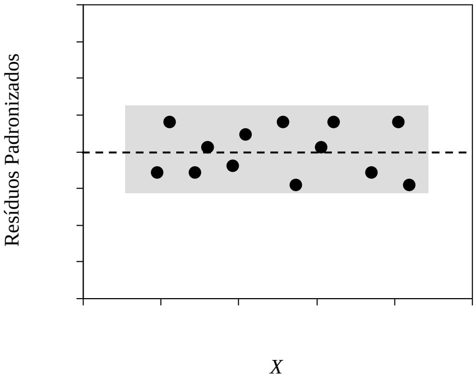
Montgomery e Runger (2018, Seção 11-7.1 e Seção 11.7-2)

- Homocedasticidade: $Var(\varepsilon_i) = \sigma^2$ para qualquer $i = 1, \ldots, n$.

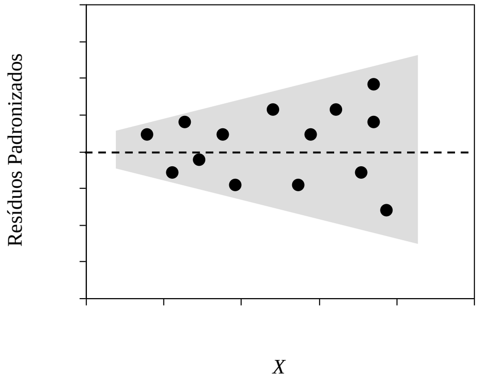
Montgomery e Runger (2018, Seção 11-7.1)

Análise de resíduos

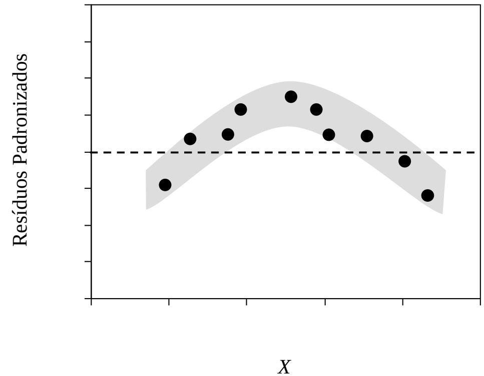
"ideal"



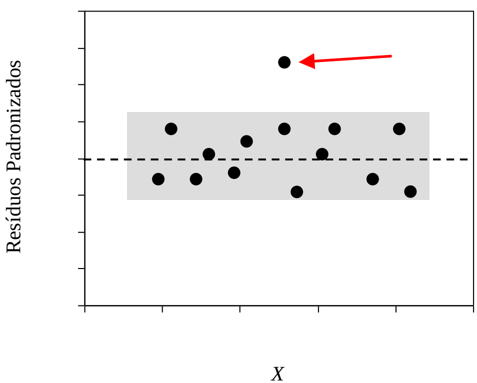
σ^2 não constante



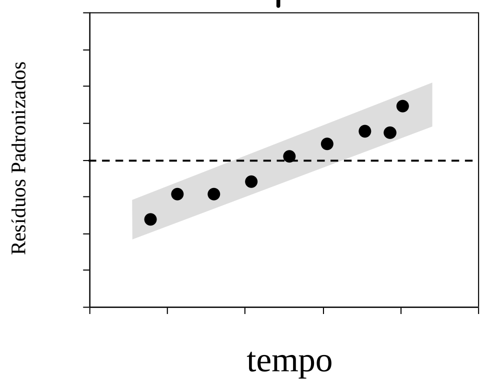
não linearidade



"outlier"

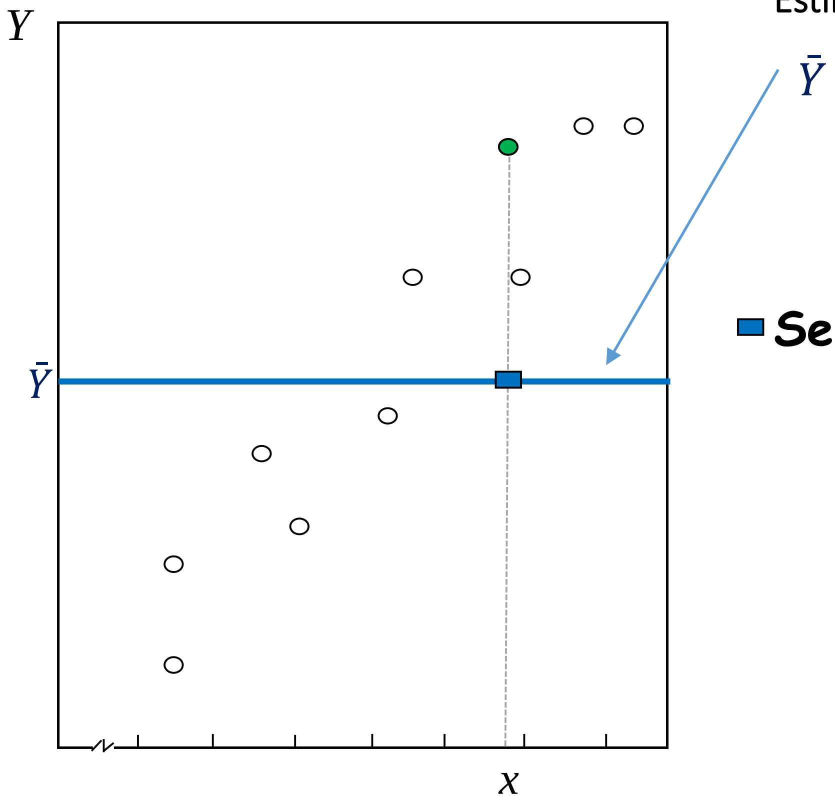


não independência



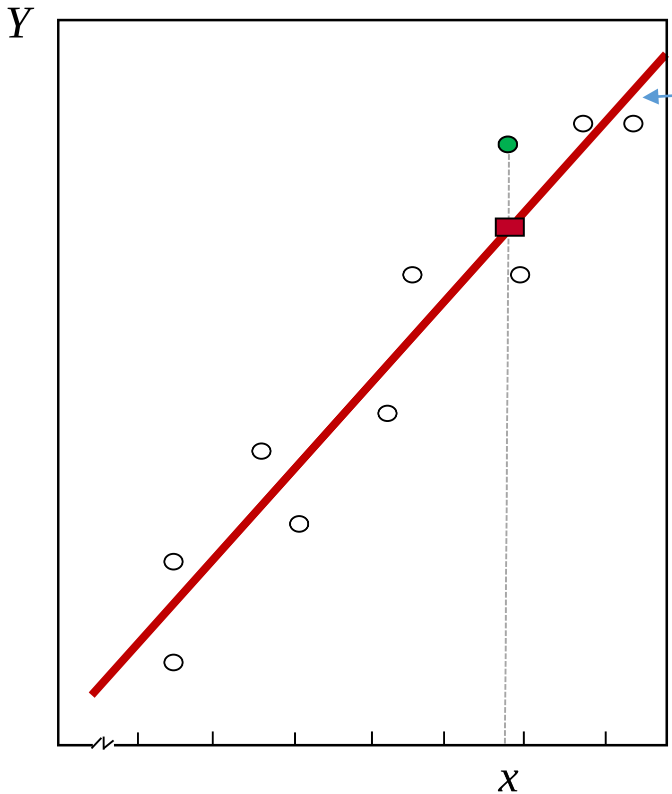
Qualidade do ajuste

Estimar valor de Y: considerando a média $\bar{Y}$.



- **Seria uma bom valor predito?**

Qualidade do ajuste

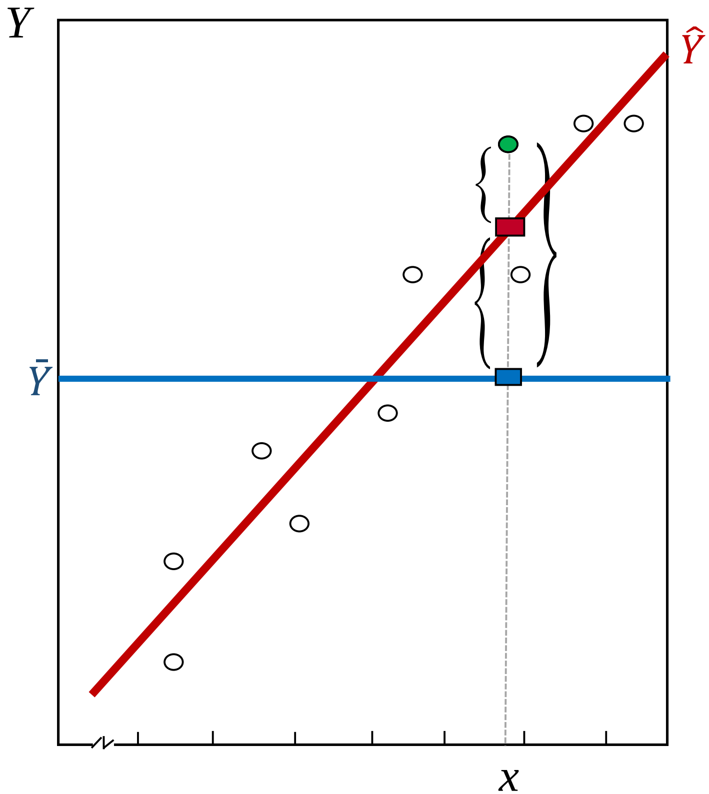


Estimar valor de Y: considerando a reta $\hat{Y}$.

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

Seria uma bom valor predito?

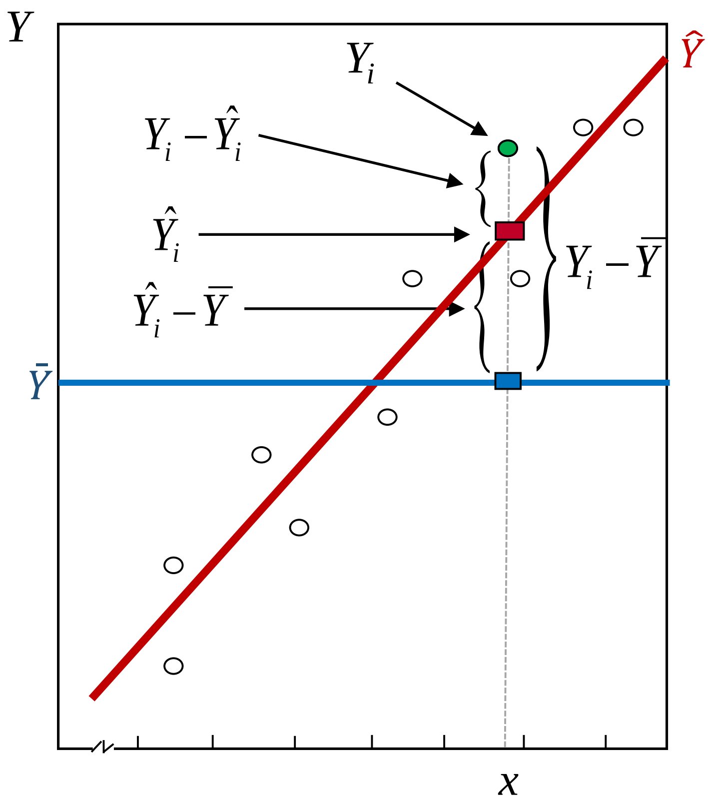
Qualidade do ajuste



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

SQT = SQReg + SQRes

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

$0 \leq R^2 \leq 1$

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

Atividade

- Download do notebook pelo Github
- Fazer individual e discutir em grupo
- Usar arquivo:

`Aula26_Atividade_....ipynb`