# Escola Politécnica de Pernambuco

*Especialização em Ciência de Dados e Analytics*

# Estatística Computacional

## Aula 3.1 – Modelos de Regressão – PARTE I


Prof. Dr. Rodrigo Lins Rodrigues

***rodrigo.linsrodrigues@ufrpe.br***

# Recapitulando o que vimos

- Estatística Descritiva ou Exploratória:
  - ✓Medidas de tendência central;
  - ✓Medidas de dispersão;
  - ✓Representação gráfica;
  - ✓Representação tabular;
  - ✓Correlação;
- Testes de hipóteses;
  - ✓Para um valor de referência;
  - ✓Para populações independentes;
  - ✓Para dados pareados;
- ANOVA;

# Conteúdo programático

- ✓ Entendimento sobre Regressão Linear Simples e Múltipla;
- ✓ Diferenças entre Correlação x Regressão;
- ✓ Conhecer os pressupostos da regressão;
- ✓ Estimar parâmetros do modelo;
- ✓ Compreender o coeficiente de determinação;
- ✓ Modelo de Regressão Logística;
- ✓ Prática computacional de Regressão.

# ... Sobre as avaliações

- **Avaliações Teóricas:**
  - ✓ Exercícios em sala de aula – 40%
  - ✓ Participação em sala – 10%

- **Avaliações Práticas**
  - ✓ Realização de projetos práticos – 25%
  - ✓ Apresentação dos resultados – 25%

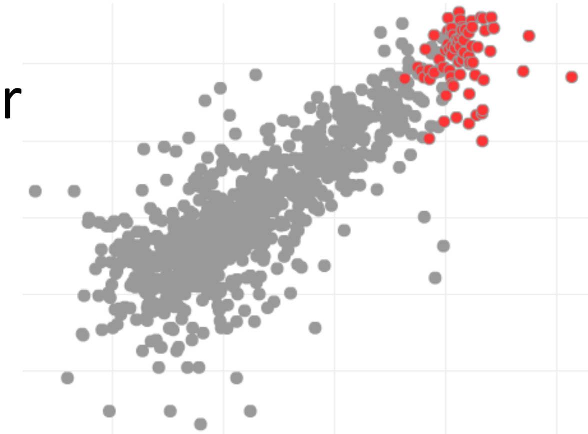
# Introdução a Regressão

# Introdução a Regressão

- A finalidade de uma análise de regressão é **estimar valores de uma variável**, com base em **valores conhecidos de outra** variável:

  ✓ Um economista pode tentar explicar as variações na procura de automóveis usados em termos de desemprego;

  ✓ Um agricultor pode suspeitar que a quantidade de fertilizante por ele usada tenha influenciado a safra.

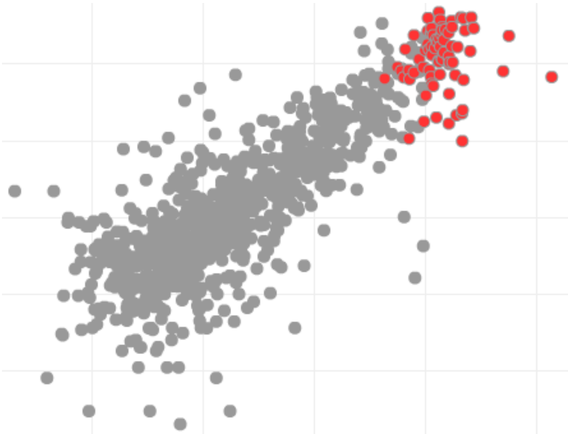


# Introdução a Regressão

- É uma das técnicas **mais utilizadas** na academia e no mercado;

- É importante enfatizar que todo modelo de regressão deve ser **definido com base na teoria e na experiência** do pesquisador;

- **Evitando relações absurdas** dentro de um contexto de pesquisa específico.



# Introdução a Regressão

Qual a diferença entre **Correlação** e **Regressão**?

# Introdução a Regressão

- **Correlação:** medida ***descritiva*** que mede força da relação entre duas variáveis quantitativas;

- **Regressão:** A finalidade é ***estimar valores*** de uma variável, com base em valores ***conhecidos da outra*** *a partir de um modelo*;

# Introdução a Regressão

- Existem diversos tipos de modelos de regressão;
- Conhecidos como G.L.M (*Modelos Generalizados*);
- Os principais são:

| Modelo de Regressão | Variável Dependente | Distribuição | Função (variável dependente) |
|---|---|---|---|
| Regressão Linear | Quantitativa | Normal | $\hat{Y}$ |
| Logística Binária | Binária | Bernoulli | $\ln\left(\frac{p}{1-p}\right)$ |
| Logística Multinominal | Qualitativa (y>2) | Binomial | $\ln\left(\frac{P_m}{1-P_m}\right)$ |
| Regressão de Poisson | Quantitativa (contagem) | Poisson | $ln(\lambda)$ |

# Introdução a Regressão

- Neste curso iremos conhecer dois tipos:

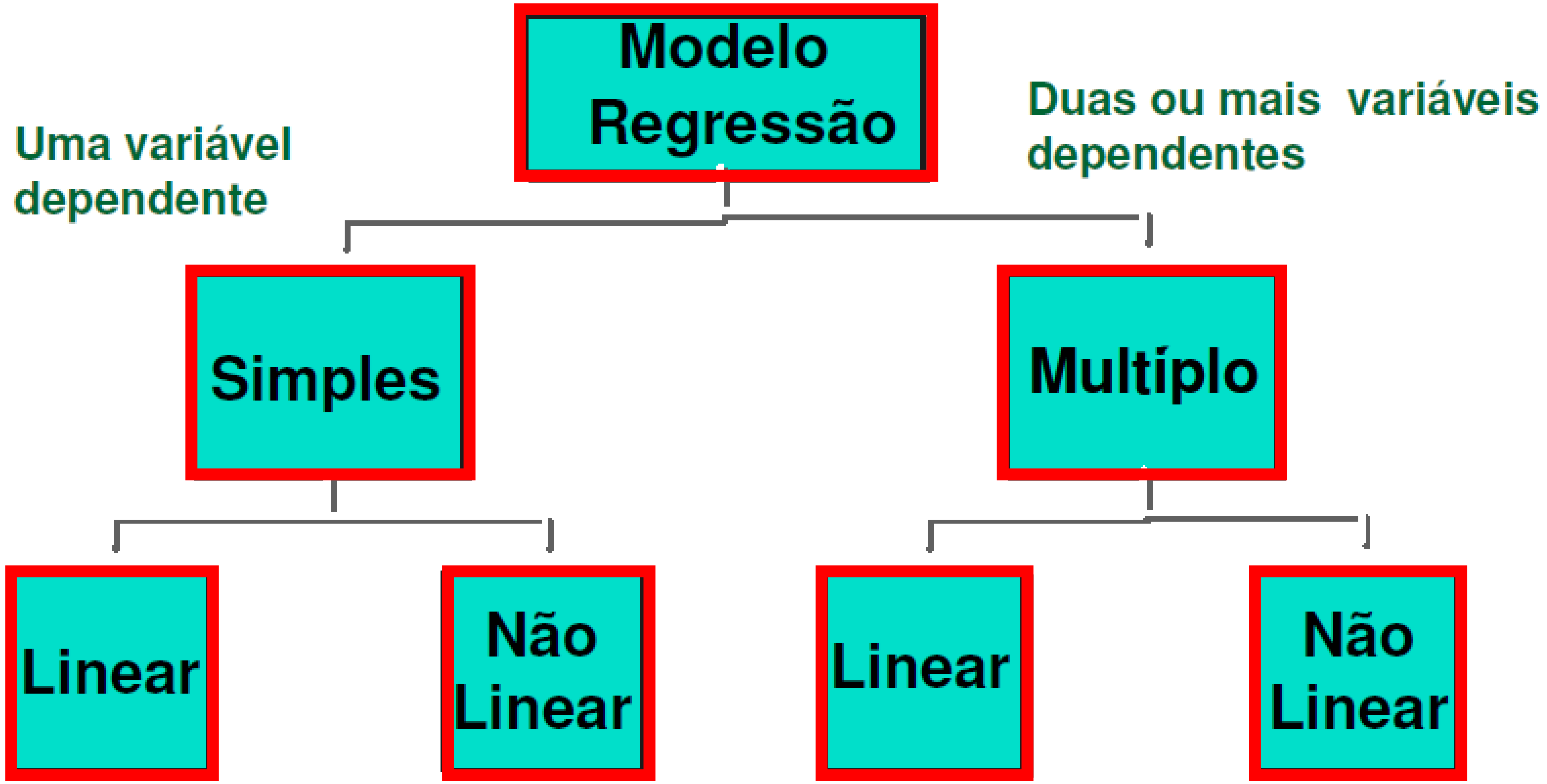
| Modelo de Regressão | Variável Dependente | Distribuição | Função (variável dependente) |
|---|---|---|---|
| **Regressão Linear** | **Quantitativa** | **Normal** | $\hat{Y}$ |
| **Logística Binária** | **Binária** | **Bernoulli** | $\ln\left(\frac{p}{1-p}\right)$ |
| Logística Multinominal | Qualitativa (y>2) | Binomial | $\ln\left(\frac{P_m}{1-P_m}\right)$ |
| Poisson | Quantitativa (contagem) | Poisson | $ln(\lambda)$ |

# Introdução a Regressão

- Neste curso iremos conhecer dois tipos:

| Modelo de Regressão | Função (variável dependente) |
|---|---|
| **Regressão Linear** | $\hat{Y} = \alpha + \beta_1 . X_{1i} + \beta_2 . X_{2i} + \cdots + \beta_k . X_{ki}$ |
| **Logística Binária** | $\ln\left(\frac{p}{1-p}\right) = \hat{Y} = \alpha + \beta_1 . X_{1i} + \beta_2 . X_{2i} + \cdots + \beta_k . X_{ki}$ |

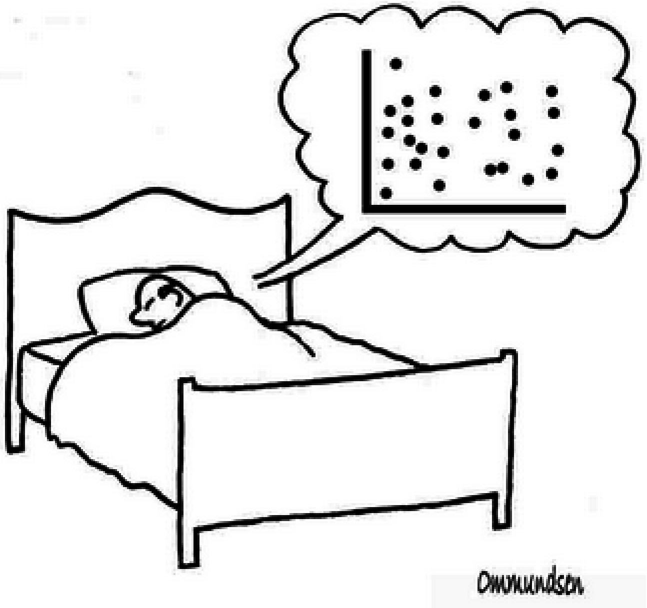
# Introdução a Regressão



# Introdução a Regressão

- **Regressão Linear Simples:**
  - ✓É um tipo de regressão linear com apenas **uma variável independente** e apenas **uma variável dependente**;

  - ✓Boa parte do nosso raciocínio processa funções e modelos linear simples:
    - Se a inflação do mês que vem for maior, então compraremos menos;
    - Se eu fizer mais horas de esporte por semana, poderei reduzir meus níveis de colesterol;
    - Se eu estudar mais, poderei ter mais oportunidades de aumentar minha renda;



# Introdução a Regressão

- **Regressão Linear Simples:**
  - ✓Para o exemplo “aumento de estudo”, outras variáveis afetam a renda;
  - ✓A regressão linear simples, simplifica a relação:

$$\boxed{renda = \hat{Y} = \alpha + \beta . edu + \mu}$$

  - ✓Normalmente busca-se identificar os sinais e valor de $\beta$;
  - ✓Para analisar se a influência é positiva ou negativa.

# Introdução a Regressão

- **Regressão Linear Múltipla**
  - ✓ Apresenta a mesma lógica da Regressão Linear Simples;
  - ✓ Com a inclusão de mais de uma variável explicativa ($X_i$) no modelo;
  - ✓ A utilização de muitas variáveis explicativas dependerá da experiência e do bom senso do pesquisador;
  - ✓ É importante que seja analisado criteriosamente o poder de cada variável inserida no modelo;

# Introdução a Regressão

- **Exemplo Regressão Linear Múltipla**
  - ✓ O gestor de uma fábrica de chocolates está interessado em avaliar o impacto da produção de embalagens para chocolates ($emb$) e da elevação da produção ($ton$) nos custos indiretos ($ci$) da empresa;
  - ✓ Neste caso existem duas variáveis explicativas;
  - ✓ Portanto o modelo de regressão múltipla é definido por:

$$\boxed{ci = \alpha + \beta_1 . emb + \beta_2 . ton + \mu}$$


# Modelo de Regressão Linear

# Modelo de Regressão Linear

1. Determinar como duas variáveis se relacionam;
2. Estimar a **função que determina a relação** entre as variáveis;
3. Usar a equação ajustada para **prever valores da variável dependente**:

Modelo de Regressão Linear Simples

$$Y_i = \alpha + \beta X_i + \mu_i$$

# Modelo de Regressão Linear

**Pressupostos do modelo de Regressão Linear**

- A relação entre *X* e *Y* é Linear;
- Os valores de X são fixos, isto é, X não é uma variável aleatória;
- A média dos erros é nula, isto é:

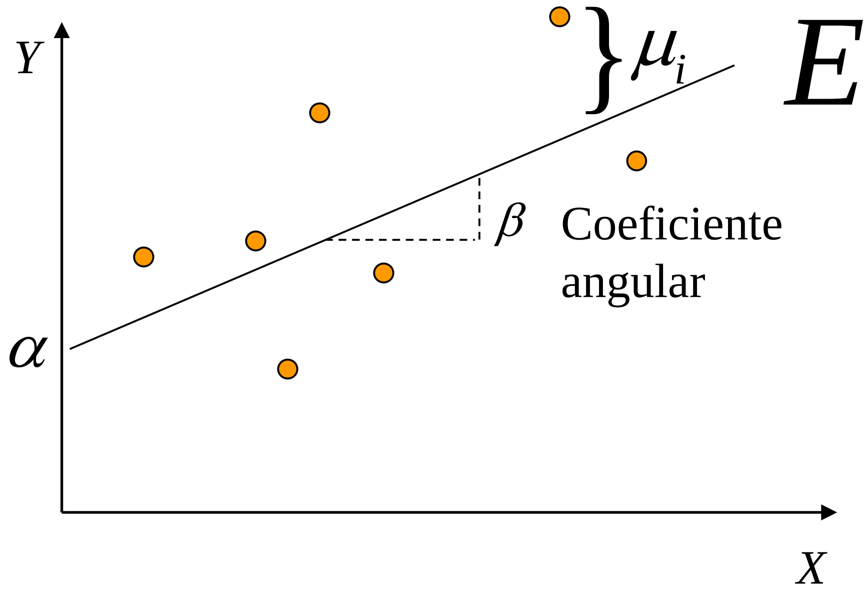
$$E(\mu_i) = 0 \qquad i = 1, 2, \ldots, n$$

- O erro em uma observação é **não correlacionado** com o erro em qualquer outra observação;

$$Var(\mu_i) = E(\mu_i) - [E(\mu_i)]^2 = E(\mu_i^2) = \sigma^2$$

- Os erros têm **distribuição normal**

# Modelo de Regressão Linear




$$E(Y) = \alpha + \beta x_i$$

Inclinação populacional

Intercepto populacional

Erro Aleatório

$$Y_i = \alpha + \beta X_i + \mu_i$$

Variável Dependente

Variável Independente

# Modelo de Regressão Linear

- **Estimadores $\alpha$ e $\beta$**

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

$$\alpha = \bar{Y} - \beta.\bar{X}$$

# Modelo de Regressão Linear

- **Exemplo 1:**
  - ✓ Deseja-se saber, para uma turma de 10 alunos, qual a influência da distância percorrida para se chegar à escola em relação ao tempo de percurso;
  - ✓ Elaborou-se um questionário e aplicou para os 10 alunos da turma;
  - ✓ Com as variáveis: (*y*)*Tempo para se chegar a escola* e (x) *Distância percorrida até a escola*.

# Modelo de Regressão Linear

- **Exemplo 1:**
  - ✓A tabela abaixo mostra os resultados declarados pelos alunos:

| Estudante | Tempo para chegar à escola (min) | Distância percorrida até a escola (km) |
|---|---|---|
| Gabriela | 15 | 8 |
| Dalila | 20 | 6 |
| Gustavo | 20 | 15 |
| Letícia | 40 | 20 |
| Luiz Otávio | 50 | 25 |
| Leonor | 25 | 11 |
| Ana | 10 | 5 |
| Antônio | 55 | 32 |
| Júlia | 35 | 28 |
| Mariana | 30 | 20 |

# Modelo de Regressão Linear

- **Exemplo:**
  - ✓Na verdade estamos interessados em modelar a equação que regula o fenômeno “tempo de percurso até a escola”;
  - ✓Sabemos que outras vereáveis influenciam o tempo de percurso, mas iremos delimitar em um problema linear simples;
  - ✓ Queremos modelar o problema da seguinte forma:

$$\boxed{tempo = f(dist)}$$

# Modelo de Regressão Linear

- **Exemplo:**
  - ✓Logo, o modelo de regressão simples, será:

$$tempo_i = \alpha + \beta . dist_i + \mu_i$$

- É necessário encontrar os valores dos parâmetros α e $\beta$:

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

$$\alpha = \bar{Y} - \beta . \bar{X}$$

# Modelo de Regressão Linear

- **Exemplo:**
  - ✓Constrói-se a tabela de valores:

| Estudante | Tempo ($Y_i$) | Distância ($X_i$) | $Y_i - \bar{Y}$ | $X_i - \bar{X}$ | $(X_i - \bar{X}).(Y_i - \bar{Y})$ | $(X_i - \bar{X})^2$ |
|---|---|---|---|---|---|---|
| Gabriela | 15 | 8 | -15 | -9 | 135 | 81 |
| Dalila | 20 | 6 | -10 | -11 | 110 | 121 |
| Gustavo | 20 | 15 | -10 | -2 | 20 | 4 |
| Letícia | 40 | 20 | 10 | 3 | 30 | 9 |
| Luiz Otávio | 50 | 25 | 20 | 8 | 160 | 64 |
| Leonor | 25 | 11 | -5 | -6 | 30 | 36 |
| Ana | 10 | 5 | -20 | -12 | 240 | 144 |
| Antônio | 55 | 32 | 25 | 15 | 375 | 225 |
| Júlia | 35 | 28 | 5 | 11 | 55 | 121 |
| Mariana | 30 | 20 | 0 | 3 | 0 | 9 |
| Soma | 300 | 170 | | | 1155 | 814 |
| Média | 30 | 17 | | | | |

# Modelo de Regressão Linear

- **Exemplo:**
  - ✓Por meio da planilha construída podemos calcular os parâmetros α e $\beta$:

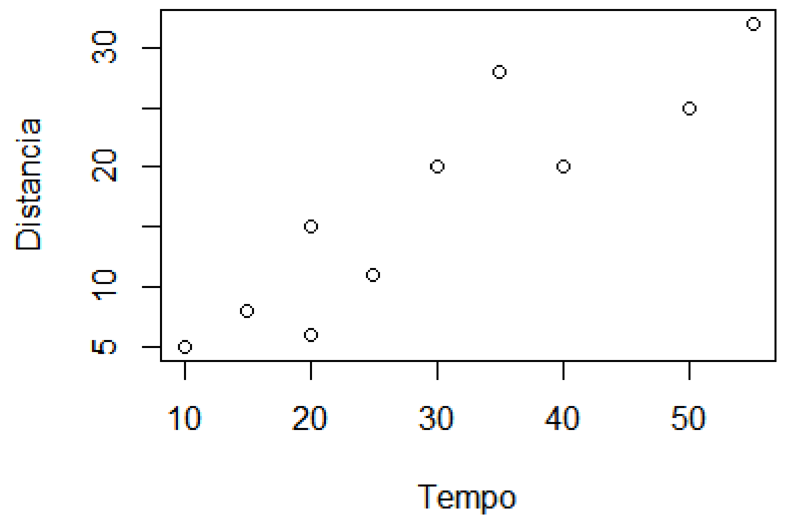
$$\beta = \frac{\sum_{i=1}^{n}(X_i - \bar{X}).(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \Rightarrow \beta = \frac{1155}{814} \Rightarrow \boxed{\beta = 1{,}4189}$$

$$\alpha = \bar{Y} - \beta.\bar{X} \Rightarrow \alpha = 30 - 1{,}4189 * 17 \Rightarrow \boxed{\alpha = 5{,}8784}$$

# Modelo de Regressão Linear

- **Exemplo:**
  - ✓Logo, a equação de regressão linear simples será:

$$\boxed{tempo_i = 5{,}8784 + 1{,}4189 * dist_i}$$



# Modelo de Regressão Linear

- O **coeficiente de determinação** é uma medida da proporção da variabilidade em uma variável que é explicada pela variabilidade da outra;
- O coeficiente de determinação é definido por:

$$R^2 = \frac{\sum_{i=1}^{n}\left(\hat{Y}_i - \bar{Y}\right)^2}{\sum_{i=1}^{n}\left(\hat{Y}_i - \bar{Y}\right)^2 + \sum_{i=1}^{n}(u_i)^2}$$

- O coeficiente está entre $0 \leq R^2 \leq 1$.

# Modelo de Regressão Linear

- Voltando ao exemplo 1 em que queríamos ver a relação de (*y)Tempo para se chegar a escola* e (x) *Distância percorrida até a escola*...

- Queremos agora verificar a **adequação do modelo linear** para os dados da pesquisa;

- Para isto precisamos construir a tabela par ao calculo do **coeficiente** de **determinação** ou **ajuste** $R^2$.

# Modelo de Regressão Linear

- Construindo a tabela, temos...

| Estudante | Tempo ($Y_i$) | Distância ($X_i$) | $\hat{Y}_i$ | $u_i(Y_i - \hat{Y}_i)$ | $(Y_i - \hat{Y}_i)^2$ | $(u_i)^2$ |
|---|---|---|---|---|---|---|
| Gabriela | 15 | 8 | 17,23 | -2,23 | 163,08 | 4,97 |
| Dalila | 20 | 6 | 14,39 | 5,61 | 243,61 | 31,45 |
| Gustavo | 20 | 15 | 27,16 | -7,16 | 8,05 | 51,3 |
| Letícia | 40 | 20 | 34,26 | 5,74 | 18,12 | 32,98 |
| Luiz Otávio | 50 | 25 | 41,35 | 8,65 | 128,85 | 74,8 |
| Leonor | 25 | 11 | 21,49 | 3,51 | 72,48 | 12,34 |
| Ana | 10 | 5 | 12,97 | -2,97 | 289,92 | 8,84 |
| Antônio | 55 | 32 | 51,28 | 3,72 | 453,00 | 13,81 |
| Júlia | 35 | 28 | 45,61 | -10,61 | 243,61 | 112,53 |
| Mariana | 30 | 20 | 34,26 | -4,26 | 18,12 | 18,12 |
| Soma | 300 | 170 | | | **1638,85** | **361,15** |
| Média | 30 | 17 | | | | |

# Modelo de Regressão Linear

- Logo, calculamos o coeficiente de determinação:

$$R^2 = \frac{\sum_{i=1}^{n}\left(\hat{Y}_i - \bar{Y}\right)^2}{\sum_{i=1}^{n}\left(\hat{Y}_i - \bar{Y}\right)^2 + \sum_{i=1}^{n}(u_i)^2} \Rightarrow R^2 = \frac{1638{,}85}{1638{,}85 - 361{,}15} \Rightarrow \boxed{R^2 = 0{,}8194}$$

- Podemos afirmar que, para a amostra estudada, 81,94% da variabilidade do tempo para se chegar a escola pode ser explicado pela variável distância.

# Modelo de Regressão Linear

- **Teste de significância dos parâmetros:**
- Agora, uma vez que o modelo foi ajustado e todos os parâmetros estimados surgem a seguinte pergunta:

  ✓ Existe realmente alguma relação linear entre X e Y ? Como podemos responder isso estatisticamente ?

- Para responder à pergunta , observamos que se $\beta = 0$ não existe relação linear explicando Y em função de X.

# Modelo de Regressão Linear

- Assim, a estatística F é uma estatística para testar:

$$\begin{cases} H_0: \beta = 0 \\ H_1: \beta \neq 0 \end{cases}$$

$$F = \frac{\text{QMReg}}{\text{QMRes}} \sim F_{1;n-2}$$

se $H_0$ **verdadeiro** (Não existe relação linear)
se $H_0$ **falso** (existe relação linear)

# Modelo de Regressão Linear

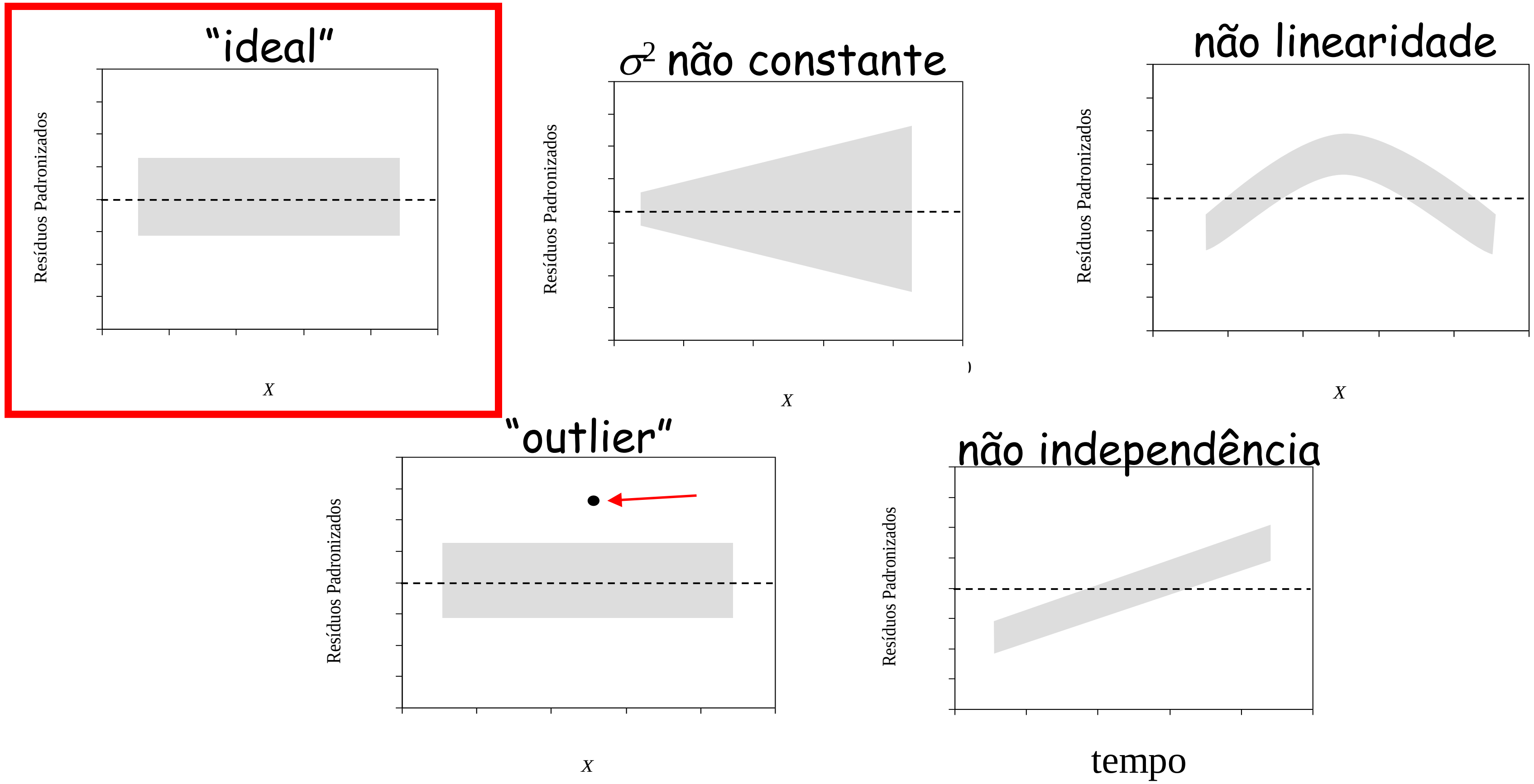
- **Análise de Resíduos:**

  ✓ É importante, após a análise de regressão, testar se os pressupostos do modelo linear se aplicam aos dados estudados;

  ✓ Resíduos representam a diferença entre o valor observado de y e o que foi predito pelo modelo de regressão;

# Modelo de Regressão Linear

- **Análise de Resíduos:**



# Modelo de Regressão Linear

- **Ajuste do modelo de regressão:**
  - Pode ser verificado através de um **gráfico de dispersão** entre X e Y;
  - Existem funções que podem ser **transformadas em modelos lineares**;
  - Existem vários tipos de funções que podemos transformar, tais como:
    - ✓ Função Potência;
    - ✓ Função Exponencial;
    - ✓ Função Hiperbólica.

# Agora é com vocês !

- Qual a diferença entre regressão linear simples e Múltipla?

- Para que serve o coeficiente de determinação e quais seu intervalo de variação ?

# Exemplos com Regressão Linear

# Exemplos com Regressão Linear

- Exemplo Regressão Linear Simples

  ✓ Para o banco de dados sobre característica de pessoas *dados.txt* na pasta Aula 3.1 - Modelos de Regressão\Scripts - Aula 3.1;

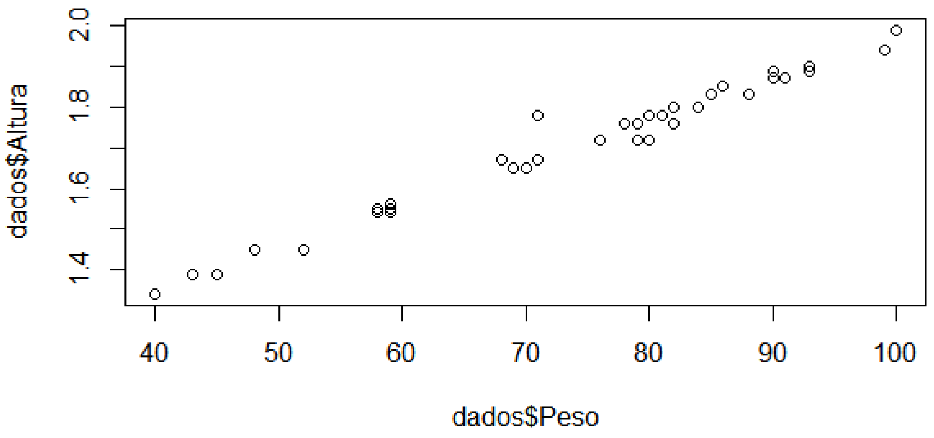
  ✓ Iremos construir um modelo de regressão simples com as variáveis *y* = *Peso* e *x* = *altura*;

  ✓ Construiremos um modelo que explique o peso de um indivíduo em função de sua altura.

# Exemplos com Regressão Linear

- Exemplo Regressão Linear Simples

```
# Ler os dados de um arquivo (interagindo com o usuário)
dados <- read.table(file.choose(),header=TRUE)

# Visão descritiva (gráfica sobre as duas variáveis)
plot(dados$Peso, dados$Altura)
```



# Exemplos com Regressão Linear

- Exemplo Regressão Linear Simples

```r
# Construindo o modelo de regressão linear múltiplo
modelo <- lm(dados$Peso~dados$Altura)

# Visualizando o modelo
modelo
```

```
Call:
lm(formula = dados$Peso ~ dados$Altura)

Coefficients:
 (Intercept)  dados$Altura
      -88.21         95.16
```

# Exemplos com Regressão Linear

- Exemplo Regressão Linear Simples

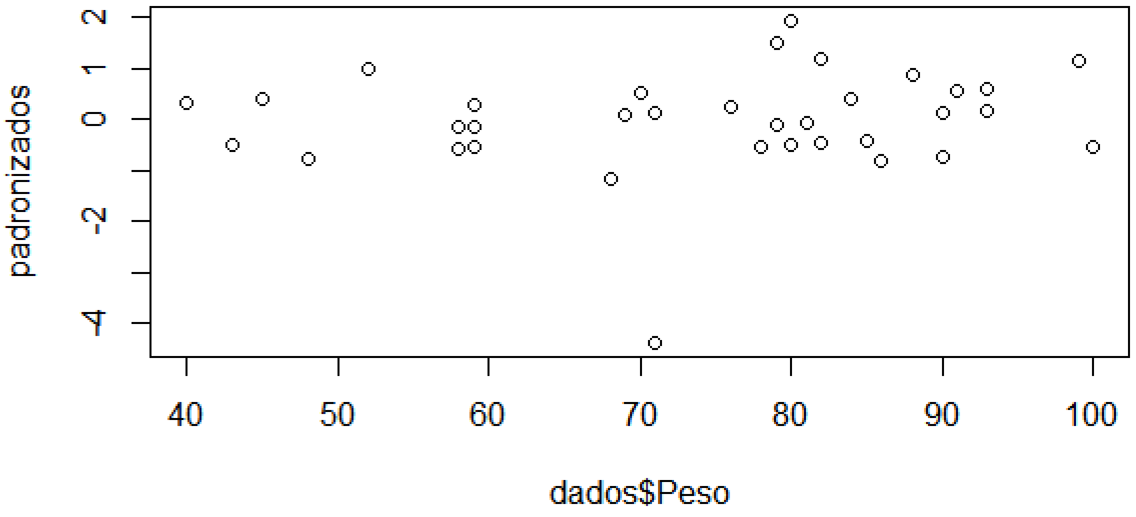
```r
# Mostrando o resultado do modelo
summary(modelo)
```

```
Call:
lm(formula = dados$Peso ~ dados$Altura)

Residuals:
     Min       1Q   Median       3Q      Max 
-10.1720  -1.1639   0.1989   1.0698   4.5376 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)   -88.214      3.940  -22.39   <2e-16 ***
dados$Altura   95.161      2.308   41.23   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.355 on 37 degrees of freedom
Multiple R-squared:  0.9787,	Adjusted R-squared:  0.9781 
F-statistic:  1700 on 1 and 37 DF,  p-value: < 2.2e-16
```

# Exemplos com Regressão Linear

- Exemplo Regressão Linear Simples

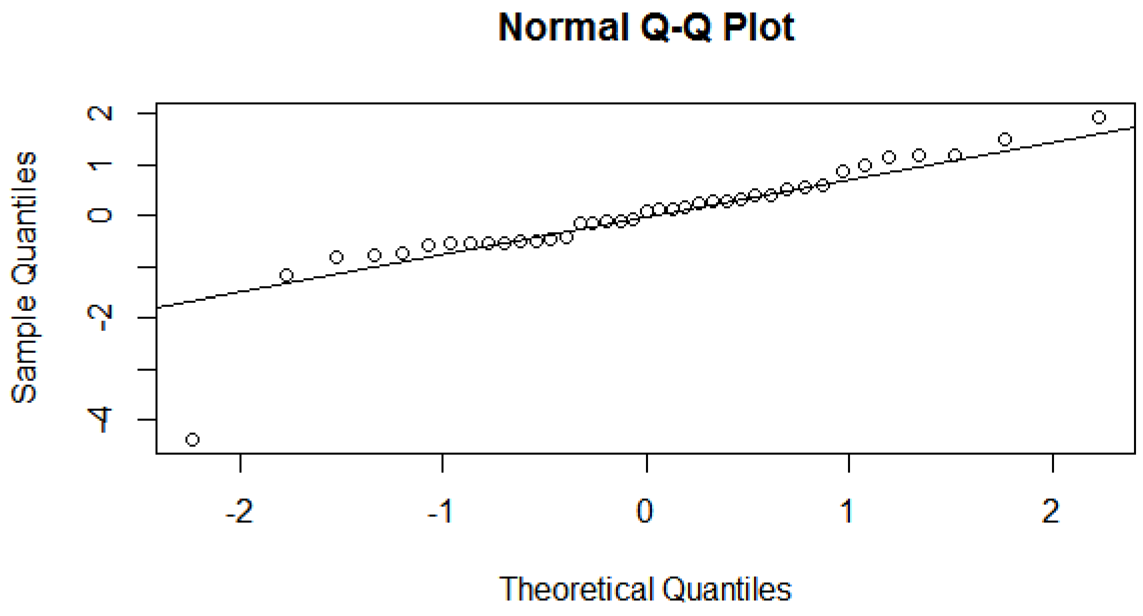
```
# Gerando os resíduos padronizados
padronizados <- rstandard(modelo)

# Plotando o gráfico dos resíduos
plot(dados$Peso, padronizados)
```



# Exemplos com Regressão Linear

- Exemplo Regressão Linear Simples

```
# Verificando a normalidade dos resíduos pelo qqplot
qqnorm(padronizados)
qqline(padronizados)
```



# Exemplos com Regressão Linear

- **Exemplo Regressão Linear Múltipla**

  ✓ Vamos utilizar o mesmo exemplo;

  ✓ Adicionaremos a variável *“Renda”*;

  ✓ Este modelo terá a variável $y$ = *Peso* e duas variáveis explicativas $x1$ = *Altura* e $x2$ = *Renda*;

# Exemplos com Regressão Linear

- **Exemplo Regressão Linear Múltipla**

```
# Ler os dados de um arquivo (interagindo com o usuário)
dados <- read.table(file.choose(),header=TRUE)

# Construindo o modelo de regressão linear múltiplo
modelo <- lm(dados$Peso~dados$Altura + dados$Renda)
```

# Exemplos com Regressão Linear

- **Exemplo Regressão Linear Múltipla**

```r
# Ler os dados de um arquivo (interagindo com o usuário)
dados <- read.table(file.choose(),header=TRUE)

# Construindo o modelo de regressão linear múltiplo
modelo <- lm(dados$Peso~dados$Altura + dados$Renda)

# Visualizando o modelo
modelo
```

```
Call:
lm(formula = dados$Peso ~ dados$Altura + dados$Renda)

Coefficients:
 (Intercept)  dados$Altura   dados$Renda
  -8.821e+01     9.516e+01     7.960e-07
```

# Exemplos com Regressão Linear

- **Exemplo Regressão Linear Múltipla**

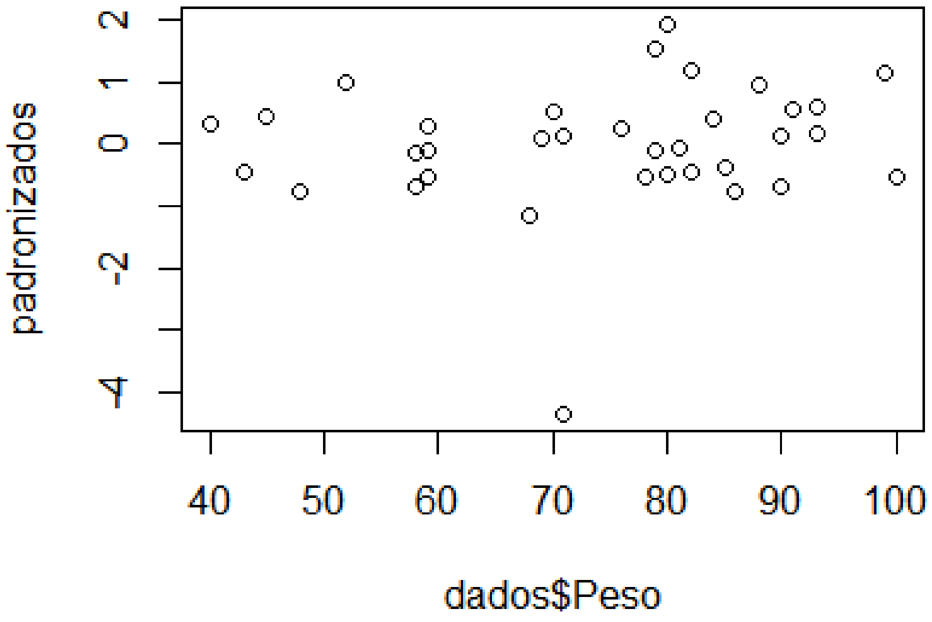
```
# Mostrando o resultado do modelo
summary(modelo)
```

```
Call:
lm(formula = dados$Peso ~ dados$Altura + dados$Renda)

Residuals:
     Min       1Q   Median       3Q      Max 
-10.1711  -1.1653   0.1996   1.0704   4.5380 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)  -8.821e+01  3.994e+00 -22.086   <2e-16 ***
dados$Altura  9.516e+01  2.349e+00  40.519   <2e-16 ***
dados$Renda   7.960e-07  2.212e-04   0.004    0.997    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.388 on 36 degrees of freedom
Multiple R-squared:  0.9787,	Adjusted R-squared:  0.9775 
F-statistic: 826.8 on 2 and 36 DF,  p-value: < 2.2e-16
```

# Exemplos com Regressão Linear

- **Exemplo Regressão Linear Múltipla**

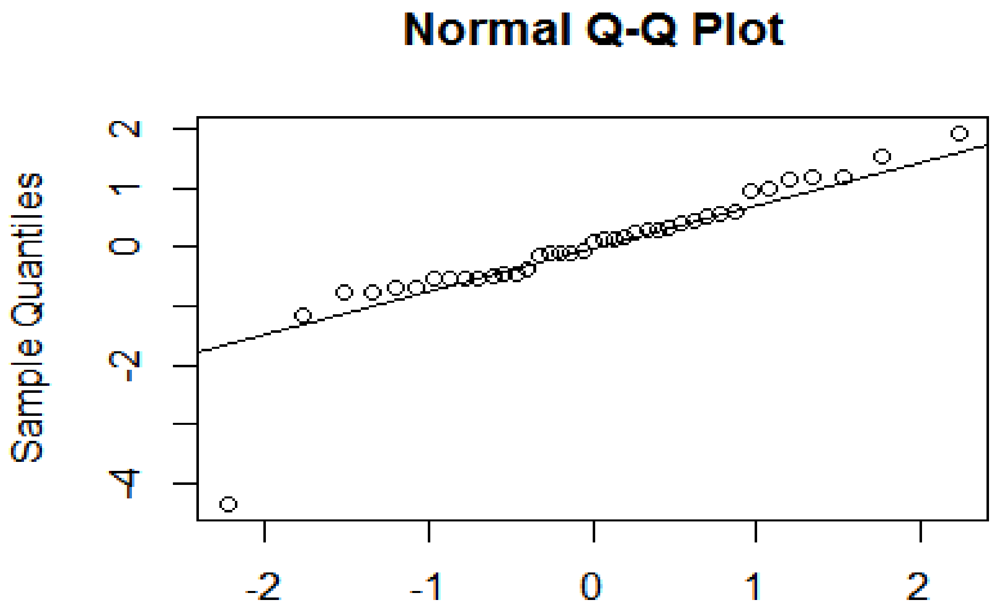
```
# Mostrando o resultado do modelo
summary(modelo)

# Gerando os residúos padronizados
padronizados <- rstandard(modelo)

# Plotando o gráfico dos resíduos
plot(dados$Peso, padronizados)
```



# Exemplos com Regressão Linear

- **Exemplo Regressão Linear Múltipla**

```
# Verificando a normalidade dos resíduos pelo qqplot
qqnorm(padronizados)
qqline(padronizados)
```



# Dúvidas ?

- **Contatos:**
  - ✓Email: rodrigo.linsrodrigues@ufrpe.br
  - ✓Facebook: /rodrigomuribec