IPS Instituto Politécnico de Setúbal
Escola Superior de Tecnologia de Setúbal

# *MÉTODOS ESTATÍSTICOS*

## Regressão Linear Simples

Licenciatura em Engenharia Informática

Departamento de Matemática
Escola Superior de Tecnologia de Setúbal
Instituto Politécnico de Setúbal
2021-2022

## Dados Bivariados

- Por vezes  a população que se pretende estudar, aparece sob a forma de pares de valores, isto é, cada indivíduo ou resultado experimental, contribui com um conjunto de dois valores.
- É o que acontece quando se pretende estudar dois atributos da mesma população visando investigar em que medida eles se relacionam, isto é, de que modo a variação de um deles exerce influência na variação do outro.
- Os atributos podem ser ambos quantitativos, ambos qualitativos ou um de cada tipo. Apenas vamos considerar atributos quantitativos.

# Objetivo

Estudar a relação entre duas

**Variáveis Quantitativas.**

## Objetivo

Estudar a relação entre duas **variáveis quantitativas**.

## Exemplos

- a relação entre a hora do dia e a temperatura atmosférica;
- a relação entre a idade e a altura das crianças;
- a relação entre o tempo de prática de atividade física e o ritmo cardíaco;
- a relação entre o tempo de estudo e a nota no teste;
- a relação entre a taxa de desemprego e a taxa de criminalidade;
- a relação entre a esperança de vida e a taxa de analfabetismo;
- ...

# Regressão Linear Simples

## Objetivo

Estudar a relação entre duas **variáveis quantitativas**.

Para atingir este objetivo vamos investigar a presença ou ausência de **relação linear** entre as duas variáveis. Essa investigação será feita de modo a:

- quantificar a força dessa relação: **correlação**;
- explicar a forma dessa relação: **regressão**.

Para quantificar a força dessa relação, **correlação**, vamos recorrer a:

- métodos gráficos: **diagrama de dispersão**;
- indicadores numéricos: **coeficiente de correlação linear**.

Para explicar a forma dessa relação, **regressão**, vamos recorrer a um

- modelo matemático: **equação da reta** $\rightarrow$ $\mathbf{y = a + bx}$.

# Regressão Linear Simples

## Objetivo

Construção de um modelo matemático que expresse a relação tipo linear existente entre duas variáveis quantitativas, tendo por base os correspondentes valores amostrais, isto é, resumir os valores amostrais através de uma reta

$$Y = a + bX$$

que dê a informação de como se refletem em $Y$ as mudanças processadas em $X$.

Neste caso tem-se:

- **Variável independente**, explicativa ou explanatória representada por $X$;
- **Variável dependente**, explicada ou resposta representada por $Y$.

# Metodologia

1. *Verificar a existência ou não de uma relação linear*

## Diagrama de Dispersão ou Nuvem de Pontos

- **Diagrama de dispersão** ou **Nuvem de pontos** é uma representação gráfica para os dados bivariados, em que cada par de dados $(x_i, y_i)$ é representado por um ponto de coordenadas $(x_i, y_i)$, num sistema de eixos coordenados.
- Este tipo de representação é muito útil, pois permite realçar algumas propriedades entre os dados, nomeadamente no que diz respeito ao tipo de correlação entre os valores de $x$ e os valores de $y$.
- Existe correlação linear quando é possível "imaginar" uma reta (com declive diferente de zero) que passa pela "nuvem" de pontos.

## Exemplo 1

Um psicólogo efetuou uma pesquisa com o objetivo de analisar a forma como as crianças aprendem um determinado jogo. Das diversas variáveis observadas foram registados o número de jogos efetuados $(X)$ e o número de erros realizados no jogo $(Y)$:

| Número de jogos | Número de erros |
|---|---|
| x | y |
| 1 | 8 |
| 2 | 5 |
| 3 | 8 |
| 4 | 7 |
| 6 | 6 |
| 7 | 3 |
| 8 | 5 |
| 9 | 2 |
| 10 | 4 |
| 11 | 2 |

1. Acha que os dados apresentam uma tendência linear?

| Número de jogos | Número de erros |
|---|---|
| x | y |
| 1 | 8 |
| 2 | 5 |
| 3 | 8 |
| 4 | 7 |
| 6 | 6 |
| 7 | 3 |
| 8 | 5 |
| 9 | 2 |
| 10 | 4 |
| 11 | 2 |

tabela de dados

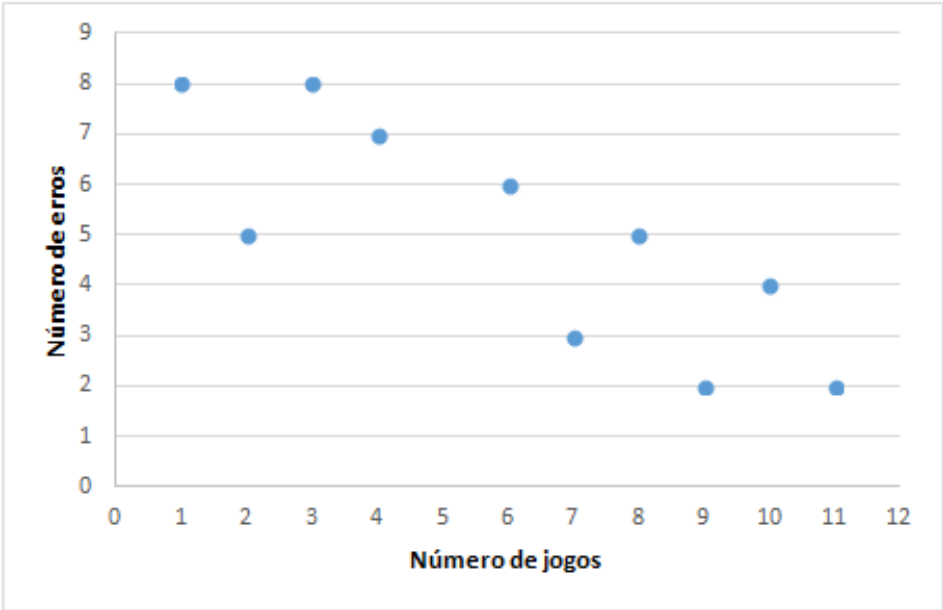



diagrama de dispersão

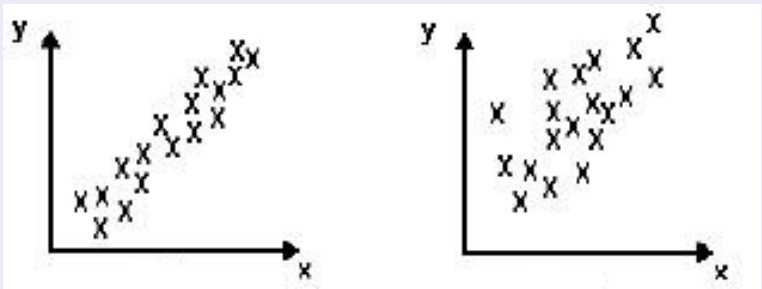
O diagrama de dispersão sugere a existência de uma relação linear entre as duas variáveis em estudo, isto é, uma relação que se traduz geometricamente através de uma reta. É possível "imaginar" uma reta que passa pela "nuvem" de pontos.

# Interpretação do Diagrama de Dispersão

## Sinal

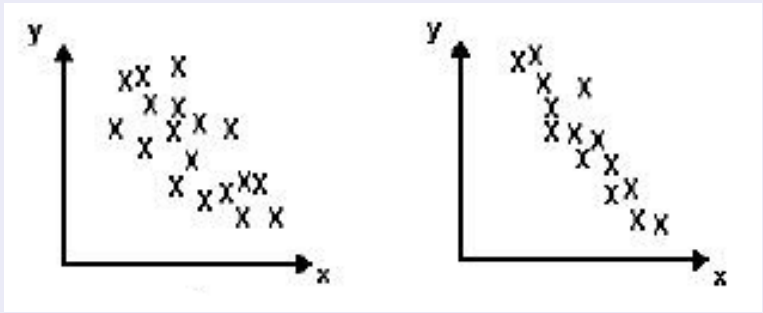
- **Correlação linear positiva**

→ A maiores valores de uma variável tendem a corresponder maiores valores da outra variável.

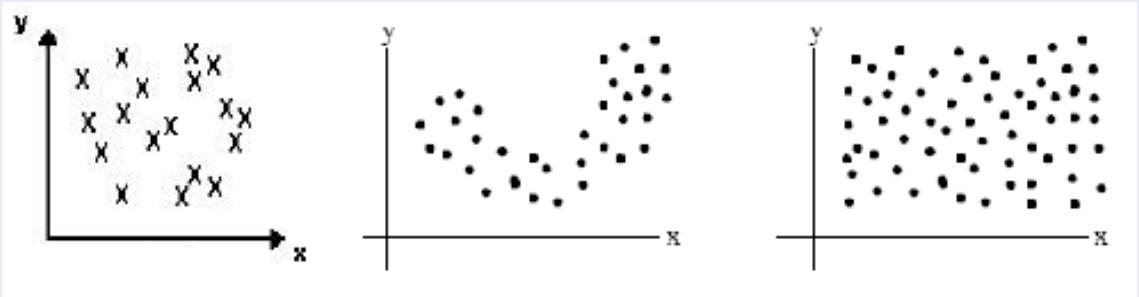



- **Correlação linear negativa**

→ A maiores valores de uma variável tendem a corresponder menores valores da outra.




- **Correlação linear nula** → Não existe associação linear entre as variáveis.



## Observação

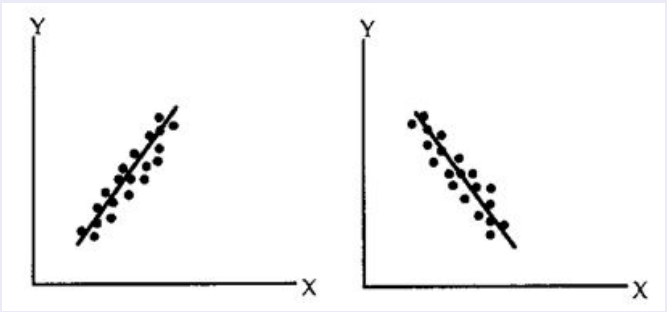
- Quando se diz que as duas variáveis **não estão linearmente correlacionadas**, não significa que as variáveis não estejam correlacionadas, apenas significa que a correlação não é linear. Neste caso poderá existir outro tipo de correlação.

- Em alguns casos, pode não existir uma relação linear entre as variáveis, mas sim quadrática, cúbica, exponencial, logarítmica,...

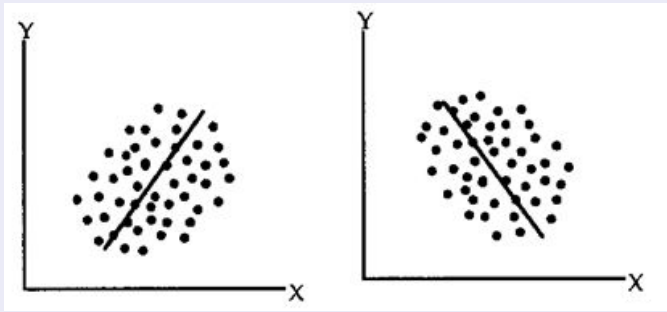
# Interpretação do Diagrama de Dispersão

## Intensidade

- A correlação é tanto mais forte quanto menor for a dispersão dos pontos em torno da linha reta, isto é, quanto mais concentrados os pontos estiverem em torno dessa reta.

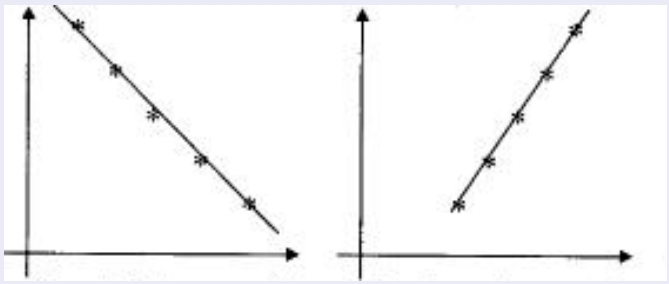



correlação linear forte




correlação linear fraca

- A correlação diz-se **perfeita** se todos os pontos coincidirem com a reta.



correlação linear perfeita

## Exemplo 1

Um psicólogo efetuou uma pesquisa com o objetivo de analisar a forma como as crianças aprendem um determinado jogo. Das diversas variáveis observadas foram registados o número de jogos efetuados $(X)$ e o número de erros realizados no jogo $(Y)$:

| Número de jogos | Número de erros |
|---|---|
| x | y |
| 1 | 8 |
| 2 | 5 |
| 3 | 8 |
| 4 | 7 |
| 6 | 6 |
| 7 | 3 |
| 8 | 5 |
| 9 | 2 |
| 10 | 4 |
| 11 | 2 |

tabela de dados

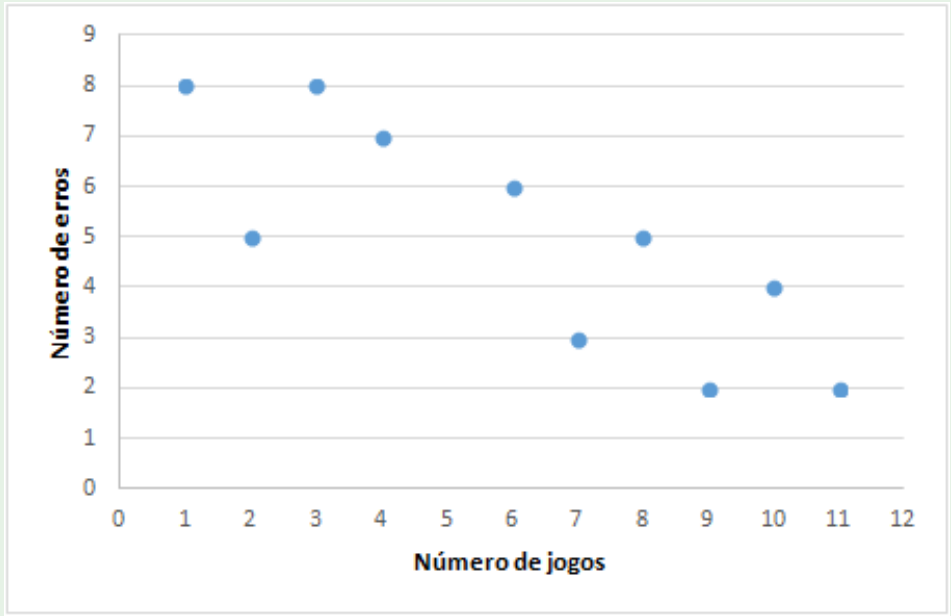



diagrama de dispersão

2. O que pode dizer sobre a tendência linear que se observa no diagrama de dispersão?

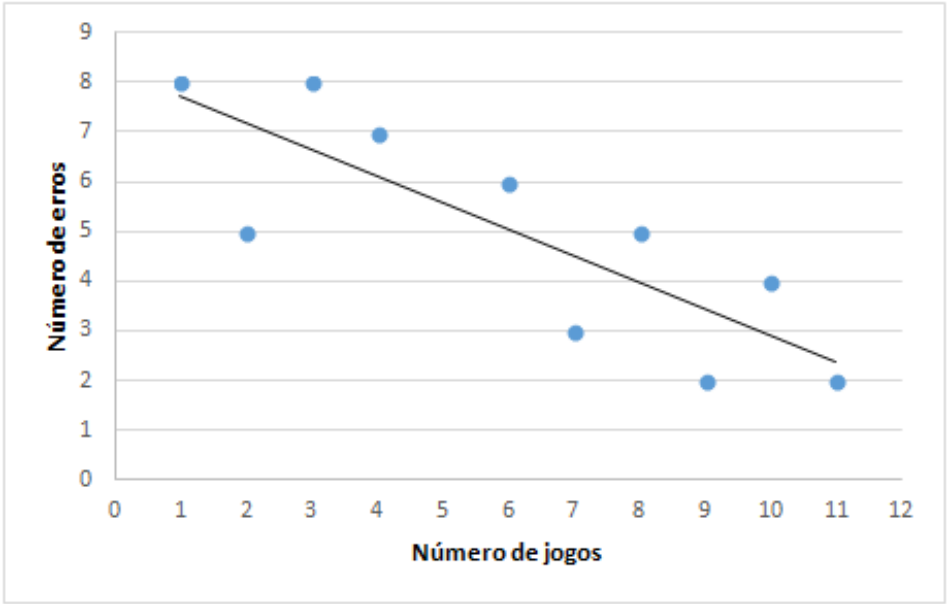



diagrama de dispersão

- O gráfico mostra associação (ou correlação), de sentido contrário, entre o número de jogos e o número de erros. Quando o número de jogos aumenta, o número de erros diminui. Diz-se que as duas variáveis estão **negativamente correlacionadas**.
- Como os pontos não se afastam muito da reta, a **correlação negativa parece ser forte** mas não é perfeita.

# Metodologia

1. *Verificar a existência ou não de uma relação linear*

## Coeficiente de Correlação Linear

- A análise gráfica da relação entre variáveis é importante, mas os olhos nem sempre são um bom juiz da intensidade de uma relação linear. Deve-se, então, utilizar uma medida numérica para complementar a análise gráfica.
- A medida que se utiliza com mais frequência para medir o grau da relação linear entre as variáveis $X$ e $Y$ é o coeficiente de correlação linear, também chamado de **coeficiente de correlação linear de Pearson**.
- O coeficiente de correlação linear mede a maior ou menor intensidade com que as variáveis se associam, quer positiva, quer negativamente. Isto é, é uma medida que avalia o quanto a "nuvem de pontos" no diagrama de dispersão aproxima-se de uma reta.
- Representa-se por $r_{xy}$, é um estimador do coeficiente de correlação linear populacional, $\rho_{XY}$.

# Coeficiente de Correlação Linear

O **coeficiente de correlação linear** (também chamado de **coeficiente de correlação linear empírico** ou **amostral**) calcula-se da seguinte forma:

$$r_{xy} = \frac{\text{covariância}_{xy}}{\sqrt{\text{variância}_x \times \text{variância}_y}} = \frac{s_{xy}}{\sqrt{s_x^2 \cdot s_y^2}} = \frac{s_{xy}}{s_x s_y} =$$

$$= \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\left(\sum_{i=1}^{n} (x_i - \overline{x})^2\right) \times \left(\sum_{i=1}^{n} (y_i - \overline{y})^2\right)}} = \frac{\sum_{i=1}^{n} x_i y_i - n\overline{x}\,\overline{y}}{\sqrt{\left(\sum_{i=1}^{n} x_i^2 - n\overline{x}^2\right) \times \left(\sum_{i=1}^{n} y_i^2 - n\overline{y}^2\right)}}$$

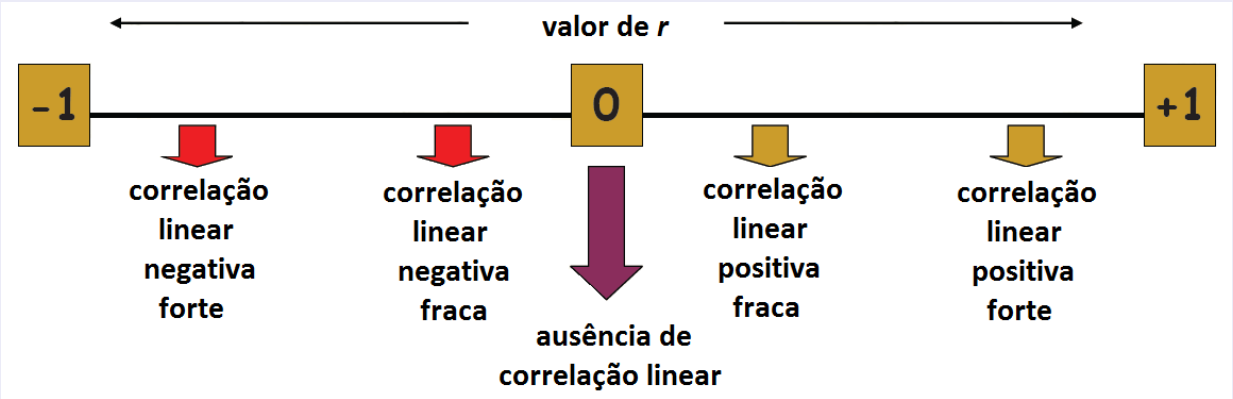
- $s_x^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})^2 = \frac{1}{n-1} \left(\sum_{i=1}^{n} x_i^2 - n\overline{x}^2\right)$ (variância amostral da variável $X$);
- $s_y^2 = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \overline{y})^2 = \frac{1}{n-1} \left(\sum_{i=1}^{n} y_i^2 - n\overline{y}^2\right)$ (variância amostral da variável $Y$);
- $s_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y}) = \frac{1}{n-1} \left(\sum_{i=1}^{n} x_i y_i - n\overline{x}\,\overline{y}\right)$ (covariância amostral entre $X$ e $Y$).

# Coeficiente de Correlação Linear

$$-1 \leq r_{XY} \leq 1$$

- $r_{XY} > 0 \rightarrow$ significa que a relação entre os valores de $x$ e os de $y$ é do mesmo sentido, isto é, a valores grandes de $x$ correspondem, de um modo geral, valores grandes de $y$ e vice-versa $\rightarrow$ **correlação linear positiva**.
- $r_{XY} < 0 \rightarrow$ a relação entre os valores de $x$ e os de $y$ é de sentido contrário, o que significa que a valores grandes de $x$, correspondem, de um modo geral, valores pequenos de $y$ e vice-versa $\rightarrow$ **correlação linear negativa**.
- $r_{XY} = 0 \rightarrow$ significa que não existe relação linear entre os valores de $x$ e os de $y$ (mas pode existir outro tipo de relação) $\rightarrow$ **não há correlação linear**.

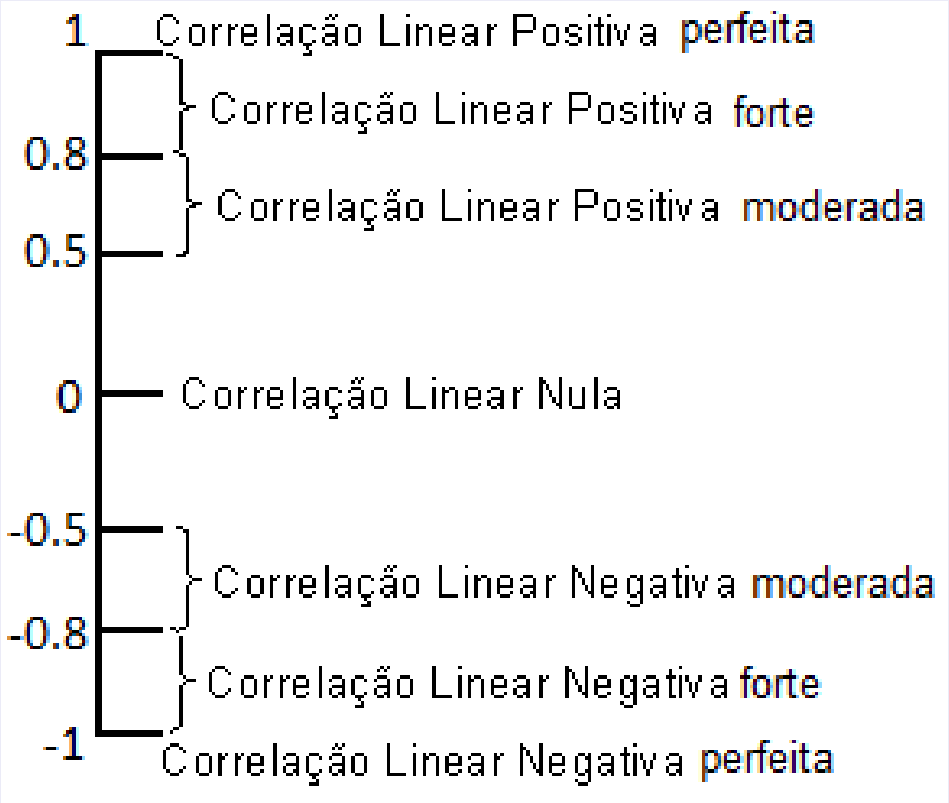
Quanto maior for $|r_{XY}|$, mais forte é a relação linear entre $X$ e $Y$.



# Coeficiente de Correlação Linear

- É possível classificar a intensidade da correlação, analisando a proximidade do coeficiente de correlação linear em relação aos valores 1 e -1.

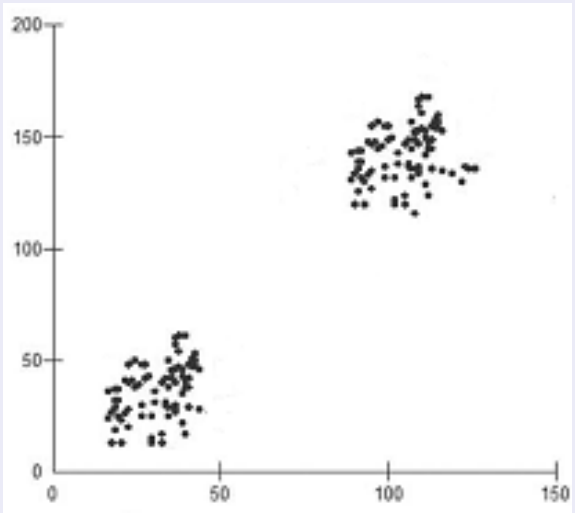
  Tabela indicativa:



# Coeficiente de Correlação Linear

## Observação

- A "confirmação" da existência de um "bom" coeficiente de correlação linear empírico entre $X$ e $Y$ deve ser sempre acompanhado pelo diagrama de dispersão.
- Por exemplo, no seguinte diagrama de dispersão pode-se observar que o modelo linear não é adequado, mas se calcular o coeficiente de correlação linear irá obter um valor próximo de 1.



## Exemplo 1

Um psicólogo efetuou uma pesquisa com o objetivo de analisar a forma como as crianças aprendem um determinado jogo. Das diversas variáveis observadas foram registados o número de jogos efetuados $(X)$ e o número de erros realizados no jogo $(Y)$:

| Número de jogos | Número de erros |
|---|---|
| x | y |
| 1 | 8 |
| 2 | 5 |
| 3 | 8 |
| 4 | 7 |
| 6 | 6 |
| 7 | 3 |
| 8 | 5 |
| 9 | 2 |
| 10 | 4 |
| 11 | 2 |

tabela de dados

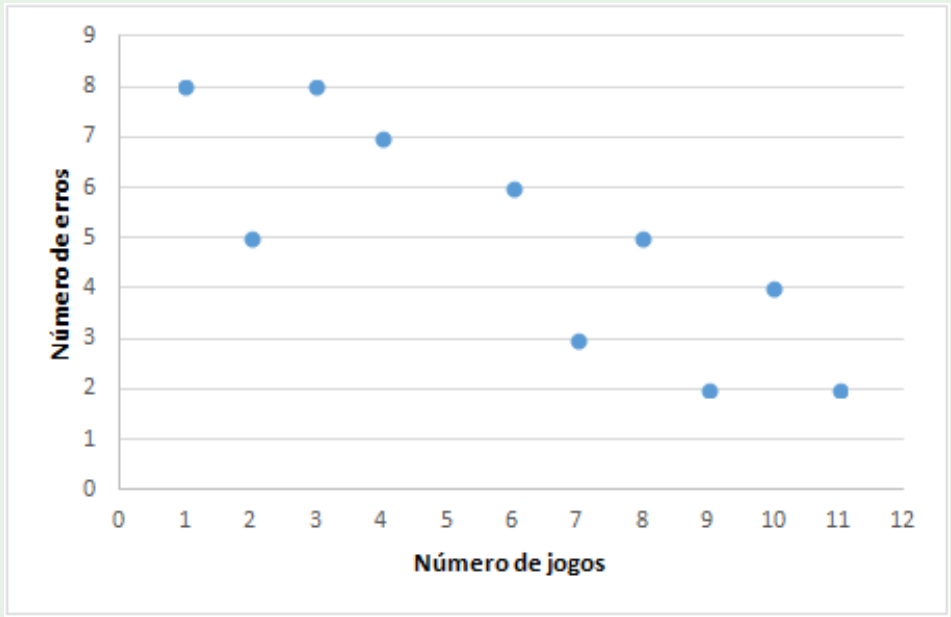



diagrama de dispersão

3. Determine o coeficiente de correlação linear.

coeficiente de correlação linear:

$$r_{XY} = \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\left(\sum_{i=1}^{n} (x_i - \overline{x})^2\right) \times \left(\sum_{i=1}^{n} (y_i - \overline{y})^2\right)}} \qquad \text{(uma possibilidade)}$$

$$\overline{x} = \frac{\sum_{i=1}^{10} x_i}{10} = \frac{61}{10} = 6.1 \qquad \overline{y} = \frac{\sum_{i=1}^{10} y_i}{10} = \frac{50}{10} = 5$$

$$r_{XY} = \frac{\sum_{i=1}^{10} (x_i - 6.1)(y_i - 5)}{\sqrt{\left(\sum_{i=1}^{10} (x_i - 6.1)^2\right) \times \left(\sum_{i=1}^{10} (y_i - 5)^2\right)}} = \frac{-58}{\sqrt{108.9 \times 46}} = -0.819$$

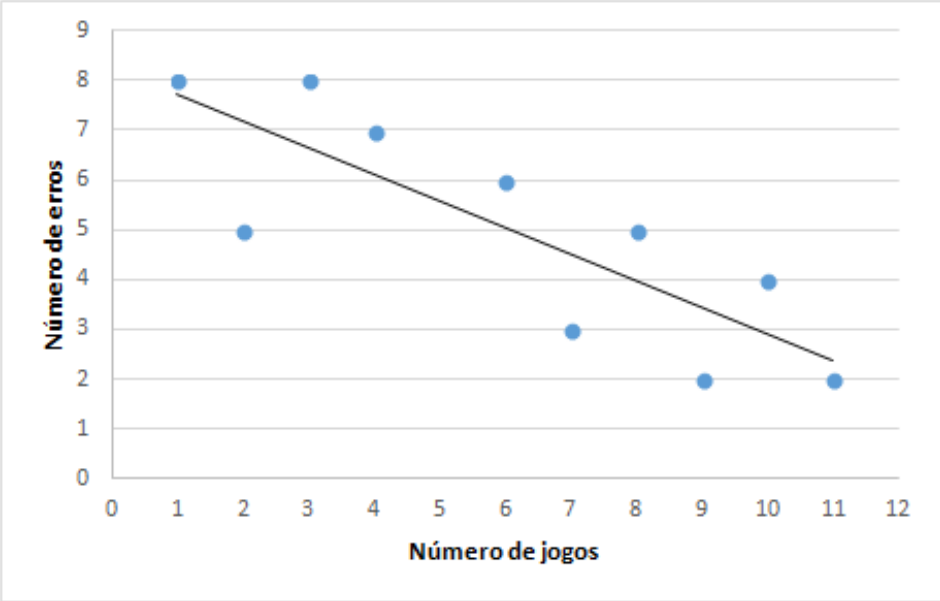



diagrama de dispersão

$$r_{XY} = -0.819$$

O coeficiente de correlação linear permite confirmar o que vimos no diagrama de dispersão, a **correlação linear** é **negativa** $(r_{XY} < 0)$ e **forte** $(-1 < r_{XY} < -0.8)$.

# Correlação

## Observação

- Correlação não significa Causalidade.
- A associação não deve ser interpretada como causa - efeito. Pode, eventualmente, haver outras variáveis, com que não estamos a contar, que contribuam para a associação linear observada.

# Metodologia

2. *Determinar a reta de regressão*

- Quando a correlação entre duas variáveis é elevada, se conhecermos o valor de uma das variáveis podemos ter uma ideia do valor que a outra irá tomar. Diz-se que podemos inferir o valor da outra variável.
- Intuitivamente, é a reta que passa através da nuvem de pontos e a divide em dois grupos semelhantes.
- A reta de regressão passa pelo ponto cujas coordenadas são, respetivamente, as médias da primeira e da segunda variáveis, isto é, o centro de gravidade da nuvem de pontos (ponto de coordenadas $(\overline{x}, \overline{y})$).

# Reta de Regressão

- A reta de regressão é o modelo matemático que resume os valores das amostras da seguinte forma

$$\widehat{y} = a + bx$$

  - a $X$ chama-se **variável independente**, explicativa ou explanatória;
  - a $Y$ chama-se **variável dependente**, explicada ou resposta.

**Atenção:** As conclusões que se tiram do diagrama de dispersão e do coeficiente de correlação linear não é alterado se trocamos as variáveis $X$ e $Y$, isto é, a existência ou não da relação linear não depende de qual variável é considerada independente. No entanto, o modelo matemático será alterado pois depende da variável que é definida como independente.

# Reta de Regressão

$$\widehat{y} = a + bx$$

- $a$ representa a ordenada na origem, isto é, indica o valor de $y$ que se espera observar quando $x = 0$ (o "local" onde a reta corta o eixo dos $yy$);
- $b$ representa o declive, isto é, a inclinação da reta. O seu valor indica em que medida $y$ muda em função de $x$, refletindo a correlação existente entre as variáveis:
  - se $b$ for positivo, a relação entre $X$ e $Y$ é positiva $\rightarrow$ **correlação linear positiva**;
  - se $b$ for negativo, a relação entre $X$ e $Y$ é negativa $\rightarrow$ **correlação linear negativa**;
- **Interpretação** de $b$: para cada aumento de uma unidade em $x$, temos um aumento médio de $b$ unidades em $y$.

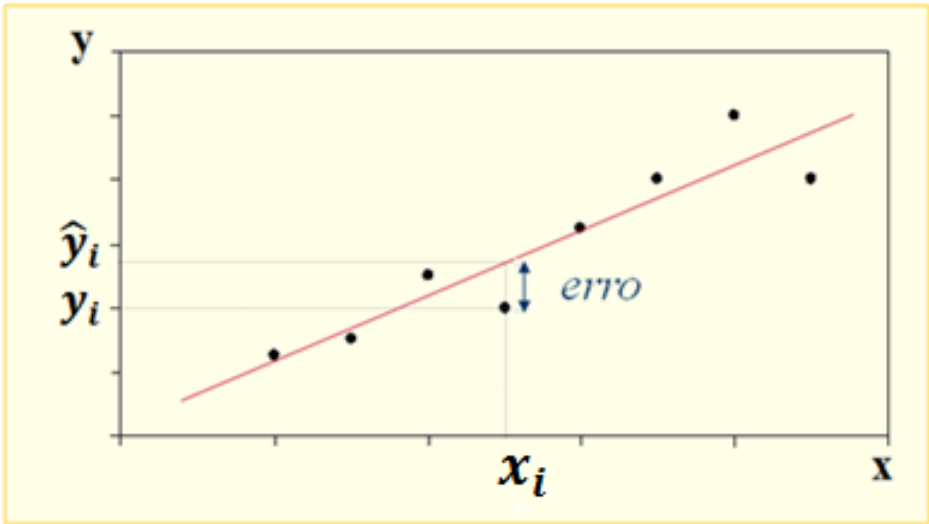
# Determinar a reta de regressão: $\widehat{y} = a + bx$

**Método dos Mínimos Quadrados** - É um dos métodos mais conhecidos que permite ajustar uma reta a uma conjunto de dados. Consiste em determinar a reta

$$\widehat{y}_i = a + bx_i, \qquad i = 1, \ldots, n$$

que minimiza a soma dos quadrados dos desvios ou erros $(e_i)$ entre os verdadeiros valores observados das ordenadas $(y_i)$ e os obtidos a partir da reta a ajustar $(\widehat{y}_i)$:

$$\min \left\{ \sum_{i=1}^{n} e_i^2 \right\} = \min \left\{ \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2 \right\}$$



# Determinar a reta de regressão: $\widehat{y} = a + bx$

e assim obtém-se a ordenada na origem $(a)$ e o declive da reta $(b)$:

$$\begin{cases} a = \overline{y} - b\overline{x} \\ b = \dfrac{s_{xy}}{s_x^2} = \dfrac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} = \dfrac{\sum_{i=1}^{n} x_i y_i - n\overline{x}\,\overline{y}}{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2} = r_{xy} \times \dfrac{s_y}{s_x} \end{cases}$$

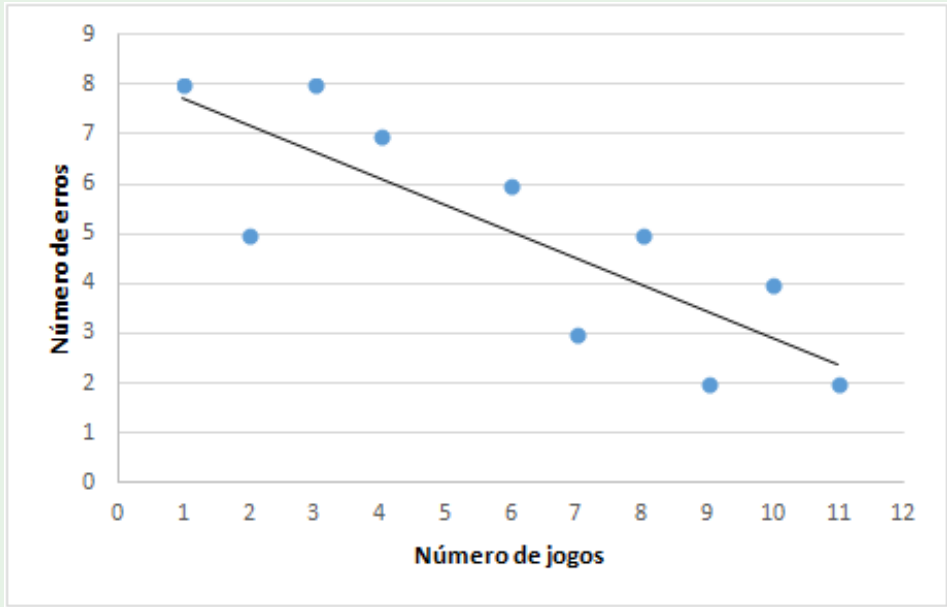
- $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ e $\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ (média amostral);
- $s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$ e $s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2$ (variância amostral);
- $s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i y_i - n\overline{x}\,\overline{y}\right)$ (covariância amostral).
- $r_{xy} = \frac{s_{xy}}{s_x s_y}$ (coeficiente de correlação linear empírico).

## Exemplo 1

Um psicólogo efetuou uma pesquisa com o objetivo de analisar a forma como as crianças aprendem um determinado jogo. Das diversas variáveis observadas foram registados o número de jogos efetuados $(X)$ e o número de erros realizados no jogo $(Y)$:

| Número de jogos | Número de erros |
|---|---|
| x | y |
| 1 | 8 |
| 2 | 5 |
| 3 | 8 |
| 4 | 7 |
| 6 | 6 |
| 7 | 3 |
| 8 | 5 |
| 9 | 2 |
| 10 | 4 |
| 11 | 2 |

tabela de dados




Coeficiente de correlação linear: $r_{XY} = -0.819$

4. Determine a reta de regressão.

Reta de regressão: $\widehat{y} = a + bx$ com $b = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$ e $a = \overline{y} - b\overline{x}$ (uma possibilidade)

$$\overline{x} = \frac{\sum_{i=1}^{10} x_i}{10} = \frac{61}{10} = 6.1 \qquad \overline{y} = \frac{\sum_{i=1}^{10} y_i}{10} = \frac{50}{10} = 5$$

$$b = \frac{\sum_{i=1}^{10} (x_i - 6.1)(y_i - 5)}{\sum_{i=1}^{10} (x_i - 6.1)^2} = \frac{-58}{108.9} = -0.5326$$

$$a = 5 - (-0.5326) \times 6.1 = 8.2489$$

Reta de regressão:

$$\widehat{y} = 8.2489 - 0.5326x$$

# Metodologia

3. *Análise dos Resíduos*

- Uma outra forma de verificar a qualidade do modelo de regressão linear é através dos erros, $e_i$, isto é, das diferenças entre os valores observados $(y_i)$ e os valores ajustados $(\widehat{y}_i)$, aos quais se chama de **resíduos**:

$$\text{resíduos} = e_i = y_i - \widehat{y}_i, \qquad i = 1, \ldots, n.$$

- Uma propriedade importante dos resíduos é o facto da sua soma ser nula,

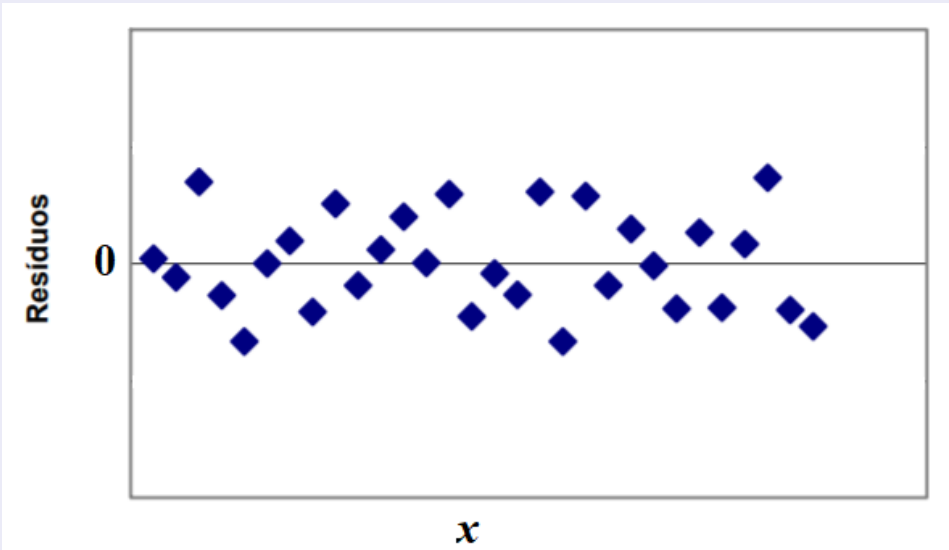
$$\sum_{i=1}^{n} e_i = 0.$$

- Uma forma de analisar os resíduos é através de um diagrama de dispersão onde se representam os pontos $(x_i, e_i)$ (no caso da regressão linear simples é equivalente a representar no diagrama os pontos $(\hat{y}_i, e_i)$), visualizando-se os desvios positivos e negativos acima e abaixo do eixo dos $xx$.

# Metodologia

3. *Análise dos Resíduos*

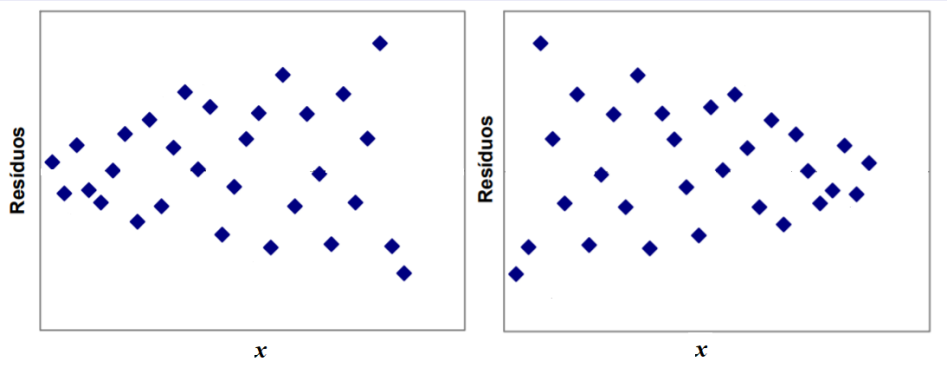
- Se os resíduos **não forem grandes** (concentrados entre -2 e 2 e poucos pontos acima de 3 ou abaixo de -3), com **variância constante** e **não tiverem um padrão bem definido** (dispersos aleatoriamente em torno de zero), é sintoma de que o **modelo ajustado é bom**:



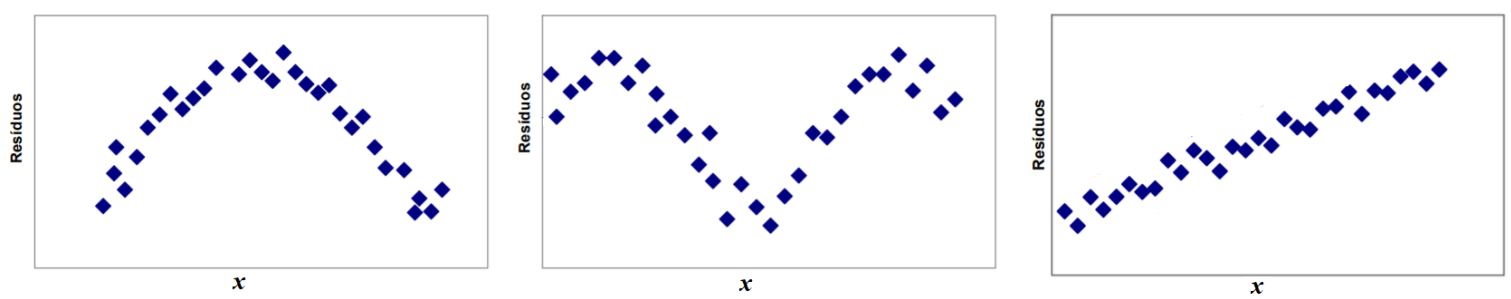
# Metodologia

3. *Análise dos Resíduos*

- Se a dispersão dos resíduos aumentar ou diminuir com os valores da variável independente $x_i$ (ou com os valores estimados da variável dependente $\hat{y}_i$), deve ser posta em causa a hipótese de variância constante dos resíduos:




- Quando os resíduos não se comportam de forma aleatória, ou seja, seguem um padrão, significa que os resíduos não são independentes:

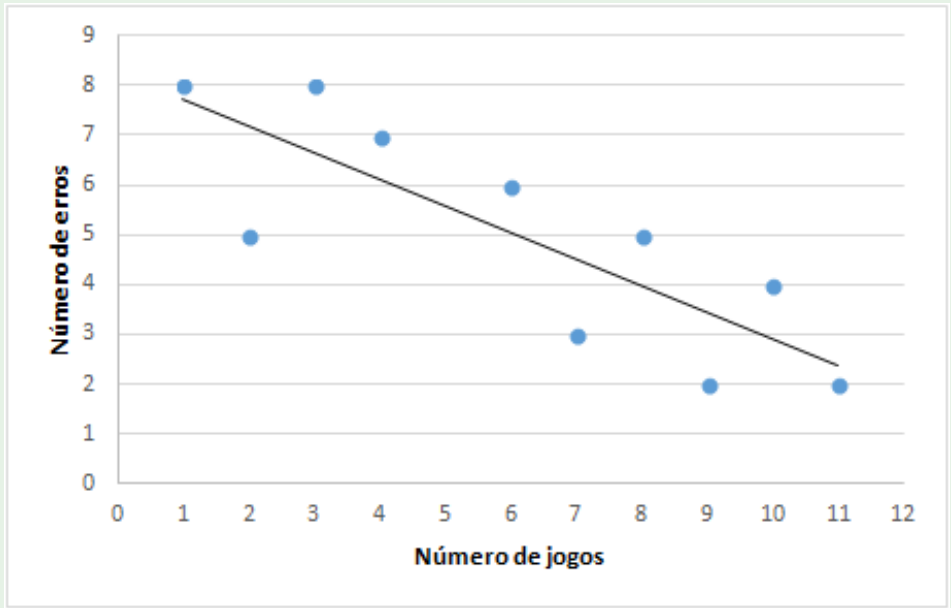


## Exemplo 1

Um psicólogo efetuou uma pesquisa com o objetivo de analisar a forma como as crianças aprendem um determinado jogo. Das diversas variáveis observadas foram registados o número de jogos efetuados $(X)$ e o número de erros realizados no jogo $(Y)$:

| Número de jogos | Número de erros |
|---|---|
| x | y |
| 1 | 8 |
| 2 | 5 |
| 3 | 8 |
| 4 | 7 |
| 6 | 6 |
| 7 | 3 |
| 8 | 5 |
| 9 | 2 |
| 10 | 4 |
| 11 | 2 |

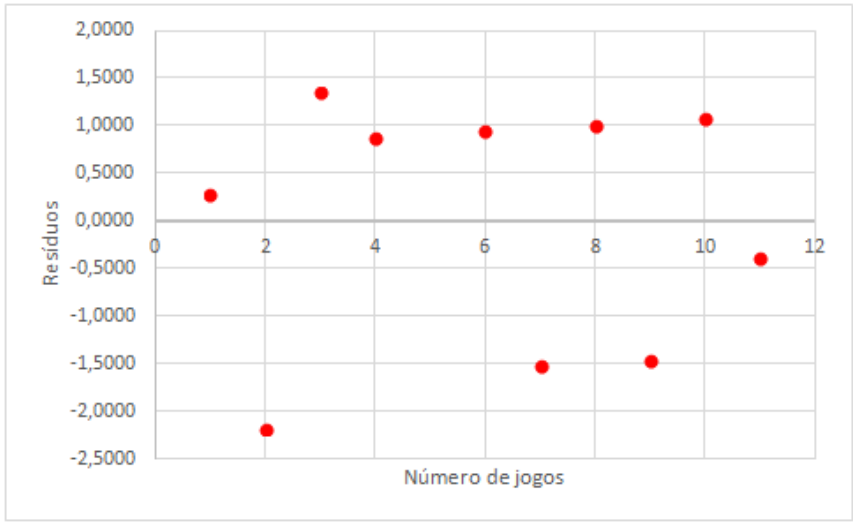
tabela de dados




Coeficiente de correlação linear: $r_{XY} = -0.819$
reta de regressão: $\widehat{y} = 8.2489 - 0.5326x$

5. Analise os resíduos.

| $x_i$ | $y_i$ | $\widehat{y} = 8.2489 - 0.5326x$ | $e_i = y_i - \widehat{y}$ |
|---|---|---|---|
| 1 | 8 | 7.7163 | 0.2837 |
| 2 | 5 | 7.1837 | $-2.1837$ |
| 3 | 8 | 6.6511 | 1.3489 |
| 4 | 7 | 6.1185 | 0.8818 |
| 6 | 6 | 5.0533 | 0.9467 |
| 7 | 3 | 4.5207 | $-1.5207$ |
| 8 | 5 | 3.9881 | 1.0119 |
| 9 | 2 | 3.4555 | $-1.4555$ |
| 10 | 4 | 2.9229 | 1.0771 |
| 11 | 2 | 2.3903 | $-0.3903$ |
| | | | $\sum_{i=1}^{10} e_i \approx 0$ |




Como os resíduos não são grandes e não apresentam um padrão bem definido, o modelo ajustado parece ser adequado.

# Metodologia

4. *Previsão*

- Depois de encontrado o modelo de regressão linear que se adapta aos dados,

$$\widehat{y} = a + bx,$$

é possível efetuar **previsões** para $\mathbf{y}$ com base em valores de $x$.

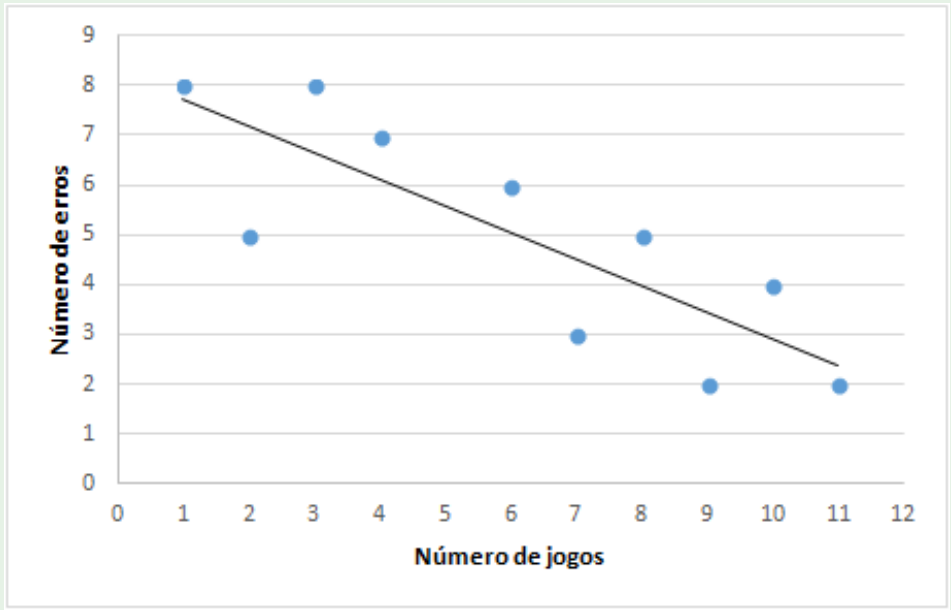
- Só deve ser feita previsão para $y$ com base em valores de $x$ dentro do intervalo analisado ou para valores muito próximos do intervalo analisado. Quando nos afastamos muito, não sabemos se a relação linear ainda se mantém, logo a previsão pode ser absurda.

## Exemplo 1

Um psicólogo efetuou uma pesquisa com o objetivo de analisar a forma como as crianças aprendem um determinado jogo. Das diversas variáveis observadas foram registados o número de jogos efetuados $(X)$ e o número de erros realizados no jogo $(Y)$:

| Número de jogos | Número de erros |
|---|---|
| x | y |
| 1 | 8 |
| 2 | 5 |
| 3 | 8 |
| 4 | 7 |
| 6 | 6 |
| 7 | 3 |
| 8 | 5 |
| 9 | 2 |
| 10 | 4 |
| 11 | 2 |

tabela de dados




Coeficiente de correlação linear: $r_{XY} = -0.819$
reta de regressão: $\widehat{y} = 8.2489 - 0.5326x$

6. Obtenha estimativas para o número de erros que uma criança comete quando joga 5 vezes o jogo e quando joga 50 vezes o jogo.

Uma estimativa para o número de erros que uma criança comete quando

1. joga 5 vezes o jogo:

$$\widehat{y}(5) = 8.2489 - 0.5326 \times 5 = 5.586 \approx 6 \text{ erros}$$

2. joga 50 vezes o jogo:

$$\widehat{y}(50) = 8.2489 - 0.5326 \times 50 = -18.381$$ ABSURDO!

Mesmo que a estimativa para 50 jogos tivesse dado um valor válido, continuaríamos a não ter qualquer confiança nessa previsão, pois o valor 50 encontra-se muito afastado dos valores observados, $[1, 11]$. Embora o modelo tenha sido considerado adequado (diagrama de de dispersão, coeficiente de correlação linear e resíduos, todos levaram a considerar o modelo adequado), nada nos garante que a reta obtida se mantém para valores afastado dos observados.

Uma estimativa para o número de erros que uma criança comete quando

1. joga 5 vezes o jogo:

$$\widehat{y}(5) = 8.2489 - 0.5326 \times 5 = 5.586 \approx 6 \text{ erros}$$

2. joga 50 vezes o jogo:

$$\widehat{y}(50) = 8.2489 - 0.5326 \times 50 = -18.381$$

$$\widehat{y}(50) = 8.2489 - 0.5326 \times 50 = -18.381 \qquad \text{ABSURDO!}$$

Mesmo que a estimativa para 50 jogos tivesse dado um valor válido, continuaríamos a não ter qualquer confiança nessa previsão, pois o valor 50 encontra-se muito afastado dos valores observados, $[1, 11]$. Embora o modelo tenha sido considerado adequado (diagrama de de dispersão, coeficiente de correlação linear e resíduos, todos levaram a considerar o modelo adequado), nada nos garante que a reta obtida se mantém para valores afastado dos observados.

Uma estimativa para o número de erros que uma criança comete quando

1. joga 5 vezes o jogo:

$$\widehat{y}(5) = 8.2489 - 0.5326 \times 5 = 5.586 \approx 6 \text{ erros}$$

2. joga 50 vezes o jogo: valor muito afastado dos valores observados

$$\widehat{y}(50) = 8.2489 - 0.5326 \times 50 = -18.381$$ ABSURDO!

Mesmo que a estimativa para 50 jogos tivesse dado um valor válido, continuaríamos a não ter qualquer confiança nessa previsão, pois o valor 50 encontra-se muito afastado dos valores observados, $[1, 11]$. Embora o modelo tenha sido considerado adequado (diagrama de de dispersão, coeficiente de correlação linear e resíduos, todos levaram a considerar o modelo adequado), nada nos garante que a reta obtida se mantém para valores afastado dos observados.

## Observação

Suponha que calculou o seguinte modelo de regressão

$$\widehat{y} = a + bx,$$

e pretende efetuar **previsões** para $\mathbf{x}$ com base em valores de $y$.

Com a reta de regressão anterior apenas faz sentido estimar valores de $y$. Como agora pretende estimar valores de $x$, então significa que $\mathbf{X}$ passa a ser a **variável dependente** e $\mathbf{Y}$ a **variável independente**, sendo necessário calcular a reta de regressão correspondente. Ou seja, ir a todas as fórmulas apresentadas anteriormente e onde está $x$ colocar $y$ e onde está $y$ colocar $x$ e assim obtém

$$\widehat{x} = a^* + b^* y$$

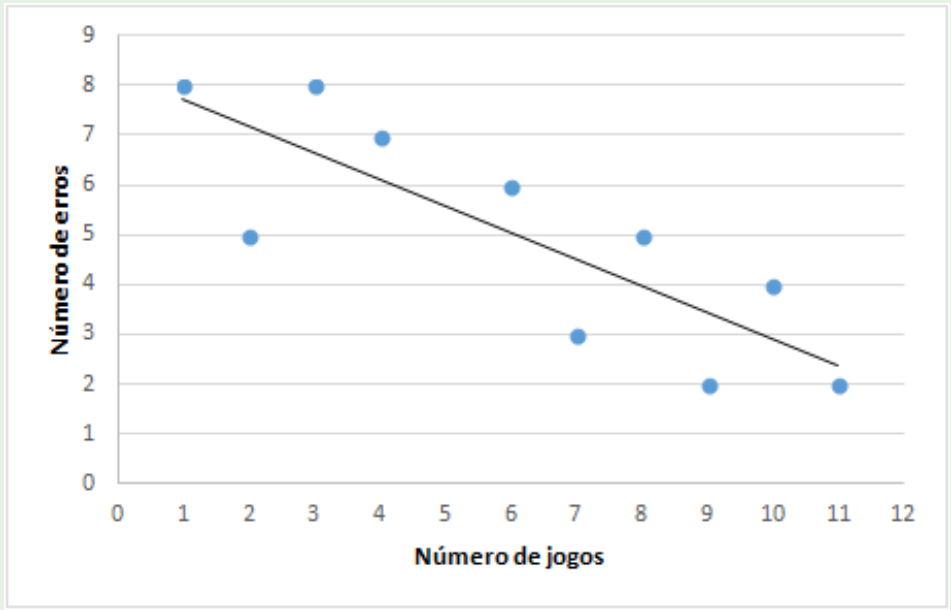
com $a^* = \overline{x} - b^* \overline{y}$ e $b^* = r_{XY} \times \frac{s_x}{s_y}$

## Exemplo 1

Um psicólogo efetuou uma pesquisa com o objetivo de analisar a forma como as crianças aprendem um determinado jogo. Das diversas variáveis observadas foram registados o número de jogos efetuados $(X)$ e o número de erros realizados no jogo $(Y)$:

| Número de jogos | Número de erros |
|---|---|
| x | y |
| 1 | 8 |
| 2 | 5 |
| 3 | 8 |
| 4 | 7 |
| 6 | 6 |
| 7 | 3 |
| 8 | 5 |
| 9 | 2 |
| 10 | 4 |
| 11 | 2 |

tabela de dados




Coeficiente de correlação linear: $r_{XY} = -0.819$
reta de regressão: $\widehat{y} = 8.2489 - 0.5326x$

7. Suponha que a criança errou apenas 1 vez e pretende obter uma estimativa do número de jogos que a criança fez. Como deve fazer?

Como agora pretende-se estimar valores de $x$, então significa que $X$ (número de jogos) passa a ser a variável dependente e $Y$ (número de erros) a variável independente, sendo necessário calcular a reta de regressão correspondente: $\widehat{x} = a^* + b^* y$.

Ou seja, ir a todas as fórmulas apresentadas anteriormente e onde está $x$ colocar $y$ e onde está $y$ colocar $x$:

$$b^* = \frac{\sum_{i=1}^{n} (y_i - \overline{y})(x_i - \overline{x})}{\sum_{i=1}^{n} (y_i - \overline{y})^2} \quad \text{e} \quad a^* = \overline{x} - b^* \overline{y} \quad \text{(uma possibilidade)}$$

$$b^* = \frac{\sum_{i=1}^{10} (y_i - 5)(x_i - 6.1)}{\sum_{i=1}^{10} (y_i - 5)^2} = \frac{-58}{46} = -1.2609$$

$$a^* = 6.1 - (-1.2609) \times 5 = 12.4045$$

Assim obtém-se a reta de regressão

$$\widehat{x} = 12.4045 - 1.2609y.$$

Voltando à questão:

$$\widehat{x}(1) = 12.4045 - 1.2609 \times 1 = 11.144 \approx 11 \text{ jogos}$$

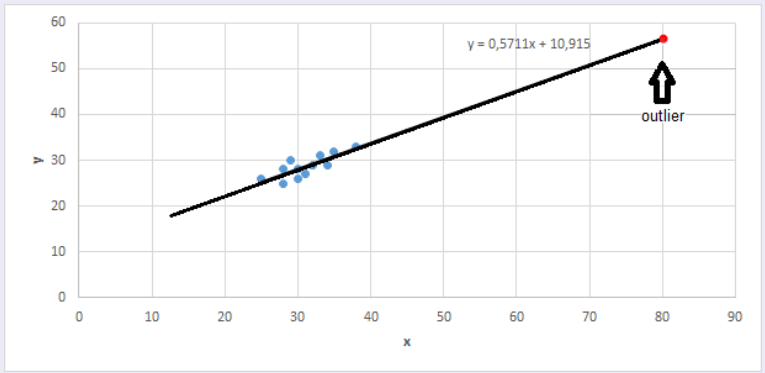
# "Outliers" ou observações discordantes

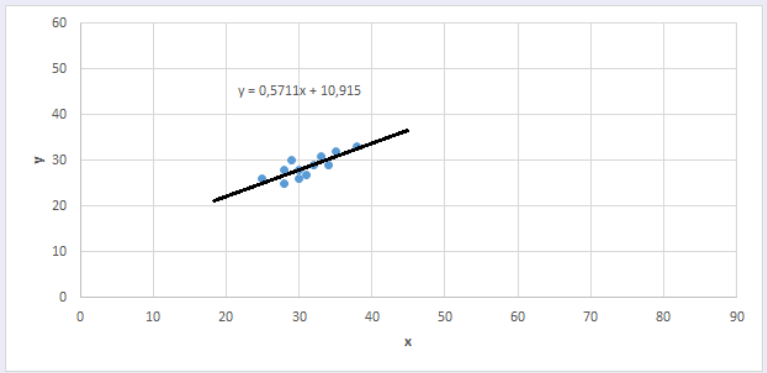
- Através das análise do Diagrama de Dispersão por vezes surgem observações que se destacam das restantes, essas observações são chamadas de "Outliers" ou observações discordantes.

- A identificação e a interpretação de "outliers" são tarefas complexas e altamente subjetivas A explicação para a existência de valores com um comportamento que se afasta nitidamente do da grande maioria dos restantes valores pode ser devido:
  - a erros humanos - essas observações devem ser corrigidas ou eliminadas do estudo;
  - à natureza do fenómeno - essas observações devem ser analisadas com cuidado.

# "Outliers" ou observações discordantes

- Observações deste tipo podem dividir-se em duas classes:
  - **"outliers" não influentes** - a sua existência não altera o modelo linear ajustado




  - **"outliers" influentes** - a sua existência altera o modelo linear ajustado