# Correlação e Regressão aplicado a IMÓVEIS

Prof. Antonio Estanislau Sanches

2018

# Objetivo

Apresentar as ferramentas de Análise Estatística disponíveis no Microsoft Excel aplicadas à Engenharia de Avaliações, utilizando-as em exemplos de resolução de problemas de regressão linear simples e regressão linear múltipla

Pressupõe-se prévios conhecimentos de estatística, tais como: distribuições bilaterais e unilaterais, colinearidade, análise de resíduos, testes de significância, análise de variância e regressão linear

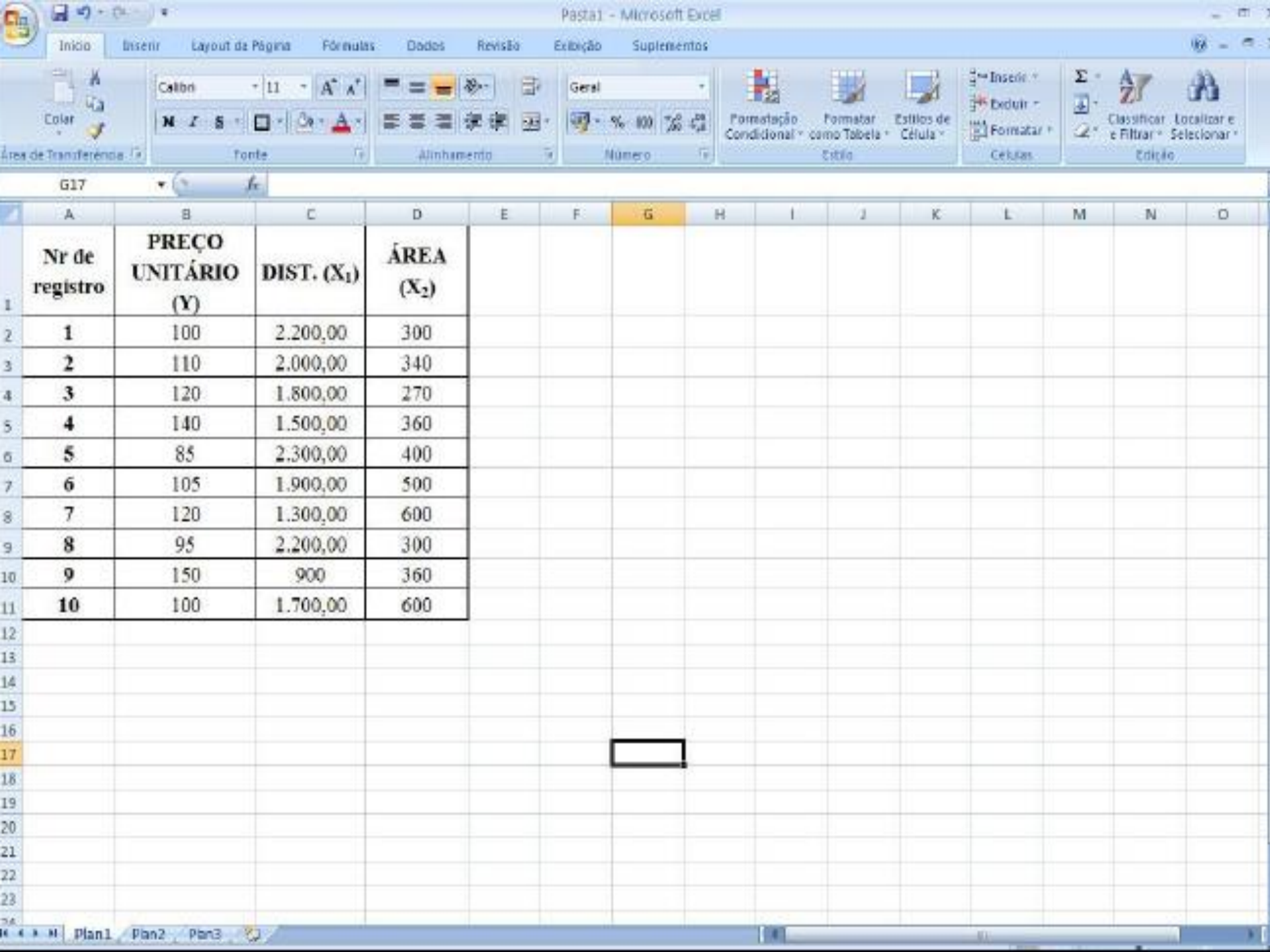
Consideremos o exemplo onde estamos interessados em avaliar um lote de 300 $m^2$ de área, situado a uma distância de cerca de 2.100 m de um ponto valorizante. Os atributos de diferenciação levantados compreendem a localização do lote, através da distância do mesmo em metros ao referido ponto e a área do lote. As características do imóvel avaliando bem como a dos imóveis da amostra estão apresentados na tabela abaixo:

| REGISTRO N.º | VARIÁVEL DEPENDENTE PREÇO UNITÁRIO (Y) | VARIÁVEIS INDEPENDENTES OU EXPLICATIVAS | |
|---|---|---|---|
| | | DIST. ($X_1$) | ÁREA ($X_2$) |
| 1 | 100,00 | 2.200,00 | 300,00 |
| 2 | 110,00 | 2.000,00 | 340,00 |
| 3 | 120,00 | 1.800,00 | 270,00 |
| 4 | 140,00 | 1.500,00 | 360,00 |
| 5 | 85,00 | 2.300,00 | 400,00 |
| 6 | 105,00 | 1.900,00 | 500,00 |
| 7 | 120,00 | 1.300,00 | 600,00 |
| 8 | 95,00 | 2.200,00 | 300,00 |
| 9 | 150,00 | 900,00 | 360,00 |
| 10 | 100,00 | 1.700,00 | 600,00 |

Inicialmente poderíamos supor que o atributo distância ao ponto valorizante não seja influenciante no valor do lote e que a influência da área seja diretamente proporcional a esse valor, o que possibilitaria a resolução do problema por estatística descritiva.

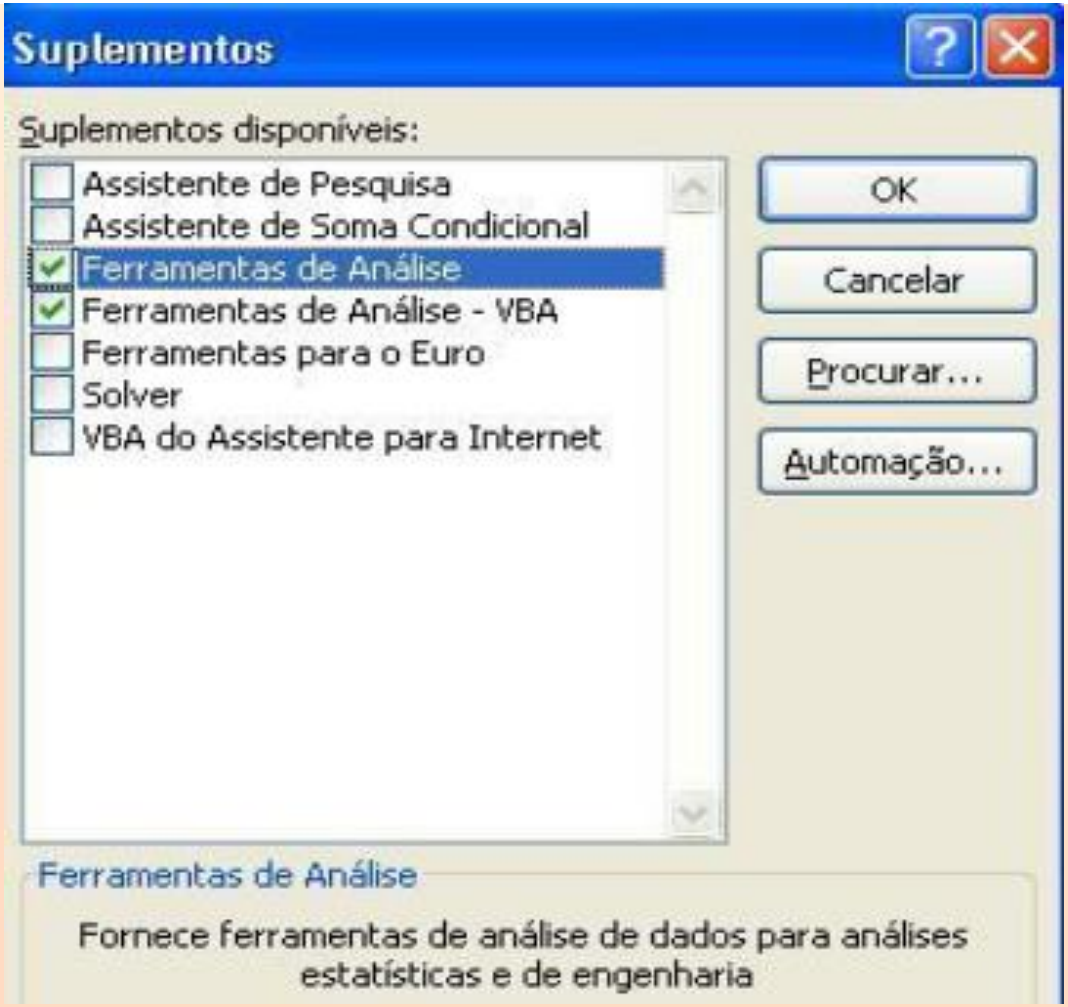
Entretanto, procuraremos levar em consideração o atributo distância ao ponto valorizante na formação dos preços, obtendo uma equação de regressão linear simples, relacionando o preço unitário (PU) com a distância (DIST) do tipo: $\hat{Y}_i = B_0 + B_1 X_1$, que terá o seguinte aspecto:

$$\hat{PU} = B_0 + B_1 * DIST$$



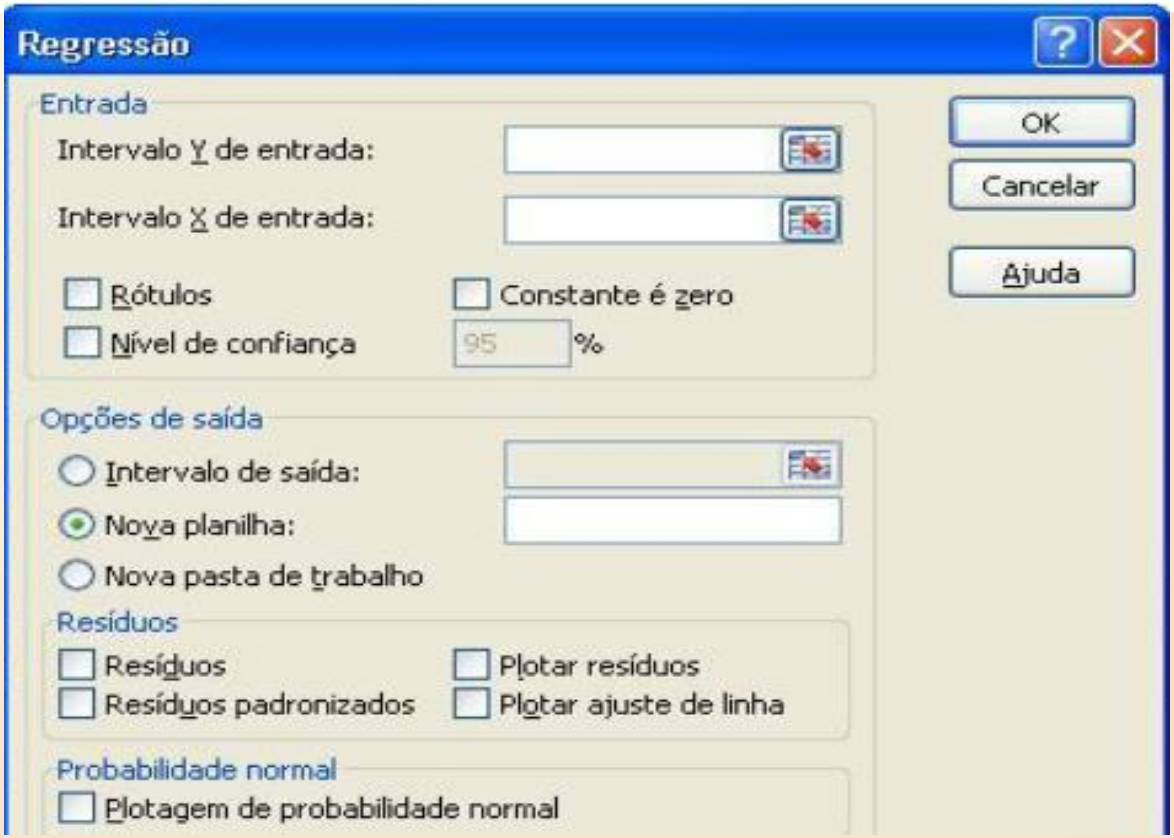
| Nr de registro | PREÇO UNITÁRIO (Y) | DIST. ($X_1$) | ÁREA ($X_2$) |
|---|---|---|---|
| 1 | 100 | 2.200,00 | 300 |
| 2 | 110 | 2.000,00 | 340 |
| 3 | 120 | 1.800,00 | 270 |
| 4 | 140 | 1.500,00 | 360 |
| 5 | 85 | 2.300,00 | 400 |
| 6 | 105 | 1.900,00 | 500 |
| 7 | 120 | 1.300,00 | 600 |
| 8 | 95 | 2.200,00 | 300 |
| 9 | 150 | 900 | 360 |
| 10 | 100 | 1.700,00 | 600 |

Antes de iniciar a Análise de Dados é necessário ativar os suplementos do Excel para que as ferramentas de análise estejam disponíveis, para ativá-las, clique com o botão direito sobre a barra de ferramentas, em seguida clique em **"Personalizar barra de tarefas de acesso rápido"**, clique em **"Suplementos"**, na opção Gerenciar, selecione a opção **"Suplementos do Excel"** e em seguida clique em **"IR"**, no menu suplementos ative os itens "**Ferramentas de Análise"** e **"Ferramentas de Análise – VBA"** , como mostrado na **Figura** em seguida clique em **"OK".**

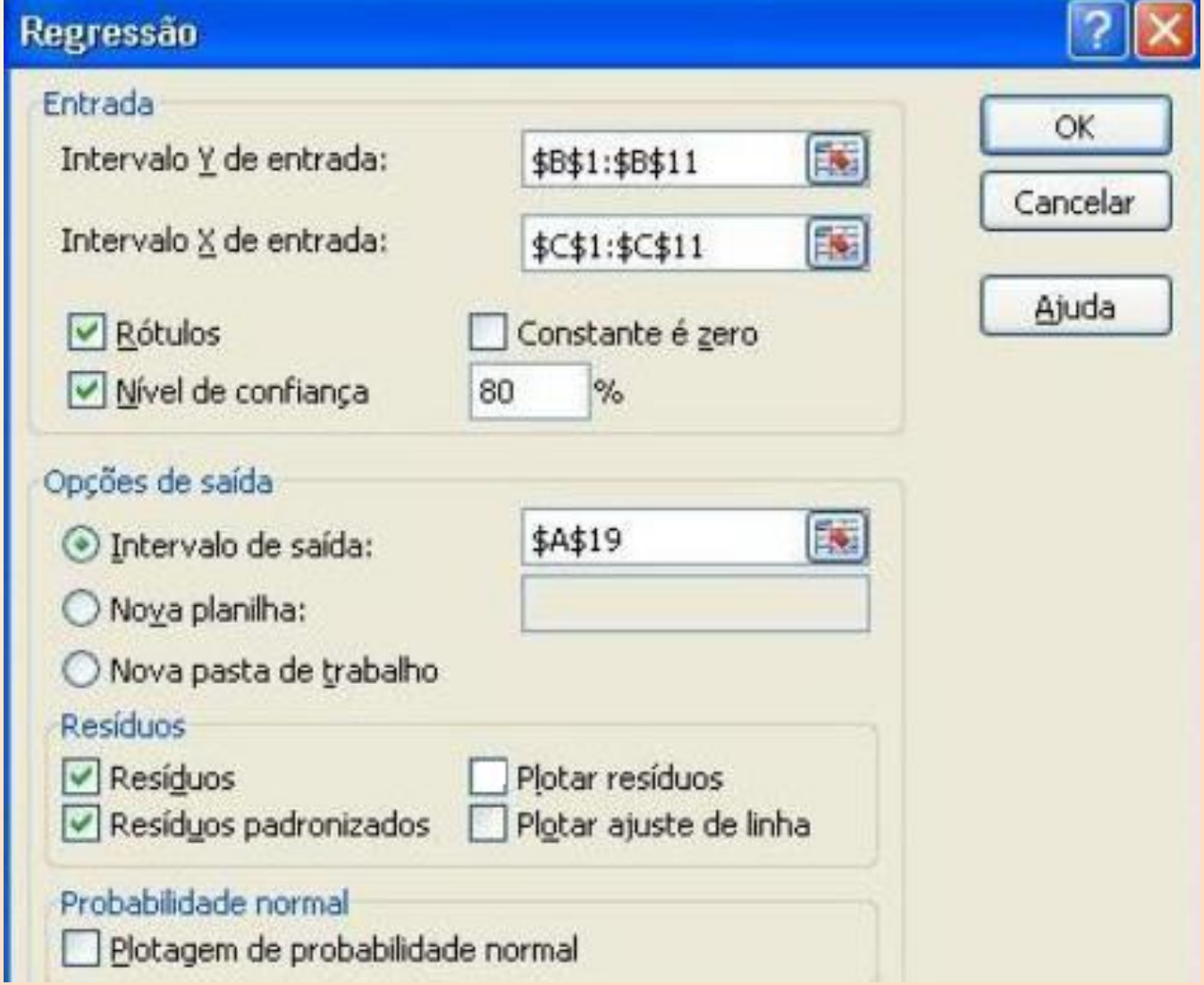


No menu **Dados** clique em **"Análise de Dados"**.

Na caixa de diálogo **Análise de Dados** dê um duplo clique em **"Regressão"**.




No campo **Intervalo Y de entrada:** deve ser digitado o intervalo que contém a variável dependente Preço Unitário (Y=PU), na região B1:B11 e no **Intervalo X de entrada:** deve ser digitado o intervalo que contém a variável independente distância (X=Dist), na região C1:C11. Marque a opção **Rótulos**, pois foram incluídos junto com os dados. Marque as opões **Resíduos** e **Resíduos Padronizados**. Clique no botão **Intervalo de Saída**, marcando a célula A19 e clique em **OK**.




**PS:** não se esqueça de alterar o Nível de confiança, de 95% para 80%, valor recomendado pela Norma NBR 14.653-2/2011..

RESUMO DOS RESULTADOS

| *Estatística de regressão* | |
|---|---|
| R múltiplo | 0,88054 |
| R-Quadrado | 0,77536 |
| R-quadrado ajustado | 0,74728 |
| Erro padrão | 10,21017 |
| Observações | 10 |

= Coef. Determinação

**Tabela 1.3**

**R-múltiplo:** Correlação entre as variáveis independentes e a variável dependente.

**R-quadrado:** Poder de explicação do modelo de regressão, no exemplo 77,5% da variabilidade dos preços é explicado pelo modelo adotado

**R-quadrado ajustado:** Idem ao R-quadrado, porém ajustado levanto em conta o numero de variáveis independentes

**Erro padrão**: É o desvio padrão do modelo, dado pela raiz quadrada da variância.

ANOVA

| | *gl* | *SQ* | *MQ* | *F* | *F de significação* |
|---|---|---|---|---|---|
| Regressão | 1 | 2878,51914 | 2878,51914 | 27,61233 | 0,00077 |
| Resíduo | 8 | 833,98085 | 104,24760 | | |
| Total | 9 | 3712,5 | | | |

| | *Coeficientes* | *Erro padrão* | *Stat t* | *valor-P* | *95% inferiores* | *95% superiores* | *Inferior 80,0%* | *Superior 80,0%* |
|---|---|---|---|---|---|---|---|---|
| Interseção | 184,16104 | 14,01440 | 13,14084 | 1,07E-06 | 151,84380 | 216,47830 | 164,58550 | 203,73660 |
| DIST. (X1) | -0,040259 | 0,00766 | -5,25474 | 0,00077 | -0,05793 | -0,02259 | -0,05096 | -0,02956 |

| Observação | Previsto(a) PREÇO UNITÁRIO (Y) | Resíduos | Resíduos padrão |
|---|---|---|---|
| 1 | 95,59121 | 4,40878 | 0,45799 |
| 2 | 103,64301 | 6,35698 | 0,66038 |
| 3 | 111,69481 | 8,30518 | 0,86276 |
| 4 | 123,77252 | 16,22747 | 1,68575 |
| 5 | 91,56531 | -6,56531 | -0,68202 |
| 6 | 107,66891 | -2,66891 | -0,27725 |
| 7 | 131,82432 | -11,82432 | -1,22834 |
| 8 | 95,59121 | -0,59121 | -0,06141 |
| 9 | 147,92792 | 2,07207 | 0,21525 |
| 10 | 115,72072 | -15,72072 | -1,63311 |

**Tabela 1.5**

Agora de posse dos coeficientes da equação de regressão linear simples, podemos concluir que a equação $\mathbf{P\hat{U} = B_0 + B_1 * DIST}$, fazendo as devidas substituições com os coeficientes, é:

$$\mathbf{P\hat{U} = 184{,}161 - 0{,}04026*DIST},$$

Que para DIST = 2.100 m resulta:

$$\mathbf{P\hat{U} = R\$\ 99{,}62\ /\ m^2}$$

# Na solução do exemplo encontramos:

| | |
|---|---|
| Nº variáveis independentes => k = | 1 |
| Nº total de variáveis => p = | 2 |
| Nº de observações => n = | 10 |
| Grau Liberdade = n - k -1 = gl = | 8 |
| Probabilidade P/ IC => $\alpha_{80\%}$ = | 80% |
| Probabilidade P/ F => $\alpha_{1\%}$ = | 1% |
| Probabilidade P/ T => $\alpha_{5\%}$ = | 5% |

| | | |
|---|---|---|
| $Y = 184{,}161 - 0{,}0403 * X_1$ | | |
| Cálc $t_{crítico}$ p/ 80% = | 1,8331 | *INVT((1-80%)/2;n-1)* |
| $Y_{avaliando}$ = | 99,62 | |
| Tamanho do $IC_t$ = | 5,9186 | *INT.CONFIANÇA.T((1-80%)/2;σ;n)* |
| Limites Inf. e Sup. do IC = | 93,6985 | 105,5358 |

**Valor do Imóvel: R$ 93,70 < PU< R$ 105,50**

# Calcular o valor do imóvel com duas variáveis independentes

| # | Vl. Unit.-Y | Dist-$X_1$ | Área-$X_2$ |
|---|---|---|---|
| 1 | 100,00 | 2200 | 300,00 |
| 2 | 110,00 | 2000 | 340,00 |
| 3 | 120,00 | 1800 | 270,00 |
| 4 | 140,00 | 1500 | 360,00 |
| 5 | 85,00 | 2300 | 400,00 |
| 6 | 105,00 | 1900 | 500,00 |
| 7 | 120,00 | 1300 | 600,00 |
| 8 | 95,00 | 2200 | 300,00 |
| 9 | 150,00 | 900 | 360,00 |
| 10 | 100,00 | 1700 | 600,00 |
| avaliando | ? | 2100 | 300,00 |

| | |
|---|---|
| Nº variáveis independentes => k = | 2 |
| Nº total de variáveis => p = | 3 |
| Nº de observações => n = | 10 |
| Grau Liberdade = n - k -1 = gl = | 7 |
| Probabilidade P/ IC => $\alpha_{80\%}$ = | 80% |
| Probabilidade P/ F => $\alpha_{1\%}$ = | 1% |
| Probabilidade P/ T => $\alpha_{5\%}$ = | 5% |

Inicia-se a solução vendo se existe colinearidade entre $X_1$ e $X_2$ calculando o Fator Inflacionário de Variância – FIV.

Para essa solução será utilizada uma nova função, chamada: **PROJ.LIN** e não a Análise de Variância - ANOVA

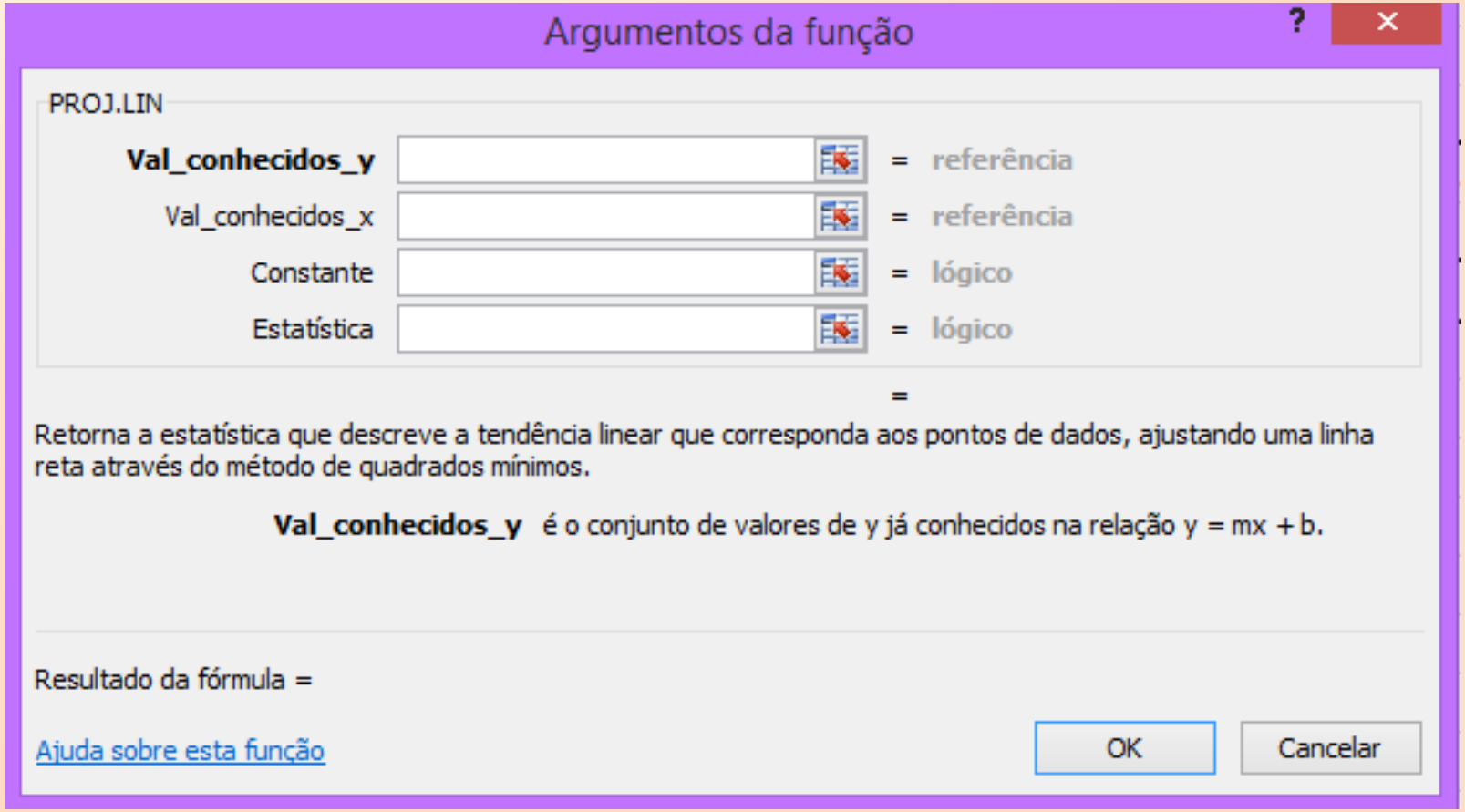
No cálculo do FIV necessitamos do coeficiente de determinação $r^2$, calculado pela PROJ.LIN, tendo $X_1$ como variável independente e $X_2$ como dependente.

$$FIV = \frac{1}{1 - r^2}$$

Inicialmente selecionamos uma região de 2 colunas por 5 linhas, onde teremos o resultado da PROJ.LIN em ***fx*.**

Em seguida, na caixa de entrada dos valores de Y, será marcada a região onde se encontram os valores da variável X1 e na caixa de entrada dos valores de X, será marcada a região onde se encontram os valores da variável X2. Nas demais caixas, Constante e Estatística, basta marcar o valor 1 (unidade).

**Não clique no botão OK**, mas use a sequência: **Ctrl + Shift + Enter**, visto que estamos tratando com matrizes



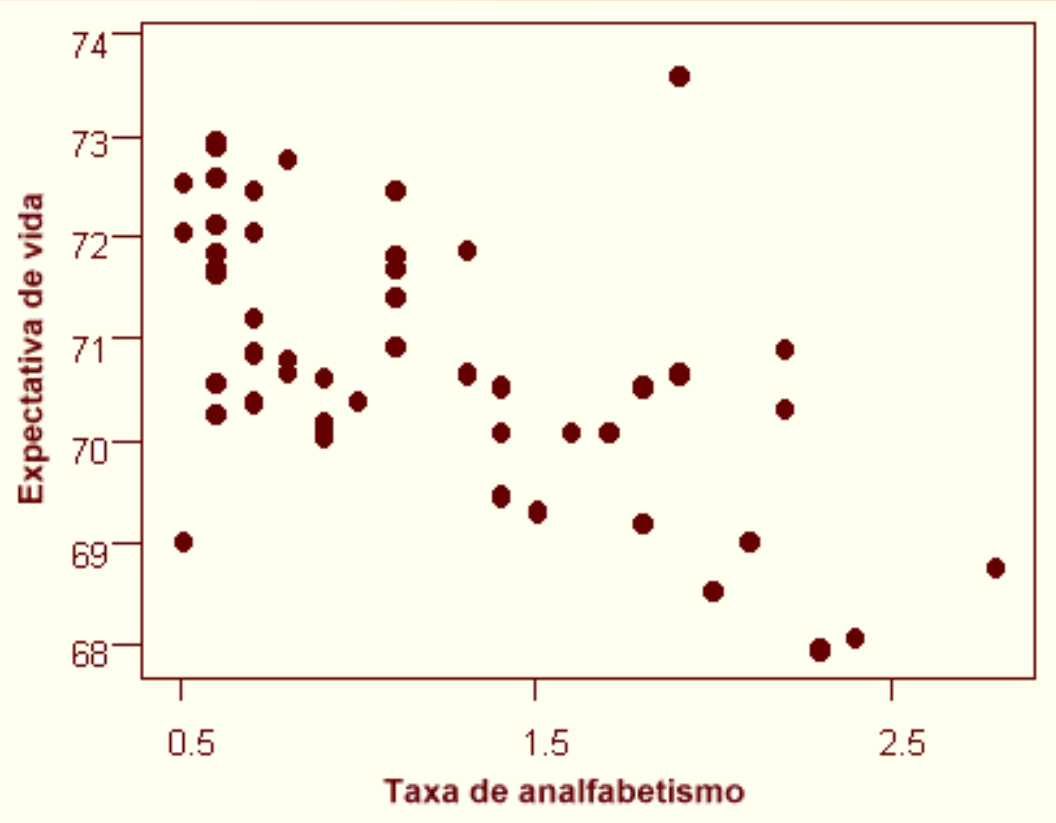
# Exemplo 3: expectativa de vida e analfabetismo

Considere as duas variáveis observadas em 50 estados norte-americanos.

Y: expectativa de vida

X: taxa de analfabetismo

# Diagrama de dispersão




**Podemos notar que, conforme aumenta a taxa de analfabetismo (X), a expectativa de vida (Y) tende a diminuir. Nota-se também uma tendência linear.**

# Cálculo da correlação

$\bar{Y}$= 70,88 (média de Y) e $S_Y$ = 1,342 (desvio padrão de Y)

$\bar{X}$= 1,17 (média de X) e $S_x$ = 0,609 (desvio padrão de X)

$\Sigma X_i Y_i$ = 4122,8 ; sendo n= 50

**Correlação entre X e Y:**

$$r = \frac{\sum_{1}^{n} X_i\, Y_i - n\,\bar{X}\,\bar{Y}}{(n-1)\,S_x\,S_y}$$

$$r = \frac{4122{,}8 - 50 \cdot 70{,}88 \cdot 1{,}17}{49 \cdot 1{,}342 \cdot 0{,}609} = \frac{-23{,}68}{40{,}047} = -0{,}59$$

# Reta ajustada:

$$\hat{Y} = a + bX$$

**O que são a e b?**

**a: intercepto**

**b: inclinação**

**Interpretação de b:**

**Para cada aumento de uma unidade em X, temos um aumento médio de b unidades em Y.**

# Reta ajustada (método de mínimos quadrados)

**Os coeficientes a e b são calculados da seguinte maneira:**

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

e

$$a = \overline{Y} - b.\overline{X}$$

Para os valores: $\overline{Y} = 7,38$ $\Sigma X_i Y_i = 509,12$

$\overline{X} = 1,17$ $n = 50$ $S_x = 0,609$

Calcular “a” ; “b” ; equação da reta e $\hat{y}$ p/ X=1,50:

**a = 2,398 ; b = 4,258 ; $\hat{y}$=2,398 + 4,258 X e $\hat{y}_{X=1,50}$= 8,79**

# No exemplo 2,

a reta ajustada é:

$$\hat{y} = 2{,}398 + 4{,}258\,X$$

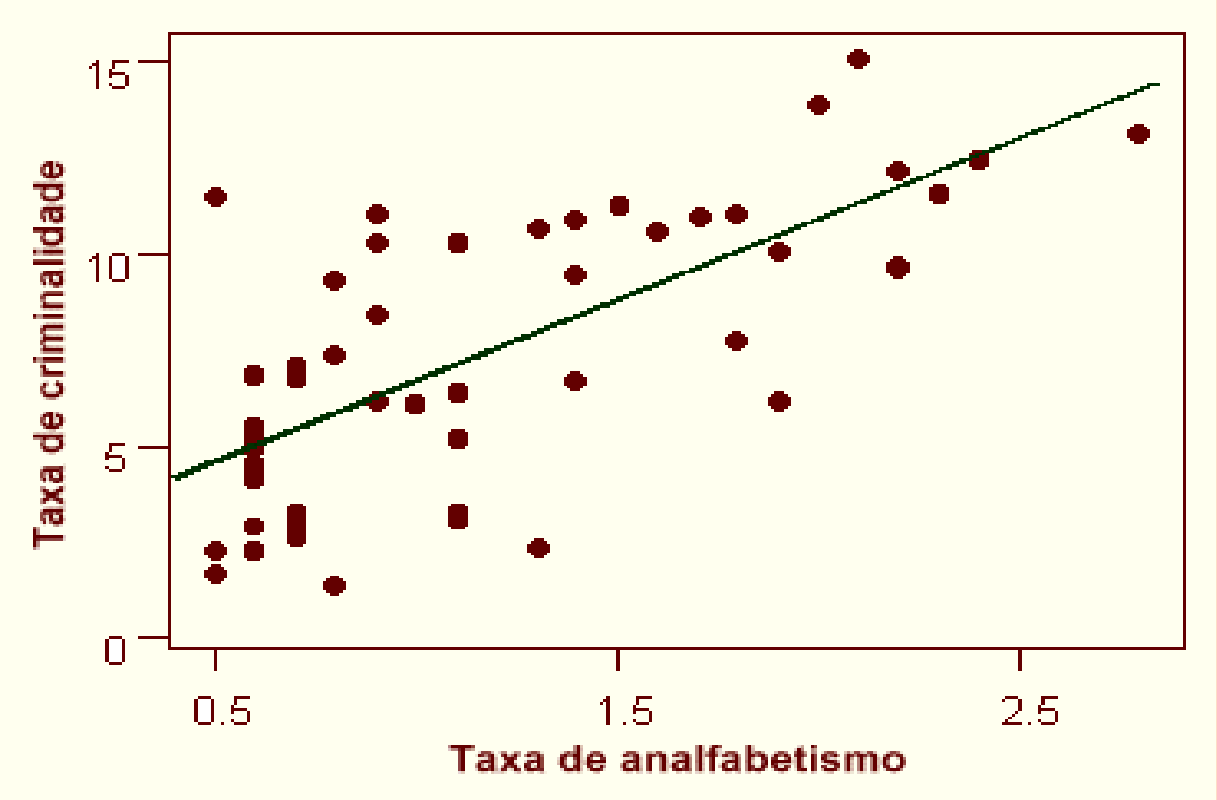
^
Y : valor predito para a taxa de criminalidade

X : taxa de analfabetismo

**Interpretação de b:**

Para um aumento de uma unidade na taxa do analfabetismo (X), a taxa de criminalidade (Y) aumenta, em média, 4,258 unidades.

# Graficamente, temos




## Como desenhar a reta no gráfico?

# No exemplo 3,

**Uma outra reta ajustada é:**

$$\hat{Y} = 72{,}395 - 1{,}296\ X$$

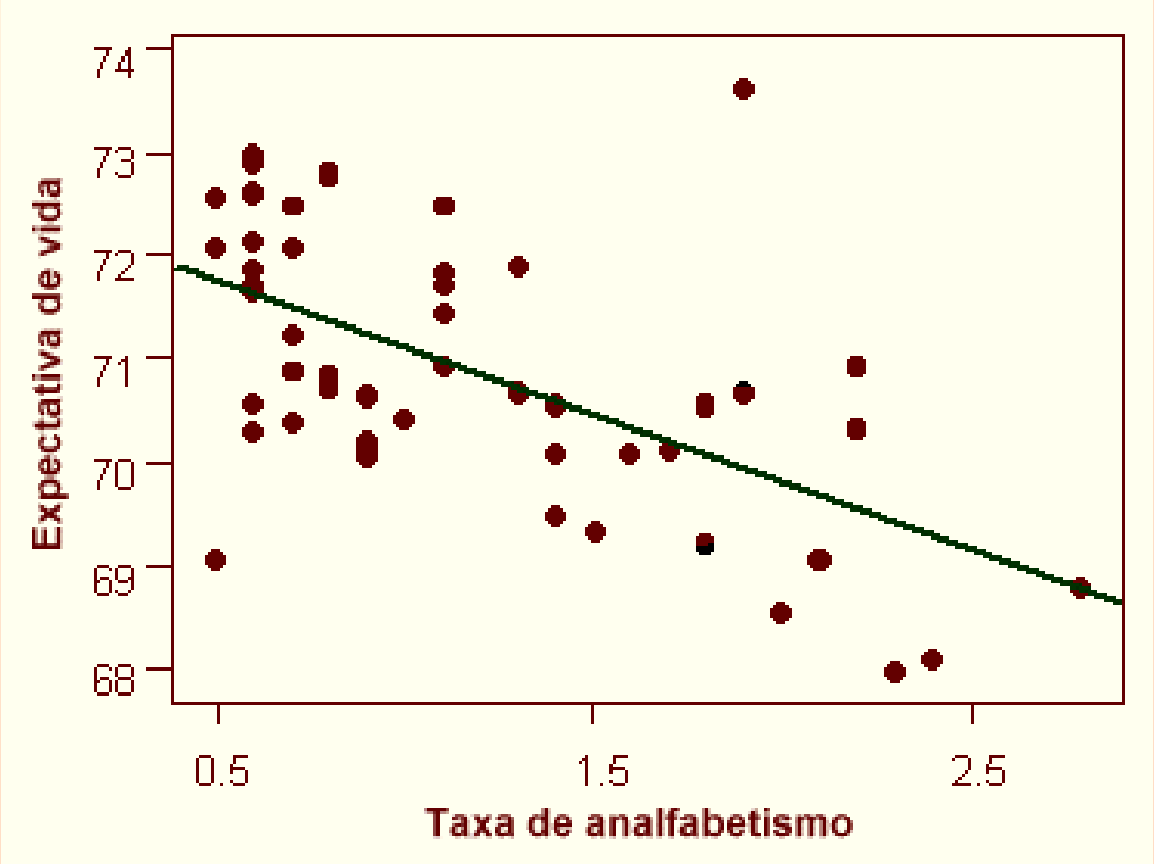
**$\hat{Y}$ : valor predito para a expectativa de vida**

**X : taxa de analfabetismo**

**Interpretação de b:**

**Para um aumento de uma unidade na taxa do analfabetismo (X), a expectativa de vida (Y) diminui, em média, 1,296 anos.**

# Graficamente, temos



# Exemplo 4: consumo de cerveja e temperatura

Y: consumo de cerveja diário por mil habitantes, em litros.

X: temperatura máxima (em ºC).

As variáveis foram observadas em nove localidades com as mesmas características demográficas e sócio-econômicas.

# Dados:

| Localidade | Temperatura (X) | Consumo (Y) |
|---|---|---|
| 1 | 16 | 290 |
| 2 | 31 | 374 |
| 3 | 38 | 393 |
| 4 | 39 | 425 |
| 5 | 37 | 406 |
| 6 | 36 | 370 |
| 7 | 36 | 365 |
| 8 | 22 | 320 |
| 9 | 10 | 269 |

# Calcule:

**a)** *r* = Coef. Correl Person;
**b)** reta de regressão e
**c)** consumo previsto para uma temperatura de 25ºC

$$r = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{(n-1)S_X S_Y}$$

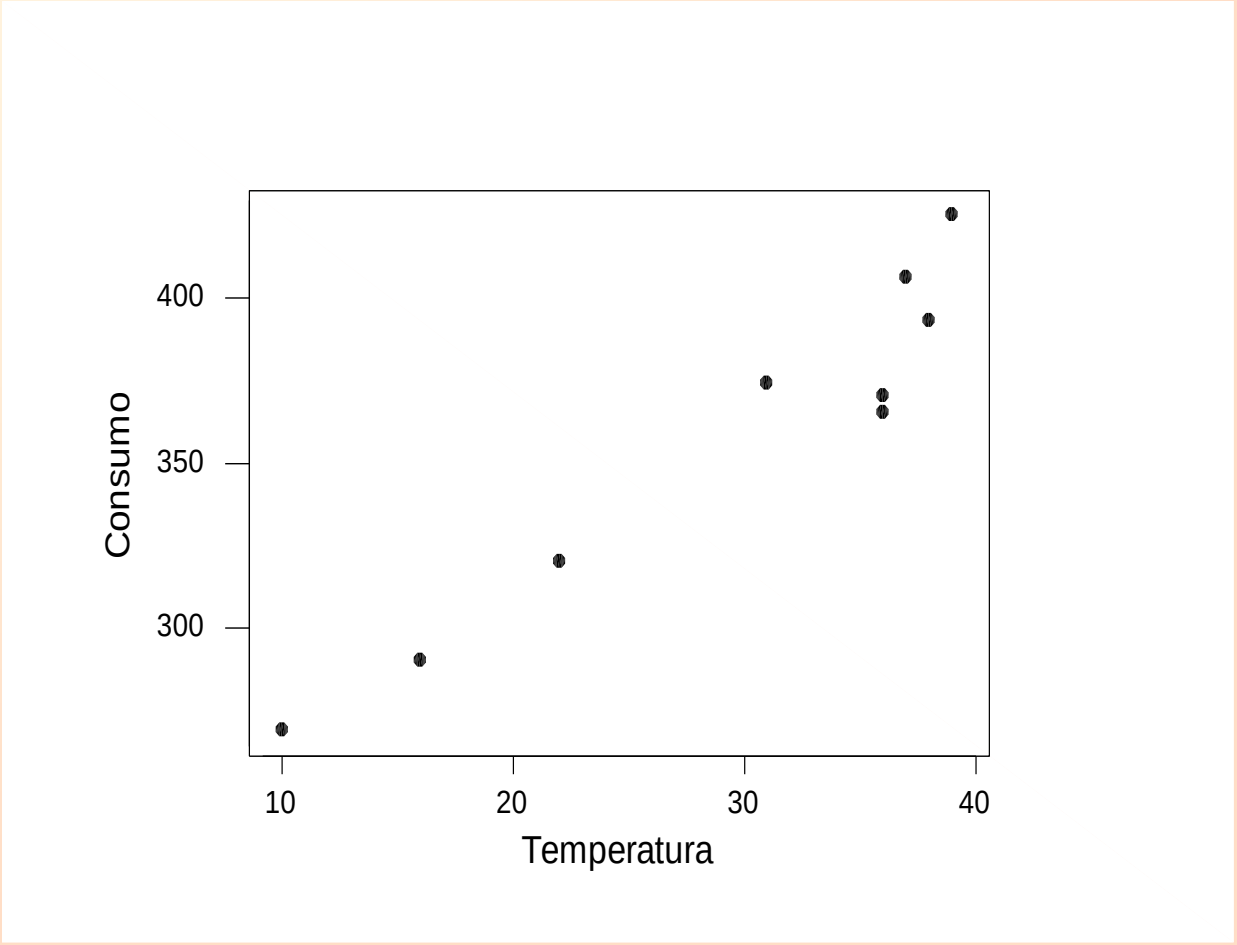
$$b = \frac{\sum_{i=1}^{n} X_i Y_i - n.\overline{X}.\overline{Y}}{(n-1).S_X^2}$$

$$a = \overline{Y} - b.\overline{X}$$

$$\hat{Y} = a + bX$$

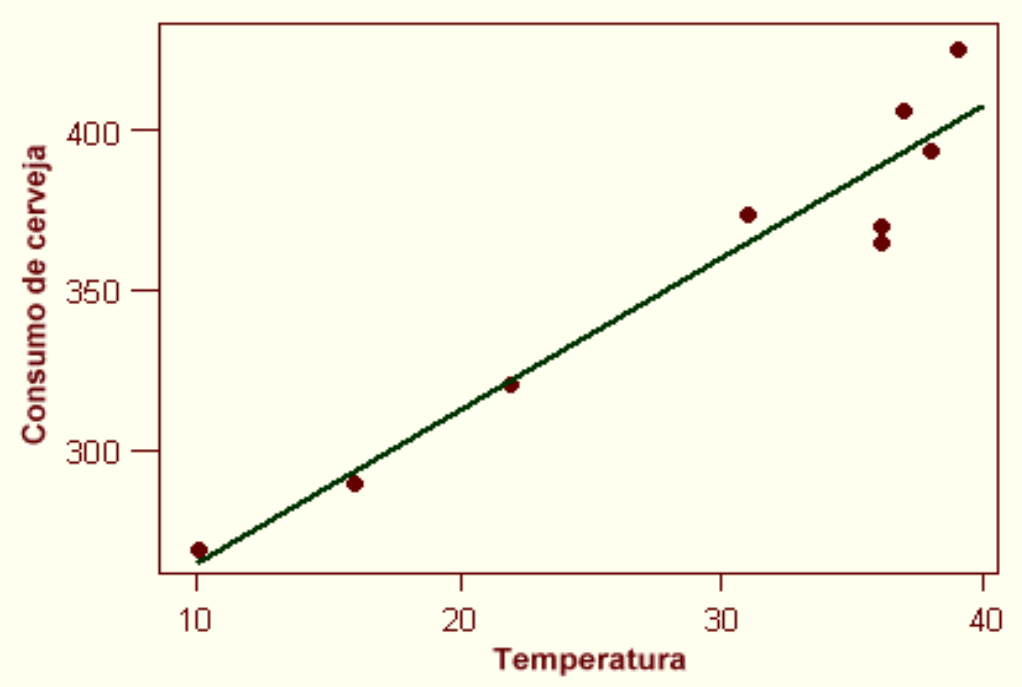
| X | Y |
|---|---|
| °C (temp) | Consumo |
| 16 | 290 |
| 31 | 374 |
| 38 | 393 |
| 39 | 425 |
| 37 | 406 |
| 36 | 370 |
| 36 | 365 |
| 22 | 320 |
| 10 | 269 |

# Diagrama de dispersão




**A correlação entre X e Y é r = 0,962.**

**A reta ajustada é:** $\hat{Y} = 217{,}37 + 4{,}74\,X$




**Qual a interpretação de b?**

**Aumentando-se um grau de temperatura (X), o consumo de cerveja (Y) aumenta, em média, 4,74 litros por mil habitantes.**

**Qual o consumo previsto para uma temperatura de 25ºC?**

$$\hat{Y} = 217{,}37 + 4{,}74.25 = 335{,}83 \text{ litros}$$

F I M