Estatística

Professora: Patrícia Ferreira Paranaíba

Regressão e Correlação

- Para estudar a relação entre duas (ou mais) variáveis quantitativas utilizamos a análise de regressão e correlação destas variáveis.
  - Altura e peso - espera-se que quanto mais alto mais pesado é o individuo;
  - Quantidade de memória RAM e tempo de processamento - espera-se que com mais memória RAM tenha-se um tempo menor de processamento.
- Regressão linear simples é o estudo que busca ajustar uma equação a um conjunto de dados de forma que a relação entre duas variáveis quantitativas possa ser expressa matematicamente.
- Correlação é um número entre -1 e 1 que mede o grau relacionamento entre duas variáveis quantitativas.
- Definimos um conjunto de variáveis (x, y), sendo x a variável independente e y a variável dependente. A primeira forma de verificar a relação de duas variáveis é traçar o gráfico de dispersão do dados.

Gráfico de dispersão

- O gráfico de dispersão contém uma variável independente representada no eixo horizontal e a variável dependente representada no eixo vertical.

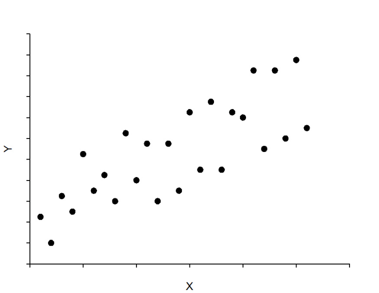

Figura: Indícios de correlação positiva, aumentando x, y também aumenta.

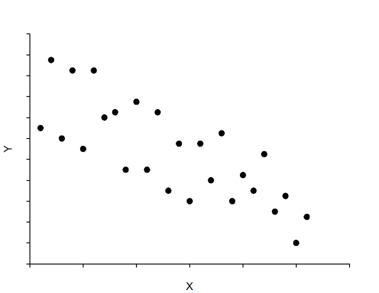

Figura: Indícios de correlação negativa, aumentando x, y diminui.

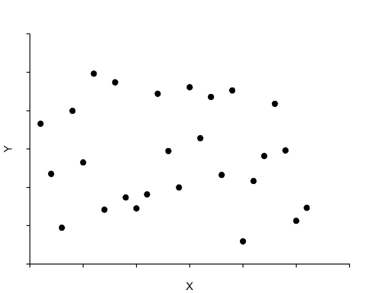

Figura: Indícios de ausência de correlação.

- O gráfico de dispersão da um ideia da existência de correlação, entretanto não apresenta qual a magnitude da correlação. Para determinar a magnitude da correlação utilizamos o coeficiente de correlação populacional (ρ). Em geral trabalhamos com amostras, e para estimar o coeficiente de correlação populacional pode-se utilizar o coeficiente de correlação amostral.

$$r = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}}$$

sendo que:

- $r > 0$ - correlação positiva;
- $r < 0$ - correlação negativa;
- $r = 0$ - ausência de correlação.

- Desta forma, deve ser realizado um teste de hipóteses para o coeficiente populacional, com base no resultado obtido na amostra, que pode ser definido da seguinte maneira:

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

- Rejeita-se H_0 se $|t_c| > t_{\frac{\alpha}{2}}$, em que:

$$t_c = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}}$$

nesse caso $v = n - 2$ graus de liberdade.

Exemplo: Em uma pesquisa feita com 7 famílias com renda bruta mensal entre 10 e 25 salários mínimos observou-se:

- X: renda bruta mensal (em salários mínimos).
- Y: porcentagem da renda bruta gasta com assistência médica.

x	10	12	14	16	18	20	22
y	11,8	10,2	12,1	13,2	15,1	15,4	15,6

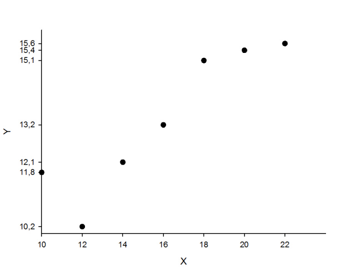

Tabela: Tabela auxiliar para o calculo da correlação

Observação	x	y	$(x-\overline{x})$	$(y-\overline{y})$	$(x-\overline{x})(y-\overline{y})$	$(x-\overline{x})^2$	$(y-\overline{y})^2$
1	10	11,8	-6	-1,5	9	36	2,25
2	12	10,2	-4	-3,1	12,4	16	9,61
3	14	12,1	-2	-1,2	2,4	4	1,44
4	16	13,2	0	-0,1	0	0	0,01
5	18	15,1	2	1,8	3,6	4	3,24
6	20	15,4	4	2,1	8,4	16	4,41
7	22	15,6	6	2,3	13,8	36	5,29
Total	112	93,4			49,6	112	26,25

$$
\begin{aligned}
\overline{x} &= \frac{\sum_{i}^{n} x_i}{n} = \frac{112}{7} = 16 \\
\overline{y} &= \frac{\sum_{i}^{n} y_i}{n} = \frac{93,4}{7} = 13,3 \\
r &= \frac{\sum_{i}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i}(x_i - \bar{x})^2 \sum_{i}(y_i - \bar{y})^2}} \\
&= \frac{49,6}{\sqrt{112 \times 26,25}} = 0,9148
\end{aligned}
$$

Verificou que o valor da correlação é r=0,9148.

- Vamos testar a hipótese se este valor é diferente de zero:

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

- Temos $v = n - 2 = 7 - 2 = 5$ graus de liberdade

$$t_c = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} = \frac{0,9148}{\sqrt{\frac{1-0,9148^2}{5}}} = 5,06$$

- Tomando-se $\alpha = 0,05$, temos $t_{0,025;5} = 2,571$.
- Como $|t_c| > t_{\frac{\alpha}{2}}$, rejeita-se H_0 ao nível de 5% de significância. Logo a correlação é diferente de zero e é igual a 0,9148.

Regressão linear simples

- A função que expressa a relação linear entre X e Y é dada por

$$y = a + bx + \epsilon$$

  em que:
  - a é coeficiente linear, interpretado como o valor da variável de dependente quando a variável inpendente é igual a 0;
  - b é coeficiente de regressão, interpretado como acréscimo na variável dependente para a variação de uma unidade na variável;
  - ϵ são os erros aleatórios de uma população normal, com média 0 e variância constante.
- Em geral, uma medição tem imperfeições que dão origem a um erro no resultado da medição.
- O erro aleatório se origina de variações temporais ou espaciais e ocorre de forma imprevisível. Os efeitos de tais variações (daqui para a frente denominaremos efeitos aleatórios) são a causa de variações em observações repetidas da grandeza.

Estimadores: Método dos Mínimos Quadrados

- Os estimadores para os coeficientes são:

$$a = \overline{y} - b\overline{x} \quad b = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2}$$

- Após ajustar o modelo de regressão deve-se realizar um teste de hipótese para verificar se os coeficientes são diferentes de zero:

$$\begin{cases} H_0 : \quad a = 0 \qquad H_0 : \quad b = 0 \\ H_1 : \quad a \neq 0 \qquad H_1 : \quad b \neq 0 \end{cases}$$

- A análise de variância é uma técnica utilizada para testar o ajuste da equação como um todo, ou seja, um teste para verificar se a equação de regressão obtida é significativa ou não.

Tabela: Análise de Variância para Regressão Linear Simples

Fontes de Variação	GL	Soma de Quadrados (SQ)	Quadrado Médio (QM)	Fc
Regressão	1	SQRegressão	QMRegressão	QMRegressão/
Erro	n-2	SQErro	QMErro	QMErro
Total	n-1	SQTotal		

$$\begin{aligned}
\text{SQTotal} &= \sum_i (y_i - \bar{y})^2 \\
\text{SQRegressão} &= b^2 \sum_i (x_i - \bar{x})^2 \\
\text{SQErro} &= \text{SQTotal} - \text{SQRegressão} \\
\text{QMRegressão} &= \text{SQRegressão} \\
\text{QMErro} &= \frac{SQErro}{n-2}
\end{aligned}$$

- O teste de hipótese para avaliar se o modelo de regressão é significativo é feito da seguinte forma:
  - Estabelecer o nível de significância α;
  - Obter o valor tabelado F_α;
  - Rejeita-se a hipótese H_0, se $F_c > F_\alpha$.
- O coeficiente de determinação r^2, é definido por:

$$r^2 = \frac{\text{SQRegressão}}{\text{SQTotal}} \quad 0 < r^2 < 1$$

  representa a porcentagem da variação total que é explicada pela equação de regressão, quanto maior o seu valor melhor.

Exemplo: Utilizando o exemplo da renda bruta mensal (em salários mínimos) e a porcentagem da renda bruta gasta com assistência médica.

$$
\begin{aligned}
b &= \frac{\displaystyle\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2} \\
&= \frac{49,6}{112} = 0,44 \\
a &= \overline{y} - b\overline{x} \\
&= 6,26
\end{aligned}
$$

- Assim a equação de regressão é igual a

$$y = 6,26 + 0,44x$$

- Vamos verificar se a regressão é significativa

$$
\begin{aligned}
\text{SQTotal} &= \sum_i (y_i - \bar{y})^2 = 26,25 \\
\text{SQRegressão} &= \frac{\left(\sum_i (x_i - \bar{x})(y_i - \bar{y})\right)^2}{\sum_i (x_i - \bar{x})^2} \\
&= \frac{(49,6)^2}{112} = 21,97 \\
\text{SQErro} &= \text{SQTotal} - \text{SQRegressão} \\
&= 26,25 - 21,97 = 4,28
\end{aligned}
$$

Tabela: Análise de Variância para Regressão Linear Simples

Fontes de Variação	GL	Soma de Quadrados (SQ)	Quadrado Médio (QM)	Fc	F_α
Regressão	1	21,97	21,97	25,55	6,60
Erro	5	4,28	0,86		
Total	6	26,25			

- Como o $F_c > F_\alpha$, rejeita-se H_0, logo o modelo de regressão linear é significativo.
- Obtendo o r^2

$$r^2 = \frac{\text{SQRegressão}}{\text{SQTotal}} = \frac{21,97}{26,25} = 0,8370 = 83,70\%$$

- Assim verifica-se que é a renda bruta explica $83,70\%$ da variação do gasto com assistência médica.