Aula 12

Banco do Brasil (Escriturário - Agente de Tecnologia) Probabilidade e Estatística - 2023 (Pós-Edital)

Autor:
Equipe Exatas Estratégia Concursos

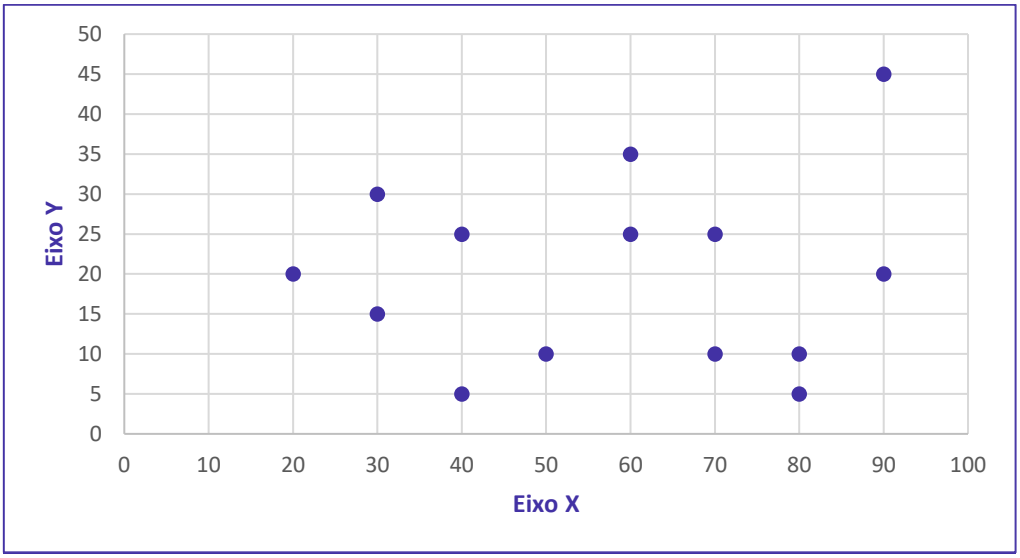
05 de Janeiro de 2023

Índice

Correlação Linear

Neste tópico estudaremos a correlação linear. A correlação é usada para indicar a força que mantém unidos dois conjuntos de valores. Por meio da análise da correlação linear, buscamos identificar se existe alguma relação entre duas ou mais variáveis, ou seja, se as alterações nas variáveis estão associadas umas com a outras.

Para avaliar a existência de correlação podemos recorrer a uma forma de representação gráfica bem simples, que chamamos de **gráfico de dispersão**. Basicamente, ela é uma representação de pares ordenados em um plano cartesiano, composto por um eixo vertical (ordenada) e um eixo horizontal (abcissa).

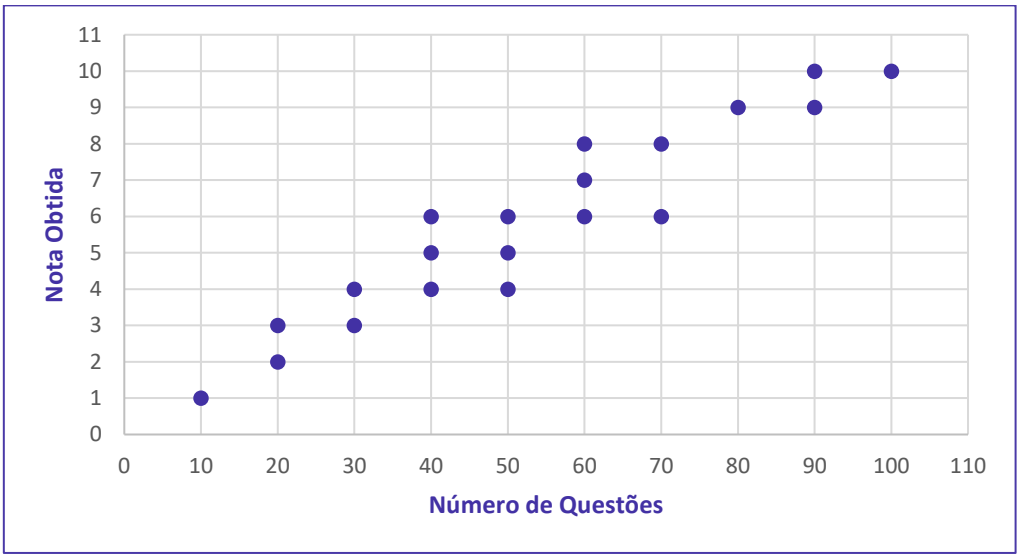


A título exemplificativo, as informações da tabela a seguir referem-se ao número de questões resolvidas por um determinado aluno e a nota obtida por ele em uma avaliação. Observe que quanto maior o número de questões resolvidas, maior é a nota obtida na avaliação.

Aluno	Número de Questões (X)	Nota Obtida (Y)
1	20	2
2	60	8
3	30	3
4	50	6
5	40	4
6	70	8
7	80	9
8	90	10
9	40	6

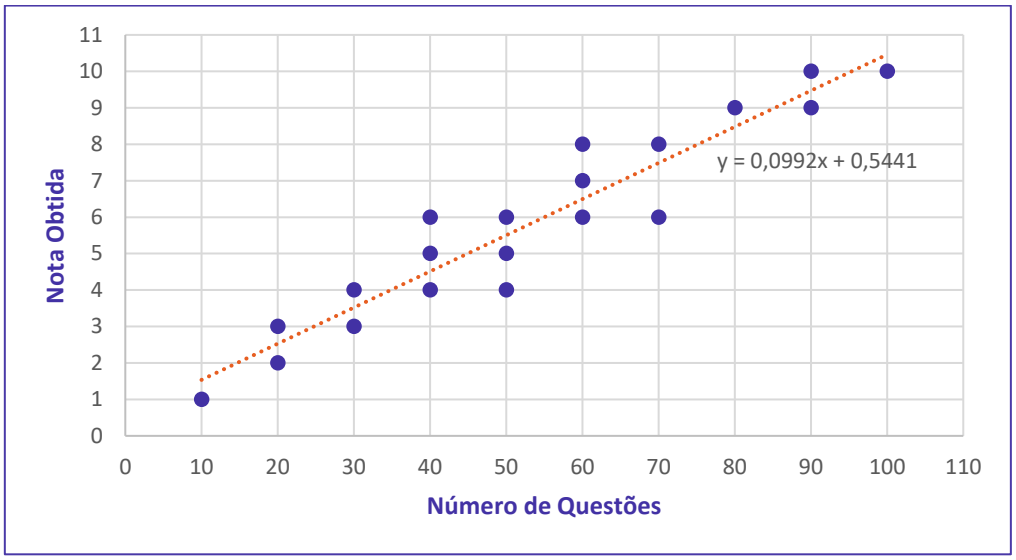
10	30	4
11	10	1
12	60	6
13	50	5
14	70	6
15	90	9
16	100	10
17	20	3
18	40	5
19	60	7
20	50	4

A representação desses dados em formato de diagrama de dispersão sugere a existência de uma **relação linear positiva (variação no mesmo sentido)** entre as duas variáveis:



Neste exemplo, percebemos que a relação dos dados agrupados é quase linear. Por isso, se traçarmos uma reta de tendência no gráfico, observaremos que os pontos se comportarão em torno da reta.

Assim:



Há situações em que essa relação linear não é tão evidente. Um exemplo disso é quando os pontos estão mais dispersos. Nesse caso, para identificarmos a relação existente entre as variáveis, usamos o coeficiente de correlação linear de Pearson, definido por r.

Por fim, devemos ter em mente que a correlação linear pode ser:

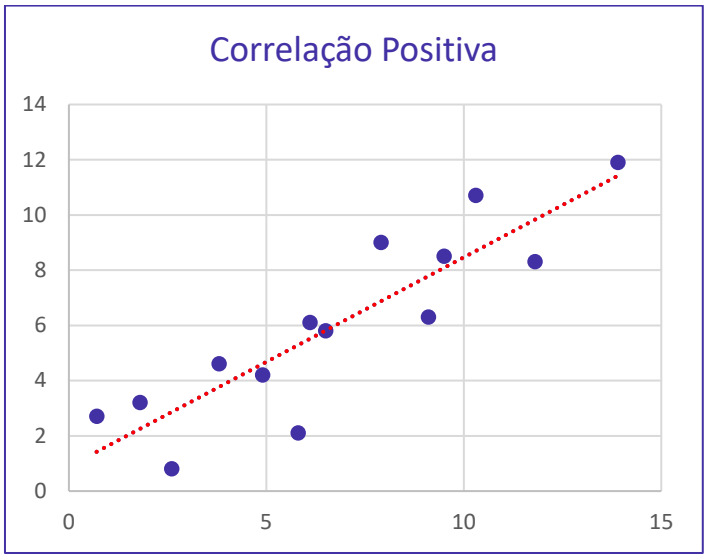
a) direta ou positiva – quando temos dois fenômenos que variam no mesmo sentido. Se aumentarmos ou diminuirmos um deles, o outro também aumentará ou diminuirá;

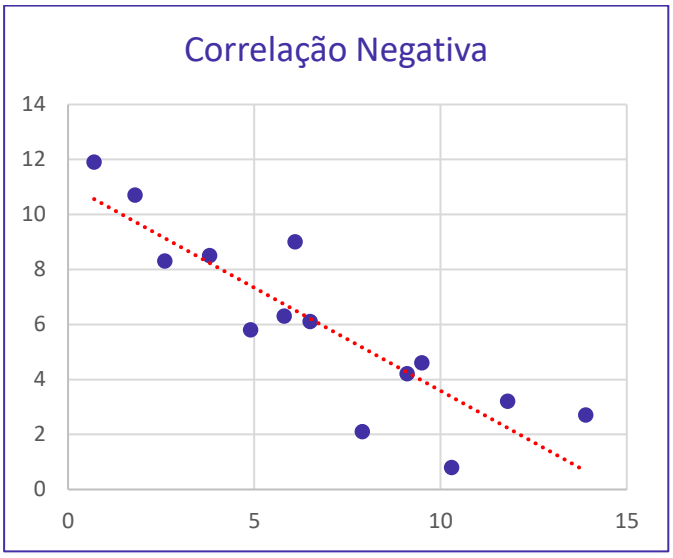
b) inversa ou negativa – quando temos dois fenômenos que variam em sentido contrário. Se aumentarmos ou diminuirmos um deles, acontecerá o contrário com o outro, no caso, diminuirá ou aumentará;

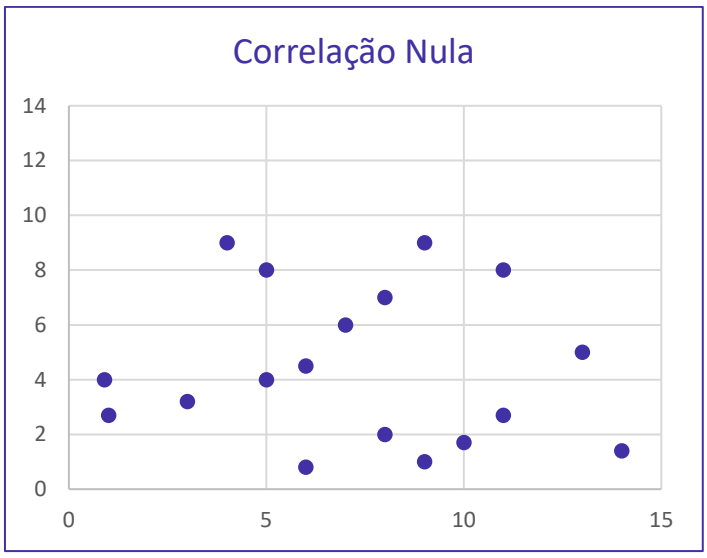
c) inexistente ou nula – quando não existe correlação ou dependência entre os dois fenômenos. Nessa situação, o valor do coeficiente de correlação linear será zero ($r = 0$) ou um valor aproximadamente igual a zero ($r \cong 0$); e

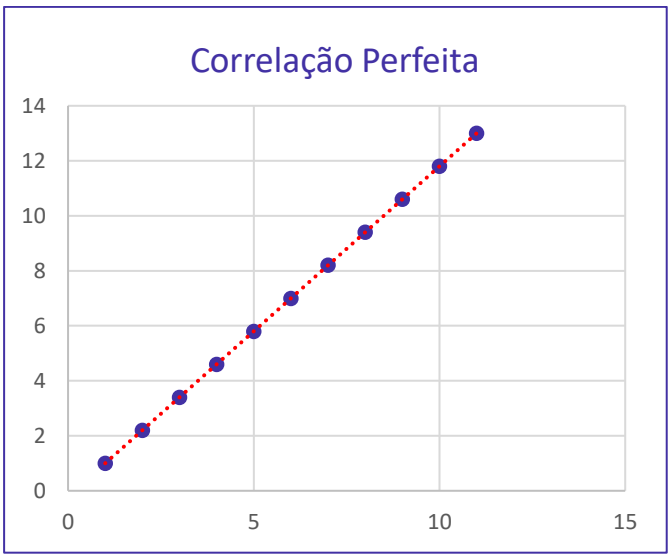
d) perfeita – quando os fenômenos se ajustam perfeitamente a uma reta.

As figuras a seguir ilustram essas quatro situações:









Coeficiente de Correlação de Pearson

O coeficiente de correlação linear de Pearson é adotado para medir o quão forte é a relação linear entre duas variáveis. Esse coeficiente é calculado pela seguinte expressão:

$$r = \frac{\sum_{i=1}^{n}[(X_i - \bar{X}) \times (Y_i - \bar{Y})]}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \times \sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

Os somatórios dessa fórmula podem ser simplificados, o que facilita a resolução de muitas questões. Por isso, é muito importante que vocês aprendam a expressão a seguir:

$$\sum_{i=1}^{n}[(X_i - \bar{X}) \times (Y_i - \bar{Y})] = \sum_{i=1}^{n}(X_i \times Y_i) - n \times \bar{X} \times \bar{Y}$$

Para facilitar a compreensão e internalização, vou apresentar um raciocínio que podemos adotar para deduzir a fórmula alternativa mostrada anteriormente.

Primeiro, precisamos aplicar a propriedade distributiva:

$$\sum_{i=1}^{n}[(X_i - \bar{X}) \times (Y_i - \bar{Y})] = \sum_{i=1}^{n}[X_i \times Y_i - X_i \times \bar{Y} - \bar{X} \times Y_i + \bar{X} \times \bar{Y}]$$

Agora, precisamos separar as quatro parcelas desse somatório principal. Reparem que as médias são constantes, portanto, podem sair do somatório:

$$\sum_{i=1}^{n}[(X_i - \bar{X}) \times (Y_i - \bar{Y})] = \sum_{i=1}^{n}(X_i \times Y_i) - \bar{Y} \times \left(\sum_{i=1}^{n} X_i\right) - \bar{X} \times \left(\sum_{i=1}^{n} Y_i\right) + \bar{X} \times \bar{Y} \times \sum_{i=1}^{n} 1$$

Nesse ponto, devemos lembrar que $\sum_{i=1}^{n} X_i = n \times \bar{X}$ e $\sum_{i=1}^{n} Y_i = n \times \bar{Y}$. Logo,

$$\sum_{i=1}^{n}[(X_i - \bar{X}) \times (Y_i - \bar{Y})] = \sum_{i=1}^{n}(X_i \times Y_i) - (\bar{Y} \times n \times \bar{X}) - (\bar{X} \times n \times \bar{Y}) + (\bar{X} \times \bar{Y} \times n)$$

Observem que as duas últimas parcelas se anulam:

$$\sum_{i=1}^{n}[(X_i - \bar{X}) \times (Y_i - \bar{Y})] = \sum_{i=1}^{n}(X_i \times Y_i) - (n \times \bar{X} \times \bar{Y}) - (n \times \bar{X} \times \bar{Y}) + (n \times \bar{X} \times \bar{Y})$$

Portanto,

$$\sum_{i=1}^{n}[(X_i - \bar{X}) \times (Y_i - \bar{Y})] = \sum_{i=1}^{n}(X_i \times Y_i) - n \times \bar{X} \times \bar{Y}$$

Utilizaremos essa fórmula alternativa para calcular o numerador do coeficiente de correlação. Reparem que a expressão do lado esquerdo nos obriga a calcular todos os desvios $(X_i - \bar{X})$ e $(Y_i - \bar{Y})$, enquanto a expressão do lado direito não. Nessa fórmula, n indica o número de pontos no gráfico de dispersão, isto é, o número de pares ordenados.

Na fórmula anterior, se substituirmos Y por X, teremos a seguinte expressão:

$$\sum_{i=1}^{n}[(X_i - \bar{X}) \times (X_i - \bar{X})] = \sum_{i=1}^{n}(X_i \times X_i) - n \times \bar{X} \times \bar{X}$$

$$\boldsymbol{\sum_{i=1}^{n}(X_i - \bar{X})^2 = \sum_{i=1}^{n}X_i^2 - n \times (\bar{X})^2}$$

Já, se substituirmos X por Y, iremos obter:

$$\sum_{i=1}^{n}[(Y_i - \bar{Y}) \times (Y_i - \bar{Y})] = \sum_{i=1}^{n}(Y_i \times Y_i) - n \times \bar{Y} \times \bar{Y}$$

$$\boldsymbol{\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n}Y_i^2 - n \times (\bar{Y})^2}$$

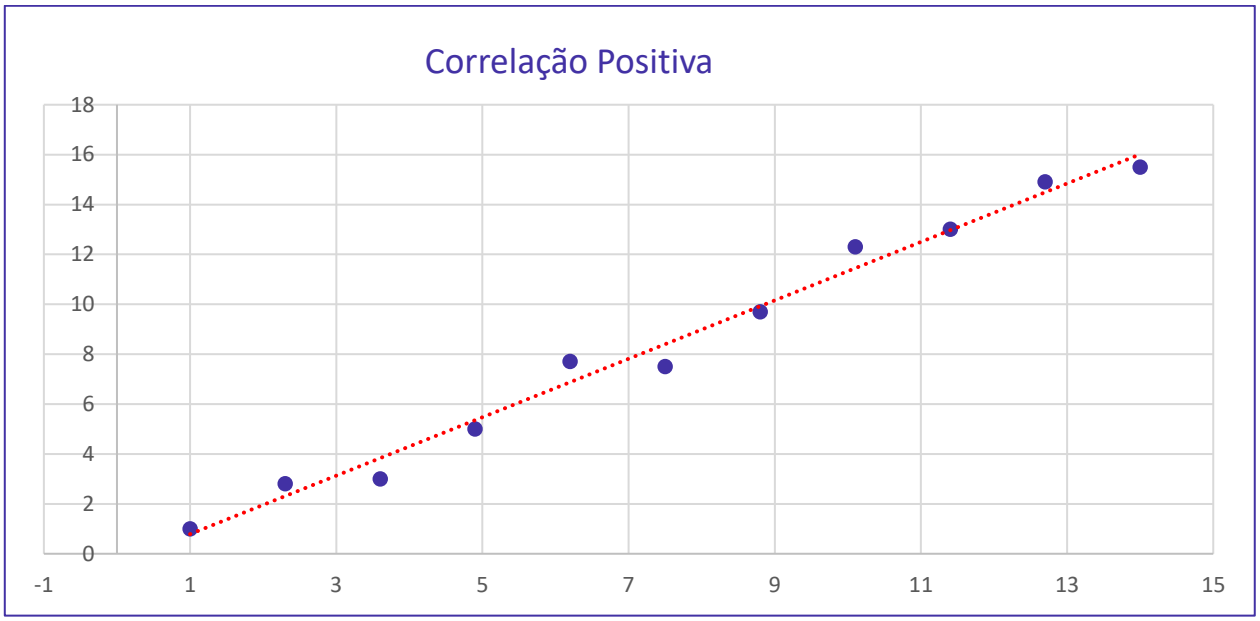
As últimas duas fórmulas são formas alternativas que podem ser empregadas no cálculo do denominador do coeficiente de correlação.

O coeficiente de correlação de Pearson pode assumir quaisquer valores entre 1 e -1, ou seja:

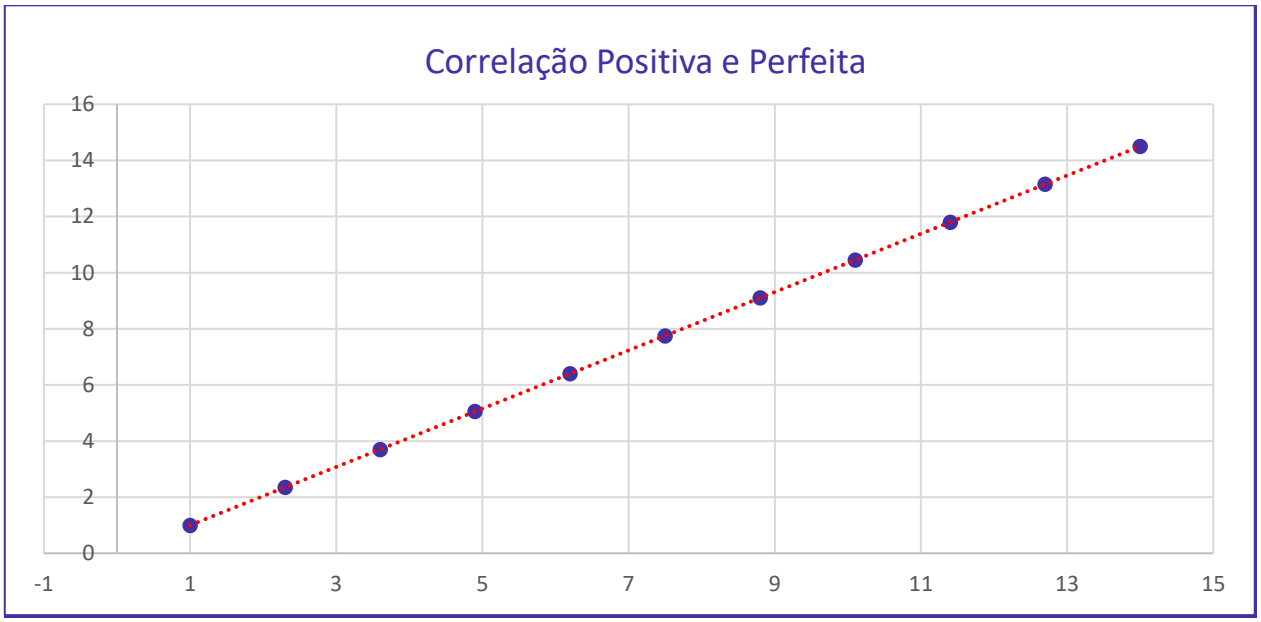
$$\boldsymbol{-1 \leq r \leq 1}$$

Assim, quanto mais próximo r estiver de 0, menor será a relação linear entre as duas variáveis. Por sua vez, quanto mais próximo r estiver de (1 ou -1), maior será a relação linear entre as duas variáveis.

O valor de r é positivo quando a variável Y tende a aumentar ou a diminuir se X também aumentar ou diminuir, respectivamente. Nessa situação, dizemos que as variáveis são positivamente correlacionadas. No exemplo a seguir, os dados estão praticamente em cima de uma reta, indicando a existência de uma correlação positiva forte, isto é, r muito próximo de 1. No caso, o coeficiente de correlação foi de 0,99267.

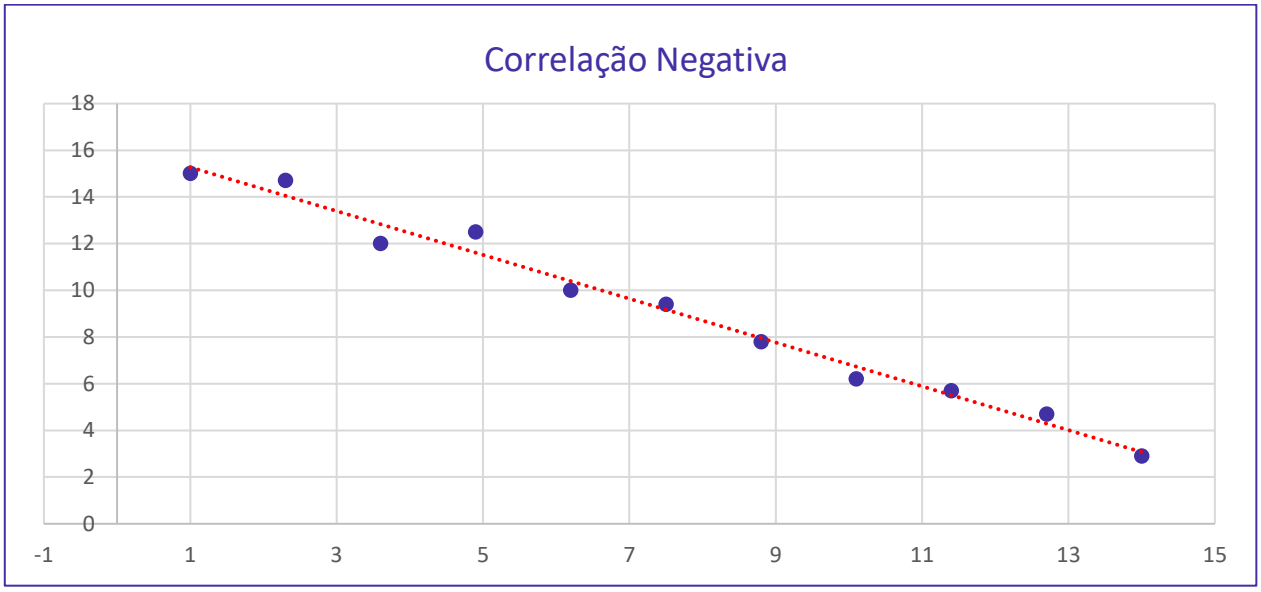


Se a correlação for positiva e todos os pontos estiverem sobre uma mesma reta, o valor de r será exatamente 1. Nesse caso, dizemos que a correlação é **positiva e perfeita**.

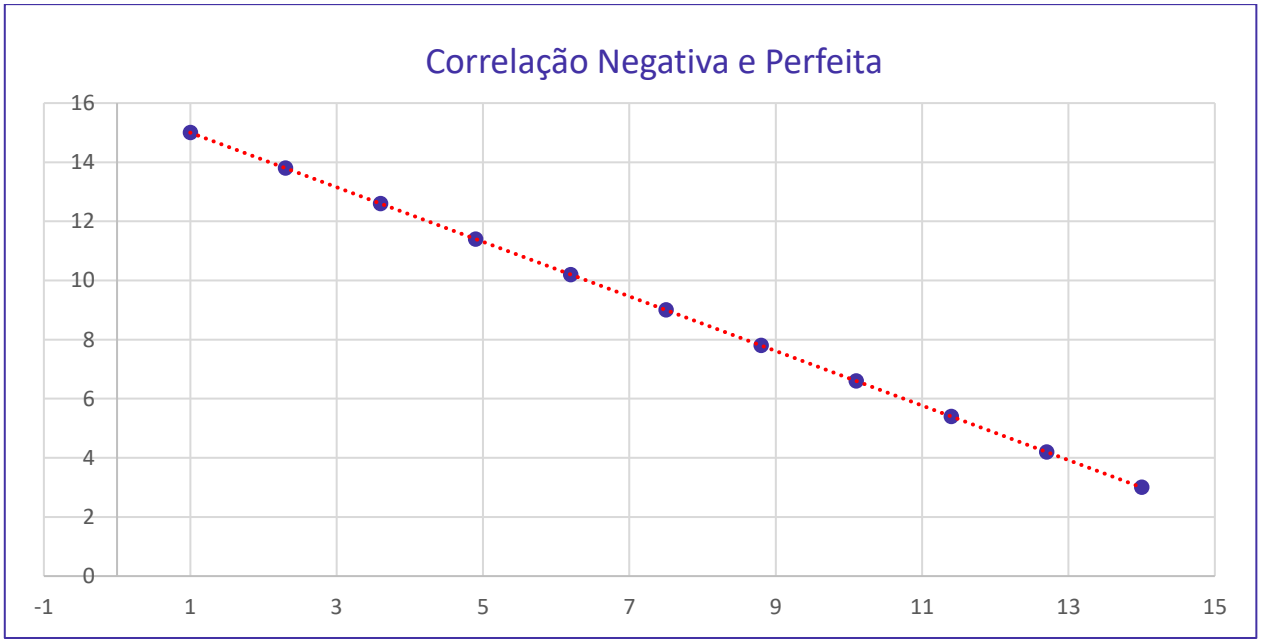


O valor de r é negativo quando a variável Y tende a diminuir ou aumentar quando X aumentar ou diminuir, respectivamente. Nessa situação, dizemos que as variáveis estão negativamente correlacionadas. No

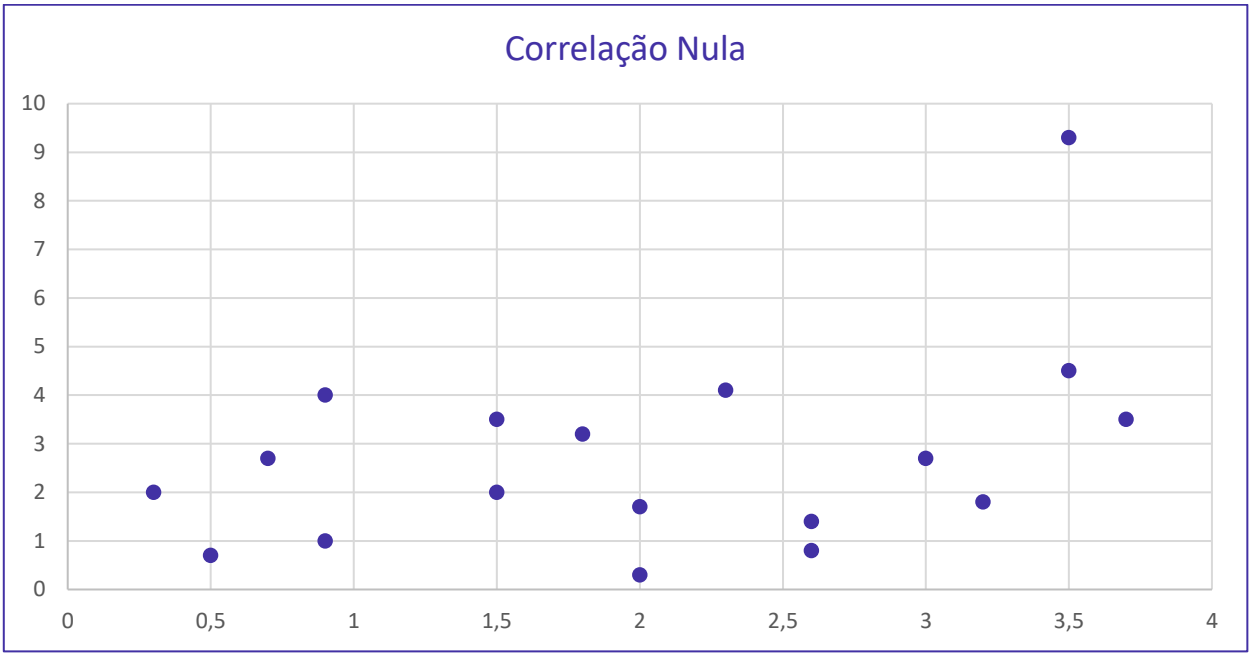
exemplo a seguir, os dados estão praticamente em cima de uma reta, indicando a existência de uma correlação negativa forte, isto é, r muito próximo de -1. No caso, o coeficiente de correlação foi de -0,9918.



Se a correlação for negativa e todos os pontos estiverem sobre uma mesma reta, o valor de r será exatamente -1. Nesse caso, dizemos que a correlação é **negativa e perfeita**.



O valor de r é zero (ou um valor muito próximo de zero) quando não existe uma relação linear entre as variáveis. No exemplo a seguir, o coeficiente de correlação é 0,1132. Nesse caso, dizemos que a **correlação linear é nula ou inexistente**.



Vejamos agora um exemplo numérico.

As questões normalmente informam os valores dos somatórios e exigem apenas a aplicação correta da fórmula. Apesar disso, vamos considerar uma tabela com 5 pares ordenados, representando as notas de 5 alunos nas disciplinas X e Y, para calcular o coeficiente de correlação pelas duas fórmulas:

Aluno	X_i	Y_i
1	6,50	7,00
2	7,50	8,00
3	8,00	8,00
4	8,50	9,00
5	9,50	10,00

Primeiro, teremos que calcular as médias de X e Y:

$$\bar{X} = \frac{6{,}50 + 7{,}50 + 8{,}00 + 8{,}50 + 9{,}50}{5} = \frac{40}{5} = 8{,}00$$

$$\bar{Y} = \frac{7{,}00 + 8{,}00 + 8{,}00 + 9{,}00 + 10{,}00}{5} = \frac{42}{5} = 8{,}40$$

Agora, calcularemos os desvios de X e Y em relação às suas médias:

Aluno	X_i	Y_i	$X_i - \bar{X}$	$Y_i - \bar{Y}$
1	6,50	7,00	6,50 - 8,00 = -1,50	7,00 - 8,40 = -1,40
2	7,50	8,00	7,50 - 8,00 = -0,50	8,00 - 8,40 = -0,40
3	8,00	8,00	8,00 - 8,00 = 0,00	8,00 - 8,40 = -0,40
4	8,50	9,00	8,50 - 8,00 = 0,50	9,00 - 8,40 = 0,60
5	9,50	10,0	9,50 - 8,00 = 1,50	10,00 - 8,40 = 1,60

▪

Vou limpar a memória de cálculo para facilitar a visualização:

Aluno	X_i	Y_i	$X_i - \bar{X}$	$Y_i - \bar{Y}$
1	6,50	7,00	-1,50	-1,40
2	7,50	8,00	-0,50	-0,40
3	8,00	8,00	0,00	-0,40
4	8,50	9,00	0,50	0,60
5	9,50	10,0	1,50	1,60

Nesse ponto, teremos que calcular o numerador e o denominador do coeficiente de correlação. Para tanto, precisaremos multiplicar os desvios de X pelos desvios de Y, bem como calcular os quadrados dos desvios:

Aluno	X_i	Y_i	$X_i - \bar{X}$	$Y_i - \bar{Y}$	$(X_i - \bar{X}) \times (Y_i - \bar{Y})$	$(X_i - \bar{X})^2$	$(Y_i - \bar{Y})^2$
1	6,50	7,00	-1,50	-1,40	(-1,50) x (-1,40) = 2,10	$(-1,50)^2 = 2,25$	$(-1,40)^2 = 1,96$
2	7,50	8,00	-0,50	-0,40	(-0,50) x (-0,40) = 0,20	$(-0,50)^2 = 0,25$	$(-0,40)^2 = 0,16$
3	8,00	8,00	0,00	-0,40	(0,00) x (-0,40) = 0,00	$(0,00)^2 = 0,00$	$(-0,40)^2 = 0,16$
4	8,50	9,00	0,50	0,60	(0,50) x (0,60) = 0,30	$(0,50)^2 = 0,25$	$(0,60)^2 = 0,36$
5	9,50	10,0	1,50	1,60	(1,50) x (1,60) = 2,40	$(1,50)^2 = 2,25$	$(1,60)^2 = 2,56$

Limpando a memória de cálculo e deixando apenas os resultados. Vejamos:

Aluno	X_i	Y_i	$X_i - \bar{X}$	$Y_i - \bar{Y}$	$(X_i - \bar{X}) \times (Y_i - \bar{Y})$	$(X_i - \bar{X})^2$	$(Y_i - \bar{Y})^2$
1	6,50	7,00	-1,50	-1,40	2,10	2,25	1,96
2	7,50	8,00	-0,50	-0,40	0,20	0,25	0,16
3	8,00	8,00	0,00	-0,40	0,00	0,00	0,16
4	8,50	9,00	0,50	0,60	0,30	0,25	0,36
5	9,50	10,0	1,50	1,60	2,40	2,25	2,56

Conhecendo esses valores, podemos calcular os somatórios da fórmula de correlação.

$$\sum_{i=1}^{5} [(X_i - \bar{X}) \times (Y_i - \bar{Y})] = 2,10 + 0,20 + 0,00 + 0,30 + 2,40 = 5,00$$

$$\sum_{i=1}^{5} (X_i - \bar{X})^2 = 2,25 + 0,25 + 0,00 + 0,25 + 2,25 = 5,00$$

$$\sum_{i=1}^{5} (Y_i - \bar{Y})^2 = 1,96 + 0,16 + 0,16 + 0,36 + 2,56 = 5,20$$

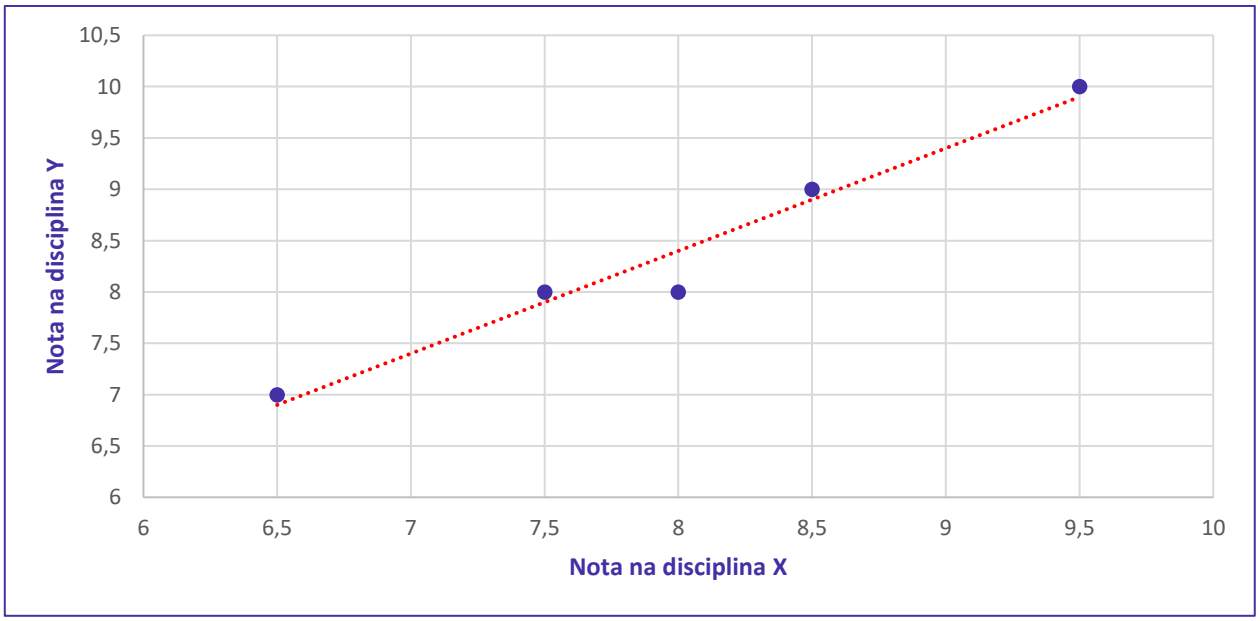
Aplicando esses valores na fórmula do coeficiente de correlação linear, temos:

$$r = \frac{\sum_{i=1}^{n}[(X_i - \bar{X}) \times (Y_i - \bar{Y})]}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \times \sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

$$r = \frac{5{,}00}{\sqrt{5{,}00 \times 5{,}20}}$$

$$r \cong 0{,}9805$$

O coeficiente de correlação linear ficou muito próximo de 1, o que implica dizer que existe uma relação linear intensa entre as notas das duas disciplinas. Vejamos o gráfico de dispersão das duas variáveis



Pronto, agora utilizaremos as fórmulas alternativas para calcular o mesmo coeficiente de correlação. Vamos relembrar a fórmula:

$$r = \frac{\sum_{i=1}^{n}[(X_i - \bar{X}) \times (Y_i - \bar{Y})]}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \times \sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

O numerador pode ser calculado mediante a aplicação da seguinte fórmula:

$$\sum_{i=1}^{n}[(X_i - \bar{X}) \times (Y_i - \bar{Y})] = \sum_{i=1}^{n}(X_i \times Y_i) - n \times \bar{X} \times \bar{Y}$$

Por sua vez, o denominador pode ser calculado por meio das seguintes fórmulas:

$$\sum_{i=1}^{n}(X_i - \bar{X})^2 = \sum_{i=1}^{n} X_i^2 - n \times (\bar{X})^2$$

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

Retornemos à tabela inicial:

Aluno	X_i	Y_i
1	6,50	7,00
2	7,50	8,00
3	8,00	8,00
4	8,50	9,00
5	9,50	10,00

Já calculamos as médias de X e Y:

$$\bar{X} = \frac{6{,}50 + 7{,}50 + 8{,}00 + 8{,}50 + 9{,}50}{5} = \frac{40}{5} = 8{,}00$$

$$\bar{Y} = \frac{7{,}00 + 8{,}00 + 8{,}00 + 9{,}00 + 10{,}00}{5} = \frac{42}{5} = 8{,}40$$

Precisamos de três colunas adicionais: $X \times Y$, X^2 e Y^2

Aluno	X_i	Y_i	$X_i \times Y_i$	X_i^2	Y_i^2
1	6,50	7,00	6,50 x 7,00 = 45,50	$(6{,}50)^2 = 42{,}25$	$(7{,}00)^2 = 49{,}00$
2	7,50	8,00	7,50 x 8,00 = 60,00	$(7{,}50)^2 = 56{,}25$	$(8{,}00)^2 = 64{,}00$
3	8,00	8,00	8,00 x 8,00 = 64,00	$(8{,}00)^2 = 64{,}00$	$(8{,}00)^2 = 64{,}00$
4	8,50	9,00	8,50 x 9,00 = 76,50	$(8{,}50)^2 = 72{,}25$	$(9{,}00)^2 = 81{,}00$
5	9,50	10,0	9,50 x 10,00 = 95,00	$(9{,}50)^2 = 90{,}25$	$(10{,}00)^2 = 100{,}00$

Limpando a memória de cálculo, ficamos com os seguintes resultados:

Aluno	X_i	Y_i	$X_i \cdot Y_i$	X_i^2	Y_i^2
1	6,50	7,00	45,50	42,25	49,00
2	7,50	8,00	60,00	56,25	64,00
3	8,00	8,00	64,00	64,00	64,00
4	8,50	9,00	76,50	72,25	81,00
5	9,50	10,0	95,00	90,25	100,00

Agora, podemos calcular os somatórios da fórmula:

$$\sum_{i=1}^{5}(X_i \times Y_i) = 45{,}50 + 60{,}00 + 64{,}00 + 76{,}50 + 95{,}00 = 341{,}00$$

$$\sum_{i=1}^{5} X_i^2 = 42{,}25 + 56{,}25 + 64{,}00 + 72{,}25 + 90{,}25 = 325{,}00$$

$$\sum_{i=1}^{5} Y_i^2 = 49{,}00 + 64{,}00 + 64{,}00 + 81{,}00 + 100{,}00 = 358$$

Já temos todas as informações necessárias para a aplicação das fórmulas alternativas.

O numerador do coeficiente de correlação é calculado por:

$$\sum_{i=1}^{5}(X_i \times Y_i) - 5 \times \bar{X} \times \bar{Y} = 341{,}00 - 5 \times 8{,}00 \times 8{,}40 = 5{,}00$$

Os termos do denominador são calculados pelas seguintes fórmulas:

$$\sum_{i=1}^{5} X_i^2 - 5 \times (\bar{X})^2 = 325 - 5 \times 8{,}00^2 = 5{,}00$$

$$\sum_{i=1}^{5} Y_i^2 - 5 \times (\bar{Y})^2 = 358 - 5 \times 8{,}40^2 = 5{,}20$$

Aplicando a fórmula do coeficiente de correlação, temos:

$$r = \frac{\sum_{i=1}^{n}[(X_i - \bar{X}) \times (Y_i - \bar{Y})]}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \times \sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

$$r = \frac{5{,}00}{\sqrt{5{,}00 \times 5{,}20}}$$

$$r \cong 0{,}9805$$

Portanto, o resultado produzido mediante a aplicação das fórmulas alternativas é exatamente o mesmo das fórmulas tradicionais.

O coeficiente de correlação linear também pode ser definido por meio das seguintes expressões:

$$\boxed{r = \frac{S_{XY}}{\sqrt{S_{XX} \times S_{YY}}}}$$

Em que $S_{XY} = \sum_{i=1}^{n}[(X_i - \bar{X}) \times (Y_i - \bar{Y})]$; $S_{XX} = \sum_{i=1}^{n}(X_i - \bar{X})^2$ e $S_{YY} = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$.

Também pode aparecer na seguinte forma:

$$\boxed{r = \frac{Cov(X,Y)}{\sigma_X \times \sigma_Y}}$$

Em que $Cov(X,Y)$ representa a covariância das variáveis X e Y; σ_X e σ_Y representam o desvio padrão, respectivamente, das variáveis X e Y.

Propriedades do Coeficiente de Correlação

1ª Propriedade

- **O coeficiente de correlação não sofre alteração quando uma constante é adicionada a (ou subtraída de) uma variável.**

Considere os dados da seguinte tabelar:

i	X_i	Y_i
1	1	6
2	2	7
3	3	8
4	4	9
5	5	10

Como vimos, primeiro temos que encontrar as médias das duas variáveis:

$$\bar{X} = \frac{1+2+3+4+5}{5} = 3$$

$$\bar{Y} = \frac{6+7+8+9+10}{5} = 8$$

Agora, vamos montar a tabela auxiliar com os desvios $(X_i - \bar{X})$ e $(Y_i - \bar{Y})$, e os respectivos produtos $(X_i - \bar{X}) \times (Y_i - \bar{Y})$, $(X_i - \bar{X})^2$ e $(Y_i - \bar{Y})^2$.

i	X_i	Y_i	$X_i - \bar{X}$	$Y_i - \bar{Y}$	$(X_i - \bar{X}) \times (Y_i - \bar{Y})$	$(X_i - \bar{X})^2$	$(Y_i - \bar{Y})^2$
1	1	6	-2	-2	4	4	4
2	2	7	-1	-1	1	1	1
3	3	8	0	0	0	0	0
4	4	9	1	1	1	1	1
5	5	10	2	2	4	4	4
Total					**10**	**10**	**10**

Aplicando esses valores na fórmula do coeficiente de correlação linear, temos:

$$r = \frac{\sum_{i=1}^{n}[(X_i - \bar{X}) \times (Y_i - \bar{Y})]}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \times \sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

$$r = \frac{10}{\sqrt{10 \times 10}}$$

$$r(X, Y) = 1$$

Logo, o coeficiente de correlação linear das variáveis X e Y é 1.

Agora, vamos adicionar 3 unidades à variável X e 5 unidades à variável Y.

i	$X_i + 3$	$Y_i + 5$
1	4	11
2	5	12
3	6	13
4	7	14
5	8	15

Novamente, temos que encontrando as médias das duas variáveis:

$$\bar{X} = \frac{4 + 5 + 6 + 7 + 8}{5} = 6$$

$$\bar{Y} = \frac{11 + 12 + 13 + 14 + 15}{5} = 13$$

Construindo a tabela auxiliar:

i	X_i	Y_i	$X_i - \bar{X}$	$Y_i - \bar{Y}$	$(X_i - \bar{X}) \times (Y_i - \bar{Y})$	$(X_i - \bar{X})^2$	$(Y_i - \bar{Y})^2$
1	4	11	-2	-2	4	4	4
2	5	12	-1	-1	1	1	1
3	6	13	0	0	0	0	0
4	7	14	1	1	1	1	1
5	8	15	2	2	4	4	4
Total					**10**	**10**	**10**

Aplicando esses valores na fórmula do coeficiente de correlação linear, temos:

$$r = \frac{10}{\sqrt{10 \times 10}}$$

$$r(X + 3, Y + 5) = 1$$

Portanto, o coeficiente de correlação não sofreu alteração quando com a adição de 3 unidades à variável X e 5 unidades à variável Y.

Essa propriedade simplifica a resolução de certas questões. Suponhamos que tivéssemos uma tabela com os seguintes valores e precisássemos calcular a correlação entre as variáveis X e Y.

i	X_i	Y_i
1	201	301
2	202	302
3	204	308
4	205	309

Reparem que podemos subtrair 200 de todos os valores de X e 300 de todos os valores de Y. Poderíamos, então, fazer uma transformação nessas variáveis, obtendo:

i	X_i	Y_i	$(X_i - 200)$	$(Y_i - 300)$
1	201	301	1	1
2	202	302	2	2
3	204	308	4	8
4	205	309	5	9

A propriedade estudada nos garante que $r(X, Y) = r(X - 200, Y - 300)$. Logo, poderíamos calcular a correlação por meio dos valores transformados.

Encontrando as médias das variáveis transformadas:

$$\bar{X} = \frac{1+2+4+5}{4} = 3$$

$$\bar{Y} = \frac{1+2+8+9}{4} = 5$$

Organizando a tabela auxiliar, teríamos:

i	X_i	Y_i	$(X_i - 200)$	$(Y_i - 300)$	$X_i - \bar{X}$	$Y_i - \bar{Y}$	$(X_i - \bar{X}) \times (Y_i - \bar{Y})$	$(X_i - \bar{X})^2$	$(Y_i - \bar{Y})^2$
1	201	301	1	1	-2	-4	8	4	16
2	202	302	2	2	-1	-3	3	1	9
3	204	308	4	8	1	3	3	1	9
4	205	309	5	9	2	4	8	4	16
Total							**22**	**10**	**50**

Aplicando esses valores na fórmula do coeficiente de correlação linear, teríamos:

$$r(X - 200, Y - 300) = \frac{22}{\sqrt{10 \times 50}} \cong 0{,}984$$

De fato, se jogássemos a tabela inicial em um software de estatística, veríamos que $r(X, Y) = 0{,}984$.

2ª Propriedade

- **O o coeficiente de correlação pode não sofrer alteração ou pode ter seu sinal alterado quando uma variável é multiplicada (ou dividida) por uma constante. Caso as constantes tenham o mesmo sinal, o valor do coeficiente de correlação não será alterado. Por outro lado, se as constantes tiverem sinais contrários, o coeficiente mudará de sinal, mas o valor permanecerá inalterado.**

Ainda com relação ao exemplo trabalhado na propriedade anterior, vamos multiplicar a variável X por 2 e a variável Y por 2.

i	$X_i \times 2$	$Y_i \times 2$
1	2	12
2	4	14
3	6	16
4	8	18
5	10	20

Encontrando as médias das duas variáveis:

$$\bar{X} = \frac{2+4+6+8+10}{5} = 6$$

$$\bar{Y} = \frac{12+14+16+18+20}{5} = 16$$

Montando a tabela auxiliar:

i	X_i	Y_i	$X_i - \bar{X}$	$Y_i - \bar{Y}$	$(X_i - \bar{X}) \times (Y_i - \bar{Y})$	$(X_i - \bar{X})^2$	$(Y_i - \bar{Y})^2$
1	2	12	-4	-4	16	16	16
2	4	14	-2	-2	4	4	4
3	6	16	0	0	0	0	0
4	8	18	2	2	4	4	4
5	10	20	4	4	16	16	16
Total					**40**	**40**	**40**

Aplicando esses valores na fórmula do coeficiente de correlação linear, temos:

$$r = \frac{40}{\sqrt{40 \times 40}}$$

$$r(2X, 2Y) = 1$$

Logo, a multiplicação por constantes de mesmo sinal não alterou o valor do coeficiente de correlação, nem implicou na alteração de seu sinal.

Se tivéssemos multiplicado a variável X por -2 e a variável Y por -2:

i	$X_i \times (-2)$	$Y_i \times (-2)$
1	-2	-12
2	-4	-14
3	-6	-16
4	-8	-18
5	-10	-20

Nessa situação, as médias das duas variáveis são:

$$\bar{X} = \frac{(-2) + (-4) + (-6) + (-8) + (-10)}{5} = -6$$

$$\bar{Y} = \frac{(-12) + (-14) + (-16) + (-18) + (-20)}{5} = -16$$

Construindo a tabela auxiliar, temos:

i	X_i	Y_i	$X_i - \bar{X}$	$Y_i - \bar{Y}$	$(X_i - \bar{X}) \times (Y_i - \bar{Y})$	$(X_i - \bar{X})^2$	$(Y_i - \bar{Y})^2$
1	-2	-12	4	4	16	16	16
2	-4	-14	2	2	4	4	4
3	-6	-16	0	0	0	0	0
4	-8	-18	-2	-2	4	4	4
5	-10	-20	-4	-4	16	16	16
Total					**40**	**40**	**40**

Aplicando esses valores na fórmula do coeficiente de correlação linear, temos:

$$r = \frac{40}{\sqrt{40 \times 40}}$$

Como as constantes possuíam sinais iguais, o sinal do coeficiente de correlação foi mantido.

$$r(-2X, -2Y) = 1$$

Novamente, a multiplicação por constantes de mesmo sinal não alterou o valor do coeficiente de correlação, nem implicou na alteração de seu sinal.

Finalmente, vamos multiplicar a variável X por 2 e a variável Y por -2, constantes com sinais contrários.

i	$X_i \times 2$	$Y_i \times (-2)$
1	2	- 12
2	4	- 14
3	6	- 16
4	8	- 18
5	10	- 20

Encontrando as médias das duas variáveis:

$$\bar{X} = \frac{2 + 4 + 6 + 8 + 10}{5} = 6$$

$$\bar{Y} = \frac{(-12) + (-14) + (-16) + (-18) + (-20)}{5} = -16$$

Organizando a tabela auxiliar, temos:

i	X_i	Y_i	$X_i - \bar{X}$	$Y_i - \bar{Y}$	$(X_i - \bar{X}) \times (Y_i - \bar{Y})$	$(X_i - \bar{X})^2$	$(Y_i - \bar{Y})^2$
1	2	-12	-4	4	-16	16	16
2	4	-14	-2	2	-4	4	4
3	6	-16	0	0	0	0	0
4	8	-18	2	-2	-4	4	4
5	10	-20	4	-4	-16	16	16

Total	**-40**	**40**	**40**

Aplicando esses valores na fórmula do coeficiente de correlação linear, temos:

$$r = \frac{-40}{\sqrt{40 \times 40}}$$

Portanto, como as constantes possuem sinais contrários, o sinal do coeficiente de correlação foi invertido.

$$r(2X, -2Y) = -1$$

Essa propriedade pode simplificar a resolução de determinadas questões. Suponhamos que tivéssemos uma tabela com os seguintes valores e precisássemos calcular a correlação entre as variáveis X e Y.

i	X_i	Y_i
1	200	300
2	350	350
3	400	450
4	450	500

Reparem que todos os valores podem ser divididos por 50. Poderíamos, então, fazer uma transformação nessas variáveis, obtendo:

i	X_i	Y_i	$\left(X_i/50\right)$	$\left(Y_i/50\right)$
1	200	300	4	6
2	350	350	7	7
3	400	450	8	9
4	450	500	9	10

A propriedade estudada nos garante que $r(X,Y) = r\left(\frac{X}{50}, \frac{Y}{50}\right)$. Logo, poderíamos calcular a correlação por meio dos valores transformados.

Encontrando as médias das variáveis transformadas:

$$\bar{X} = \frac{4+7+8+9}{4} = 7$$

$$\bar{Y} = \frac{6+7+9+10}{4} = 9$$

Organizando a tabela auxiliar, teríamos:

i	X_i	Y_i	$\left({}^{X_i}/_{50}\right)$	$\left({}^{Y_i}/_{50}\right)$	$X_i - \bar{X}$	$Y_i - \bar{Y}$	$(X_i - \bar{X}) \times (Y_i - \bar{Y})$	$(X_i - \bar{X})^2$	$(Y_i - \bar{Y})^2$
1	200	300	4	6	-3	-2	6	9	4
2	350	350	7	7	0	-1	0	0	1
3	400	450	8	9	1	1	1	1	1
4	450	500	9	10	2	2	4	4	4
Total							**11**	**14**	**10**

Aplicando esses valores na fórmula do coeficiente de correlação linear, teríamos:

$$r\left(\frac{X}{50}, \frac{Y}{50}\right) = \frac{11}{\sqrt{14 \times 10}} \cong 0{,}93$$

De fato, se jogássemos a tabela inicial em um software de estatística, veríamos que $r(X,Y) = 0{,}93$.

REGRESSÃO LINEAR SIMPLES

A regressão simples é uma continuação do conceito de correlação/covariância. A regressão tenta explicar a relação de uma variável chamada dependente, usando outra variável chamada independente.

Na regressão linear simples queremos calcular a expressão matemática que relaciona Y (variável dependente) em função de X (variável independente). Como estamos falando de regressão linear simples, trata-se da equação que representa uma reta. Essa equação pode ser escrita como:

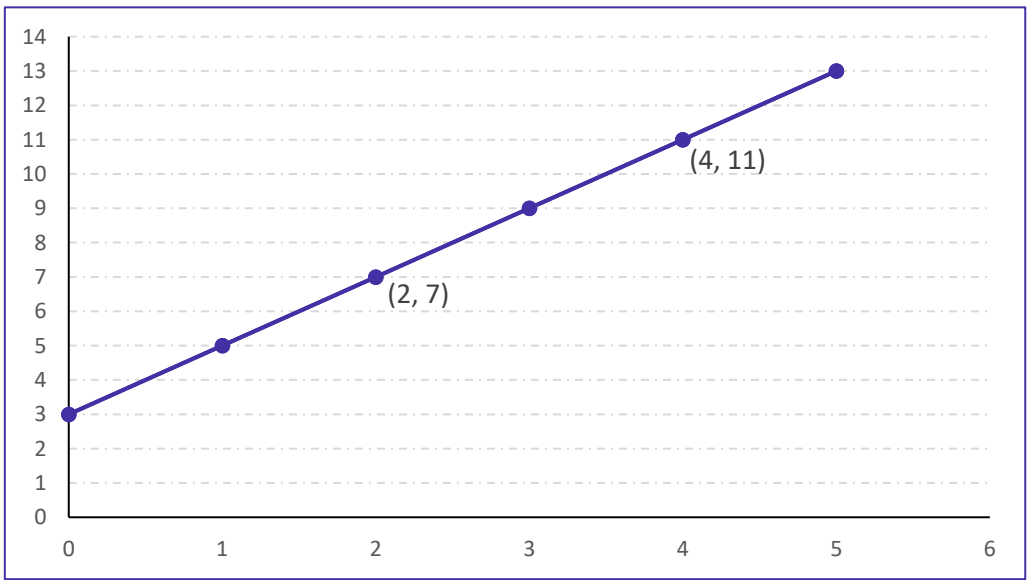
$$y = m \cdot x + b$$

O coeficiente m é conhecido como **taxa de variação** ou **coeficiente angular da reta.** Esse coeficiente indica que uma função é **crescente se $m > 0$**; **decrescente se $m < 0$**; ou **constante se $m = 0$**. Para uma reta que passa pelos pontos (x_0, y_0) e (x, y), o coeficiente angular é expresso por:

$$m = \frac{\Delta y}{\Delta x} = \frac{y - y_0}{x - x_0}$$

O coeficiente b é conhecido como **coeficiente linear da reta** e determina o ponto em que a reta intercepta o eixo y.

Vamos calcular a reta apresentada na figura abaixo, que passa pelos pontos (2, 7) e (4, 11).



O **coeficiente angular da reta (m)** é o quociente entre a variação de y e a variação de x. Podemos escolher qualquer um dos pontos como referência para o cálculo da variação, desde que tenhamos atenção na hora de aplicar os dados na fórmula. A ordem a ser considerada é sempre $x - x_0$ e $y - y_0$, em que x_0 e y_0 são as

coordenadas do ponto tomado como referência. Assim, se adotarmos o ponto $(2,7)$ como referência, teremos:

$$m = \frac{\Delta y}{\Delta x} = \frac{11 - 7}{4 - 2} = 2$$

Dessa forma, a equação da reta fica:

$$y = m \cdot x + b$$

$$y = 2 \cdot x + b$$

Para calcular o valor de b, podemos usar qualquer ponto da reta, a exemplo de (2, 7).

$$7 = 2 \cdot 2 + b$$

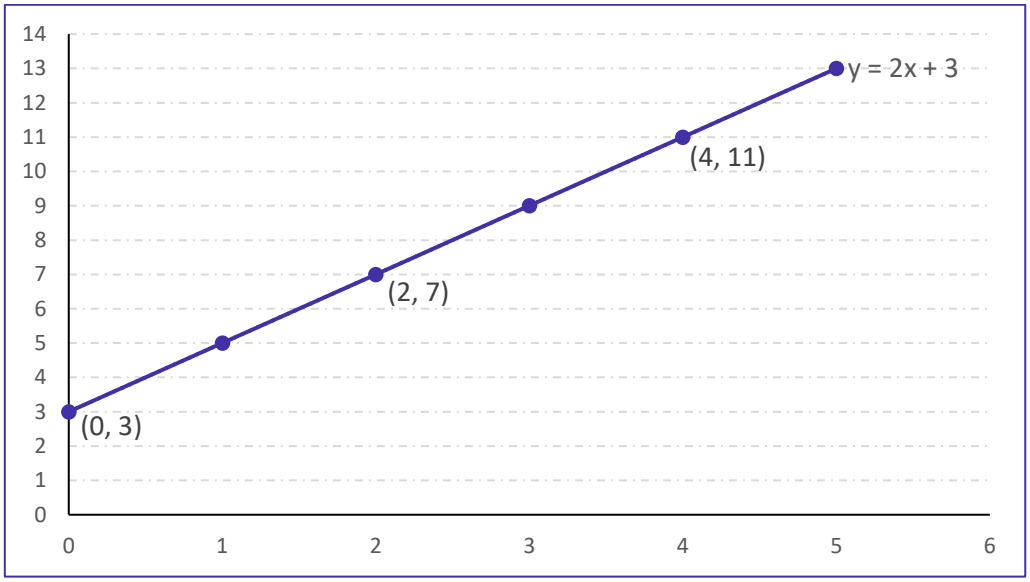
$$7 = 4 + b$$

$$b = 3$$

Logo, a expressão que representa nossa reta nesse exemplo é:

$$y = 2 \cdot x + 3$$

Como $b = 3$, a reta intercepta o eixo y no ponto (0, 3). Vejamos:



Nesse caso temos uma correlação perfeita positiva. O que aconteceria se os pontos do gráfico tivessem um pouco mais dispersos, não evidenciando uma correlação linear perfeita? Será se conseguiríamos determinar uma reta que se ajustasse a esse tipo de gráfico? A resposta é sim. Basta fazermos um pequeno ajuste na expressão usada para determinar a reta de regressão.

Observe:

$$\boldsymbol{Y_i = \alpha + \beta X_i + \varepsilon_i}$$

Em que $i = 1, 2, 3, \cdots, n$.

O termo $\alpha + \beta X_i$ é o componente de Y_i que varia linearmente, de acordo com X_i. Por sua vez, ε_i é o componente aleatório de Y_i que descreve os erros (ou desvios) cometidos quando tentamos aproximar uma série de observações X_i por meio de uma reta Y_i.

Nesse modelo, Y_i é a variável cujo comportamento desejamos prever ou explicar, sendo chamada de variável **dependente** ou **resposta**. Por outro lado, a variável X_i é utilizada para explicar o comportamento de Y_i, sendo conhecida como **independente**, **regressora**, **explanatória** ou **explicativa**.

O modelo de regressão linear requer que sejam atendidos alguns pressupostos básicos quanto à variável aleatória ε_i (erro ou desvio):

i) $\boldsymbol{E(\varepsilon_i) = 0}$**.** A média dos erros é igual a zero. Ou seja, os desvios "para cima da reta" igualam o valor dos desvios "para baixo da reta" na média.

ii) $\boldsymbol{Var(\varepsilon_i) = \sigma^2}$**.** A variância dos erros é constante. Essa propriedade é denominada de **homocedasticia**. Isso só é possível se a variável ε_i tiver variância constante. Ou seja, se ela tiver sempre a mesma variância, independente de qual o valor de X_i. Quando o modelo apresenta variâncias diferentes para o erro, temos uma situação de **heterocedasticia.**

iii) $\boldsymbol{Cov(\varepsilon_i, \varepsilon_j) = 0\ para\ i \neq j}$**.** Os erros cometidos não são correlacionados, isto é, **os desvios ε_i são variáveis aleatórias independentes**. Quando os erros não são independentes, temos uma situação denominada de **autocorrelação**.

(CESPE 2019/TJ-AM) Um estudo considerou um modelo de regressão linear simples na forma $y = 0,8x + b + \epsilon$, em que y é a variável dependente, x representa a variável explicativa do modelo, o coeficiente b denomina-se intercepto e ϵ é um erro aleatório que possui média nula e desvio padrão σ. Sabe-se que a variável y segue a distribuição normal padrão e que o modelo apresenta coeficiente de determinação R^2 igual a 85%.

Com base nessas informações, julgue o item que se segue.

O erro aleatório ϵ segue a distribuição normal padrão.

Comentários:

No modelo de regressão linear simples, as seguintes suposições sobre o erro devem ser observadas:

$E(\epsilon) = 0$, isto é, em média, o erro do modelo deve ser 0;

$Var(\epsilon) = \sigma^2$, a variância deve ser constante, isto é, deve existir homocedasticia;

$Cov(\epsilon i, \epsilon j) = 0$, os erros devem ser independentes, ou seja, não há correlação entre os erros.

Nessa questão, o único ponto que precisamos mostrar é que o $Var(\boldsymbol{\epsilon}) = 1$. O enunciado afirmou que Y segue distribuição normal padrão. De fato, Y tem distribuição $N(b + 0{,}8x + \mu; \sigma^2) = N(0{,}1)$ em que σ^2 é a variância do erro. Como Y segue uma normal padrão, então $\sigma^2 = 1$. Consequentemente, o erro também seguirá uma distribuição normal, $\epsilon \sim N(0{,}1)$.

Gabarito: Certo.

Método dos Mínimos Quadrados

O método dos mínimos quadrados diz que a reta a ser adotada deverá ser aquela que torna mínima a soma dos quadrados das distâncias da reta aos pontos experimentais, medidas no sentido da variação aleatória. Em outras palavras, devemos encontrar uma reta que minimize o somatório dos quadrados das distâncias ($\sum_{i=1}^{n} e_i^2$). O objetivo é minimizar a soma dos quadrados dos desvios.

Esse método é empregado na obtenção dos estimadores α e β de um modelo de regressão linear:

$$Y_i = \alpha + \beta X_i + \varepsilon_i.$$

A expressão usada para determinar a reta de regressão é:

$$\hat{Y}_i = a + bX_i$$

em que a e b são as estimativas dos parâmetro α e β, respectivamente.

Os erros (desvios) resultantes da aplicação do modelo de regressão linear correspondem às diferenças entre os valores observados e os valores estimados:

$$e_i = Y_i - \hat{Y}_i$$

O objetivo do método dos mínimos quadrados é minimizar o somatório dos quadrados dos desvios ($\sum_{i=1}^{n} e_i^2$):

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2$$

Por esse método, o valor de b é dado por:

$$\boldsymbol{b = \frac{\sum_{i=1}^{n}[(X_i - \bar{X}) \times (Y_i - \bar{Y})]}{\sum_{i=1}^{n}[(X_i - \bar{X})^2]}}$$

Existem outras formas mais simples de calcular o valor de b.

Para o numerador da fórmula, temos:

$$\sum_{i=1}^{n}[(X_i - \bar{X}) \times (Y_i - \bar{Y})] = \sum_{i=1}^{n}(X_i \times Y_i) - n \times \bar{X} \times \bar{Y}$$

Para o denominador da fórmula, temos:

$$\sum_{i=1}^{n}[(X_i - \bar{X})^2] = \sum_{i=1}^{n}(X_i^2) - n \times \bar{X}^2$$

Logo,

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

A reta de regressão passa pelos pontos médios $(\bar{X}, \bar{Y})$ das variáveis X e Y. Isso implica que o valor de a pode ser calculado substituindo o valor de b em:

$$\boldsymbol{a = \bar{Y} - b\bar{X}}$$

Vejamos um exemplo. A tabela a seguir apresenta as notas de 5 alunos nas disciplinas X e Y.

Aluno	X	Y	$X - \bar{X}$	$Y - \bar{Y}$	$(X - \bar{X}) \times (Y - \bar{Y})$	$(X - \bar{X})^2$	$(Y - \bar{Y})^2$
1	5	9	-2	1	-2	4	1
2	5	8	-2	0	0	4	0
3	8	10	1	2	2	1	4
4	8	7	1	-1	-1	1	1
5	9	6	2	-2	-4	4	4
Média	**7**	**8**	**Total**		**-5**	**14**	**10**

Calculando o valor de b:

$$b = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2}$$

$$b = \frac{-5}{14}$$

$$b \cong -0{,}357$$

O valor de a é calculado por:

$$a = \bar{Y} - b\bar{X}$$

$$a = 8 - (-0{,}357) \times 7$$

$$a = 8 + 2{,}499$$

$$a = 10{,}499$$

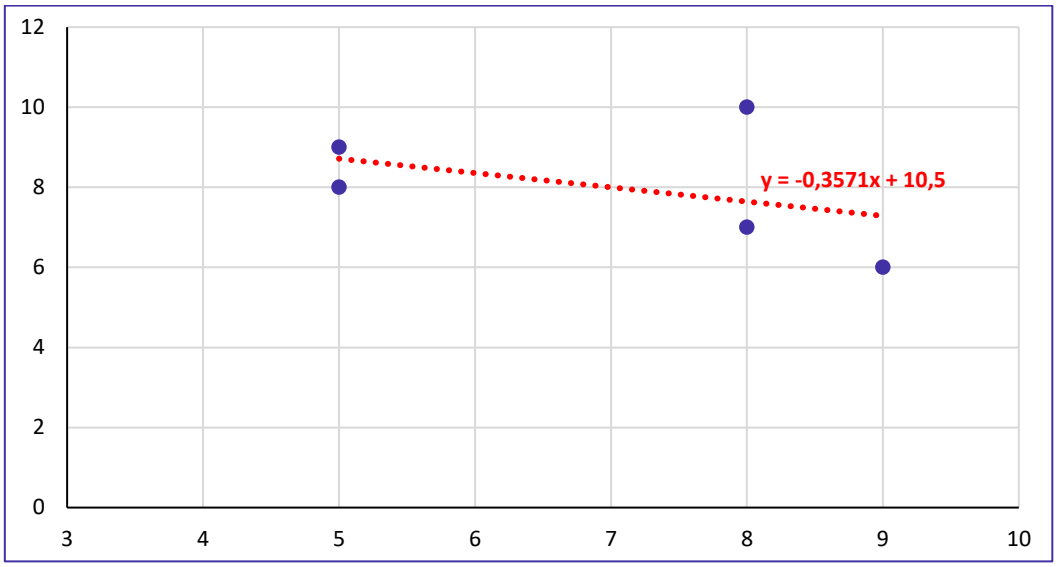
Assim, a reta de regressão estimada é:

$$\hat{Y} = 10{,}499 - 0{,}357 \times X$$

Temos dessa reta estimada que a soma dos quadrados dos desvios é mínima. Podemos montar uma nova tabela contendo os valores observados de (Y) e os valores estimados pela reta ($\hat{Y}$):

Aluno	X	Y	$\hat{Y}$
1	5	9	8,714
2	5	8	8,714
3	8	10	7,643
4	8	7	7,643
5	9	6	7,286

Montando o gráfico com os valores estimados, temos:



Percebam que a reta tem uma correlação negativa. Os pontos azuis são os pares ordenados da amostra e a reta vermelha é a reta de regressão que calculamos de tal forma que os desvios de estimativa cometidos se comportem segundo a condição de mínimos quadrados.

O coeficiente b pode ser calculado por meio da seguinte expressão:

$$b = \frac{S_{XY}}{S_{XX}}$$

Em que $S_{XY} = \sum_{i=1}^{n}[(X_i - \bar{X}) \times (Y_i - \bar{Y})]$ e $S_{XX} = \sum_{i=1}^{n}(X_i - \bar{X})^2$.

HORA DE PRATICAR!

(VUNESP 2019/MPE-SP). Um aluno teve as seguintes notas: 3; 5; 5,5; 6,5. O professor quer atribuir a nota final, escolhendo uma nota representativa desse conjunto com base no método dos mínimos quadrados. Desse modo, essa nota final será

a) 4.

b) 4,5.

c) 5.

d) 5,5.

e) 6.

Comentários:

Vamos montar uma tabela com os dados fornecidos:

X	Y	$(X-\bar{X})$	$(Y-\bar{Y})$	$(X-\bar{X})(Y-\bar{Y})$	$(X-\bar{X})^2$
1	3	-1,5	-2	3	2,25
2	5	-0,5	0	0	0,25
3	5,5	0,5	0,5	0,25	0,25
4	6,5	1,5	1,5	2,25	2,25
$\bar{X}=2{,}5$	$\bar{Y}=5$	Total		5,5	5

Pelo método dos mínimos quadrados, temos:

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

$$b = \frac{\sum(X_i-\bar{X})(Y_i-\bar{Y})}{\sum(X_i-\bar{X})^2}$$

$$b = \frac{5{,}5}{5} = 1{,}1$$

$$a = \bar{Y} - b\bar{X}$$

$$a = 5 - 1{,}1 \times 2{,}5$$

$$a = 2{,}25$$

Nossa reta de regressão será:

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

$$\hat{Y} = 2{,}25 + 1{,}1X$$

Sabendo que a reta de regressão passa pelo ponto $(\bar{X}, \bar{Y})$, então:

$$\hat{Y} = 2{,}25 + 1{,}1 \times 2{,}5$$

$$\hat{Y} = 5$$

Gabarito: C.

(CESPE 2018/ABIN) Ao avaliar o efeito das variações de uma grandeza X sobre outra grandeza Y por meio de uma regressão linear da forma $\hat{Y} = \hat{\alpha} + \hat{\beta}X$, um analista, usando o método dos mínimos quadrados, encontrou, a partir de 20 amostras, os seguintes somatórios (calculados sobre os vinte valores de cada variável):

$$\boldsymbol{\sum X = 300; \sum Y = 400; \sum X^2 = 6.000; \sum Y^2 = 12.800 \; e \; \sum (XY) = 8.400}$$

A partir desses resultados, julgue o item a seguir.

Para $X = 10$, a estimativa de Y é $\hat{Y} = 12$.

Comentários:

Inicialmente, vamos calcular os valores de $\bar{Y}$ e de $\bar{X}$:

$$\bar{Y} = \frac{\sum y}{n} = \frac{400}{20} = 20$$

$$\bar{X} = \frac{\sum x}{n} = \frac{300}{20} = 15$$

Agora, utilizaremos o método dos mínimos quadrados para determinar $\hat{\beta}$:

$$\hat{\beta} = \frac{\sum X_i Y_i - n\bar{X}\bar{Y}}{\sum X_i^2 - n\bar{X}^2}$$

$$\hat{\beta} = \frac{8400 - 20 \times 15 \times 20}{6000 - 20 \times 15^2}$$

$$\hat{\beta} = \frac{2400}{1500} = 1{,}6$$

Conhecendo $\hat{\beta}$, podemos determinar o valor de $\hat{\alpha}$:

$$\hat{\alpha} = \frac{\sum Y_i - \hat{\beta} \sum X_i}{n}$$

$$\hat{\alpha} = 20 - 1{,}6 \times 15 = -4$$

Assim, o modelo de regressão é dado por:

$$\hat{Y} = -4 + 1{,}6X$$

Para $X = 10$, temos o seguinte valor de $\hat{Y}$:

$$\hat{Y} = -4 + 1{,}6 \times 10$$

$$\hat{Y} = 12$$

Gabarito: Certo.

(FCC 2018/Pref. São Luís) Analisando um gráfico de dispersão referente a 10 pares de observações (t, Y_t) com $t = 1, 2, 3, \ldots, 10$, optou-se por utilizar o modelo linear $Y_t = \alpha + \beta t + \varepsilon_t$ com o objetivo de se prever a variável Y, que representa o faturamento anual de uma empresa em milhões de reais, no ano $(2007 + t)$. Os parâmetros α e β são desconhecidos e ε_t é o erro aleatório com as respectivas hipóteses do modelo de regressão linear simples. As estimativas de α e β (a e b, respectivamente) foram obtidas por meio do método dos mínimos quadrados com base nos dados dos 10 pares de observações citados. Se $a = 2$ e a soma dos faturamentos dos 10 dados observados foi de 64 milhões de reais, então, pela equação da reta obtida, a previsão do faturamento para 2020 é, em milhões de reais, de

a) 11,6

b) 15,0

c) 13,2

d) 12,4

e) 14,4

Comentários:

A reta calculada é expressa por:

$$\hat{Y} = a + b \times t$$

Sabemos que a soma dos faturamentos dos 10 dados observados foi de 64 milhões de reais, então, calculando a média temos:

$$\bar{Y} = \frac{64}{10} = 6{,}4.$$

Agora, vamos calcular a média de t:

$$\bar{t} = \frac{1 + 2 + 3 + 4 + 5 + 6 + 7 + 8 + 9 + 10}{10} = 5{,}5$$

Sabemos que $a = 2$ e que a reta de regressão passa pelo ponto $(\bar{t}, \bar{Y})$. Portanto, vamos encontrar o valor de b:

$$\bar{Y} = a + b\bar{t}$$

$$6{,}4 = 2 + b \times 5{,}5$$

$$b = \frac{4{,}4}{5{,}5}$$

$$b = 0{,}8$$

A reta fica assim:

$$\hat{Y} = 2 + 0{,}8t$$

Em 2020, temos que $t = 13$, pois $2020 = 2007 + 13$. Logo:

$$\hat{Y} = 2 + 0{,}8 \times 13$$

$$\hat{Y} = 12{,}4$$

Gabarito: D.

Reta Passando pela Origem

Em determinadas situações, a reta de regressão deve **passar pela origem** para que consiga se ajustar adequadamente ao modelo teórico. Quando isso ocorre, temos uma situação em que o **coeficiente linear da reta de regressão é nulo ($\alpha = 0$)**.

Nesse caso, o modelo de regressão que passa obrigatoriamente pela origem é:

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

$$Y_i = 0 + \beta X_i + \varepsilon_i$$

$$\boldsymbol{Y_i = \beta X_i + \varepsilon_i.}$$

Em que X_i é a variável independente ou explicativa; Y_i é a variável dependente ou resposta; ε_i representa os erros aleatórios e β é o parâmetro populacional a ser estimado.

Assim, a estimativa de β, pelo método dos mínimos quadrados, é:

$$\boldsymbol{b = \frac{\sum X_i \times Y_i}{\sum {X_i}^2}}$$

A reta de regressão ajustada é:

$$\boldsymbol{\hat{Y}_i = bX_i.}$$

Os desvios ou resíduos são dados por:

$$\boldsymbol{e_i = Y_i - \hat{Y}_i.}$$

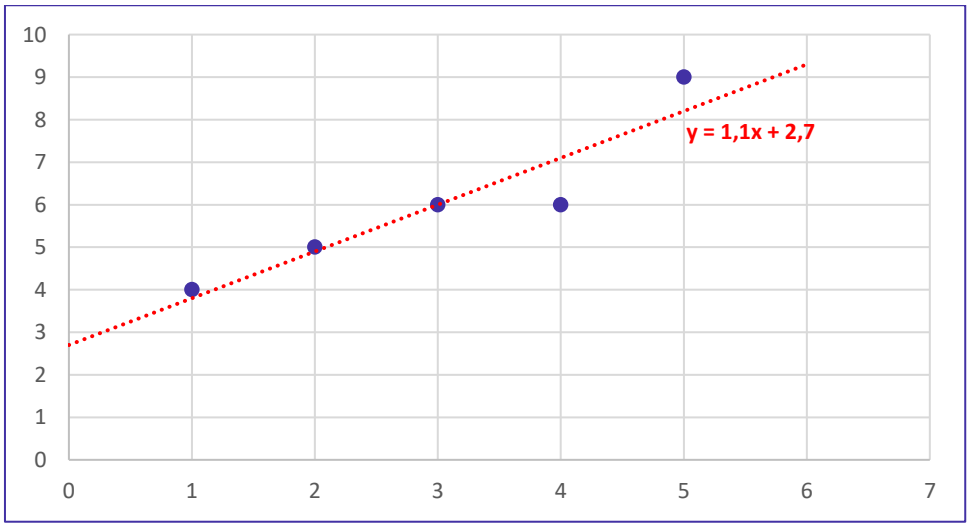
Nesse caso, não há garantia de que o somatório dos resíduos seja igual a zero.

EXEMPLIFICANDO

Calcular a reta que passa pela origem e comparar os desvios dessa abordagem com os desvios do modelo linear tradicional para os dados abaixo:

i	X	Y
1	1	4
2	2	5
3	3	6
4	4	6
5	5	9

Se utilizássemos o modelo de regressão linear tradicional, a reta que iríamos obter seria a seguinte:



Agora, vamos calcular a reta de regressão que passa pela origem e comparar com o modelo tradicional. Para isso, adicionaremos duas colunas à tabela original:

i	X	Y	$X \times Y$	X^2
1	1	4	4	1
2	2	5	10	4
3	3	6	18	9
4	4	6	24	16
5	5	9	45	25
Total			**101**	**55**

De posse dos totais dessas duas colunas, podemos estimar o valor de β pelo método dos mínimos quadrados:

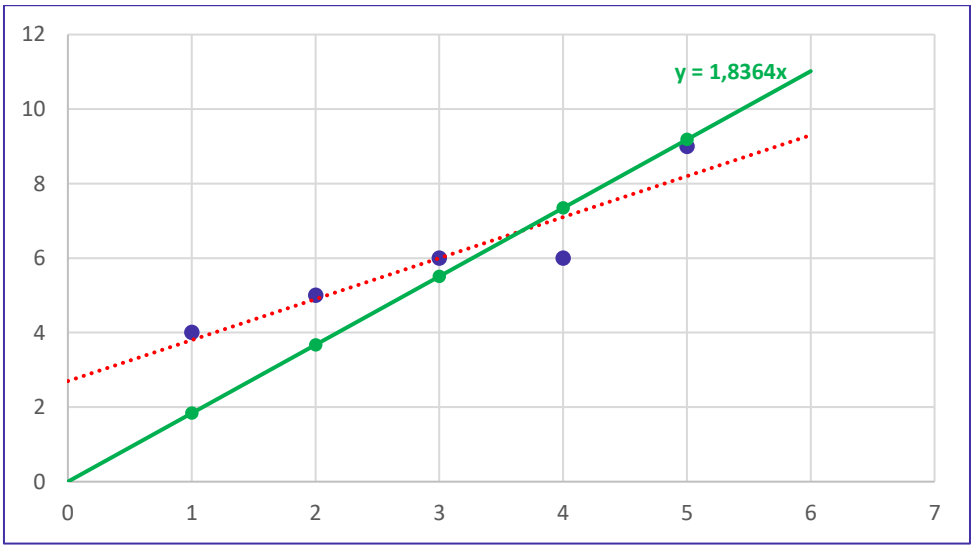
$$b = \frac{\sum X_i \times Y_i}{\sum {X_i}^2} = \frac{101}{55} \cong 1{,}836$$

Portanto, a reta de regressão ajustada é:

$$\widehat{Y_i} = bX_i$$

$$\widehat{Y_i} = 1{,}836 \times X_i$$

Vejamos como esse modelo se comporta em relação ao modelo tradicional:



Vamos, agora, comparar os resíduos da abordagem tradicional com os desvios do modelo que passa pela origem:

i	Modelo tradicional			Modelo que passa pela origem		
	Y_i	$\widehat{Y}_i$	$e_i = Y_i - \widehat{Y}_i$	Y_i	$\widehat{Y}_i$	$e_i = Y_i - \widehat{Y}_i$
1	4	3,8	0,2	4	1,8364	2,1636
2	5	4,9	0,1	5	3,6727	1,3273
3	6	6	0	6	5,5091	0,4909
4	6	7,1	-1,1	6	7,3455	-1,3455
5	9	8,2	0,8	9	9,1818	-0,1818
Total			**0**	**Total**		**2,4545**

Reparem que, no modelo que passa pela origem, não há garantia de que o somatório dos desvios seja zero.

ANÁLISE DE VARIÂNCIA DA REGRESSÃO

A estratégia que adotamos para verificar se compensa ou não utilizar um modelo de regressão linear, $Y_i = \alpha + \beta X_i + \varepsilon_i$, é observar a redução no resíduo (desvio) quando comparado com um modelo aproximadamente uniforme $Y_i = \mu + \varepsilon_i$.

Se a redução é muito pequena, significa dizer que os dois modelos são praticamente equivalentes. Isso ocorre quando a inclinação β for zero ou um valor muito pequeno, não compensando usar um modelo mais complexo. Assim, estamos interessados em testar a hipótese:

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

Se a hipótese nula é aceita, concluímos que não existe relação linear significativa entre as variáveis X e Y.

O resultado da **análise de variância da regressão** é uma tabela que resume várias medidas usadas no teste de hipóteses anterior. Para montar a tabela de análise de variância (ANOVA), precisamos conhecer: os graus de liberdade, as somas dos quadrados e os quadrados médios do modelo, dos resíduos (erros ou desvios) e total.

A seguir, veremos como construir a tabela de análise de variância da regressão.

Graus de Liberdade

O número total de graus de liberdade de uma amostra de tamanho n é:

$$GL_{Total} = n - 1$$

Como vimos anteriormente, a equação de regressão possui apenas dois parâmetros (α e β). Portanto, o número de graus de liberdade do modelo é:

$$GL_{Modelo} = 2 - 1 = 1$$

Agora, temos que descobrir o número de graus de liberdade dos resíduos. Para isso, utilizamos a seguinte relação:

$$GL_{Total} = GL_{Modelo} + GL_{Resíduos}$$

Daí, concluímos que:

$$n - 1 = 1 + GL_{Resíduos}$$

$$GL_{Resíduos} = n - 2$$

ATENÇÃO
DECORE!

O número de graus de liberdade do modelo de regressão é:

$$GL_{Modelo} = 2 - 1 = 1$$

O número de graus de liberdade dos resíduos é:

$$GL_{Resíduos} = n - 2$$

O número de graus de liberdade total é:

$$GL_{Total} = n - 1$$

Somas de Quadrados

Como vimos, a reta de regressão linear fornece uma estimativa $\hat{Y}_i$ para uma variável Y_i. Os erros (desvios) resultantes da aplicação do modelo de regressão linear correspondem às diferenças entre os valores observados e os valores estimados:

$$e_i = Y_i - \hat{Y}_i \Rightarrow Y_i = e_i + \hat{Y}_i$$

Subtraindo $\bar{Y}$ dos dois lados, temos:

$$Y_i - \bar{Y} = e_i + \hat{Y}_i - \bar{Y}$$

Agora, elevando os dois lados ao quadrado:

$$(Y_i - \bar{Y})^2 = (e_i + \hat{Y}_i - \bar{Y})^2$$

$$(Y_i - \bar{Y})^2 = e_i^2 + \left(\hat{Y}_i - \bar{Y}\right)^2 + 2 \times e_i \times (\hat{Y}_i - \bar{Y})$$

Somando tudo, temos:

$$\sum (Y_i - \bar{Y})^2 = \sum e_i^2 + \sum \left(\hat{Y}_i - \bar{Y}\right)^2 + 2 \times \sum e_i \times (\hat{Y}_i - \bar{Y})$$

É possível demonstrar que :

$$2 \times \sum e_i \times (\hat{Y}_i - \bar{Y}) = 0$$

Logo,

$$\boxed{\sum (Y_i - \bar{Y})^2 = \sum e_i^2 + \sum \left(\hat{Y}_i - \bar{Y}\right)^2.}$$

Portanto, temos que o **desvio total do modelo de regressão**, $(Y_i - \bar{Y})$, é o desvio de cada valor de Y_i em relação à média $\bar{Y}$.

$$\boxed{SQT = \sum_{i=1}^{n} (Y_i - \bar{Y})^2}$$

Assim, a **soma dos quadrados totais**, definida por $\sum (Y_i - \bar{Y})^2$, é igual a soma dos quadrados dos resíduos/desvios/erros, definida por $\sum \varepsilon_i^2$, mais a soma dos quadrados do modelo de regressão, definida na expressão por $\sum \left(\hat{Y}_i - \bar{Y}\right)^2$:

$$\boxed{SQT = SQM + SQR}$$

A parcela do desvio total que o modelo de regressão é capaz de explicar é denominada de **"desvio explicável"**. Essa parcela corresponde à a diferença entre cada valor previsto pelo modelo $(\hat{Y}_i)$ e o valor médio $(\bar{Y})$. Assim, a **soma dos quadrados do modelo de regressão** é:

$$\boxed{SQM = \sum_{i=1}^{n} \left(\hat{Y}_i - \bar{Y}\right)^2}$$

Podemos demonstrar que:

$$\boxed{SQM = b \times \sum_{i=1}^{n} [(X_i - \bar{X}) \times (Y_i - \bar{Y})]}$$

Em que b é a estimativa do coeficiente angular da reta de regressão.

A fórmula a seguir também pode ser utilizada para o cálculo de SQM:

$$SQM = b^2 \times \sum_{i=1}^{n} (X_i - \bar{X})^2$$

A parcela do desvio total que o modelo de regressão não é capaz de explicar é chamada de **"desvio não explicável"**. Essa parcela corresponde à diferença entre cada valor de Y_i e o valor previsto pelo modelo $\hat{Y}_i$. Assim, podemos definir a **soma dos quadrados dos erros (resíduos)**.

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

Em algumas questões de concurso, a notação SQR é utilizada para representar a **soma dos quadrados do modelo de regressão**, e não dos resíduos, como fizemos nesta aula. Na maioria das questões, contudo, SQR representa a **soma dos quadrados dos resíduos (erros).**

A **soma dos quadrados totais** é calculada por meio das seguintes fórmulas:

$$SQT = SQM + SQR$$

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

A **soma dos quadrados do modelo** de regressão é calculada mediante as seguintes fórmulas:

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

$$SQM = b \times \sum_{i=1}^{n} [(X_i - \bar{X}) \times (Y_i - \bar{Y})]$$

$$SQM = b^2 \times \sum_{i=1}^{n} (X_i - \bar{X})^2$$

A **soma dos quadrados dos resíduos** é calculada pela fórmula:

$$SQR = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2$$

Coeficiente de Determinação

O coeficiente de determinação mede a qualidade do ajuste proporcionado pela reta de regressão. Ele determina a parcela da variação total de Y que é explicada pelo modelo de regressão, sendo calculado pela fórmula:

$$\boxed{R^2 = \frac{SQM}{SQT}}$$

em que R é o coeficiente de correlação linear, calculado pela expressão:

$$R = \sqrt{\frac{SQM}{SQT}}$$

O coeficiente de determinação também pode ser escrito da seguinte forma:

$$\boxed{R^2 = \frac{SQM}{SQT} = \frac{SQT - SQR}{SQT} = 1 - \frac{SQR}{SQT}}$$

A análise que queremos fazer segue o mesmo raciocínio do coeficiente de correlação, assim temos que:

$$0 \leq R^2 \leq 1$$

Portanto, quanto mais próximo de 1 estiver o coeficiente de determinação, mais forte será a correlação entre as variáveis. Implica dizer que grande parte da variação de Y é explicada pela modelo de regressão linear, ou seja, a reta de regressão é capaz de explicar as diferenças entre os valores observados (Y_i) e a média ($\bar{Y}$).

Por outro lado, quanto mais próximo de 0 estiver o coeficiente de determinação, mais fraca será a correlação linear entre as variáveis. Significa dizer que grande parte da variação de Y não é explicada pelo modelo de regressão, ou seja, a reta de regressão é capaz de explicar muito pouco sobre as diferenças entre os valores observados (Y_i) e a média ($\bar{Y}$).

Coeficiente de Determinação Ajustado

O coeficiente de determinação ajustado é mais utilizado quando estamos tratando de regressão múltipla. Contudo, esse assunto também tem sido abordado em algumas questões de regressão linear simples. Assim, é importante conhecermos essa medida.

Basicamente, essa medida ajusta o coeficiente de determinação aos graus de liberdade. Ela é obtida pela divisão de SQR e SQT pelos respectivos graus de liberdade:

$$\overline{R^2} = 1 - \frac{SQR/(n-2)}{SQT/(n-1)}$$

A relação entre o coeficiente de determinação ajustado ($\overline{R^2}$) e o coeficiente de determinação tradicional (R^2) é dada por:

$$\overline{R^2} = 1 - (1 - R^2) \times \frac{(n-1)}{(n-2)}$$

Quadrados Médios

Os quadrados médios são obtidos pela divisão entre as somas dos quadrados e os respectivos graus de liberdade. Assim, temos:

a) quadrado médio do modelo (QMM):

$$QMM = \frac{SQM}{1}$$

b) quadrado médio dos resíduos (QMR):

$$QMR = \frac{SQR}{n-2}$$

c) quadrado médio total (QMT):

$$QMT = \frac{SQT}{n-1}$$

O quadrado médio dos resíduos (QMR) corresponde à estimativa da variância σ^2 residual.

Estatística F (Razão F)

Para testar $H_0: \beta = 0$ contra $H_1: \beta \neq 0$, usamos a seguinte estatística teste, denominada de estatística F (ou razão F):

$$F^* = \frac{QMM}{QMR}$$

Se o valor de F^* for significativamente grande, teremos evidências para rejeitar H_0.

Sob a hipótese H_0, F^* tem distribuição F de Snedecor, com 1 e $n-2$ graus de liberdade, em que n é o número de observações.

Dessa forma, para avaliar o teste de hipóteses, basta compararmos o valor da estatística teste com o valor crítico tabelado:

- Se $F^* > F_{crítico}$, podemos rejeitar a hipótese nula;
- Se $F^* < F_{crítico}$, não podemos rejeitar a hipótese nula.

O valor de $F_{crítico}$ é consultado em uma tabela F de Snedecor com 1 grau de liberdade no numerador e $n-2$ graus de liberdade no denominador, para um determinado nível de significância.

Tabela de Análise de Variância da Regressão

Em geral, as questões de **análise de variância da regressão** fornecem uma tabela incompleta e pedem alguma medida que está faltando. Para descobrir o valor da medida solicitada, você deve conhecer a estrutura da tabela e as fórmulas apresentadas neste tópico. A estrutura da tabela de análise de variância da regressão sempre terá o seguinte formato:

Fonte de Variação	Graus de Liberdade	Soma dos Quadrados	Quadrados Médios	Estatística F (Razão F)
Modelo	1	SQM	$QMM = \frac{SQM}{1}$	$F^* = \frac{QMM}{QMR}$
Resíduos	$n-2$	SQR	$QMR = \frac{SQR}{n-2}$	
Total	$\boldsymbol{n-1}$	$\boldsymbol{SQT}$	$\boldsymbol{QMT = \frac{SQT}{n-1}}$	

(CESPE/EBSERH/2018) Determinado estudo considerou um modelo de regressão linear simples na forma $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, em que y_i representa o número de leitos por habitante existente no município i; x_i representa um indicador de qualidade de vida referente a esse mesmo município i, para $i = 1, \cdots, n$. A componente ε_i representa um erro aleatório com média 0 e variância σ^2. A tabela a seguir mostra a tabela ANOVA resultante do ajuste desse modelo pelo método dos mínimos quadrados ordinários.

Fonte de Variação	Soma dos Quadrados	Graus de Liberdade	Média dos Quadrados	Razão F	P-valor
Modelo	**900**	**1**	**900**	**90**	**< 0,001**
Erro	**100**	**10**	**10**		
Total	**1.000**	**11**			

A partir das informações e da tabela apresentadas, julgue os itens subsequentes.

O referido estudo contemplou um conjunto de dados obtidos de $n = 11$ municípios.

Comentários:

Na análise de variância (ANOVA) da regressão, o total de graus de liberdade corresponde a $n - 1$, em que n representa o número total de amostras. Logo, podemos estabelecer que:

$$n - 1 = 11$$

$$n = 12\ munic\acute{i}pios.$$

Gabarito: Errado.

(CESPE/EBSERH/2018) Determinado estudo considerou um modelo de regressão linear simples na forma $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, em que y_i representa o número de leitos por habitante existente no município i; x_i representa um indicador de qualidade de vida referente a esse mesmo município i, para $i = 1, \cdots, n$. A componente ε_i representa um erro aleatório com média 0 e variância σ^2. A tabela a seguir mostra a tabela ANOVA resultante do ajuste desse modelo pelo método dos mínimos quadrados ordinários.

Fonte de Variação	Soma dos Quadrados	Graus de Liberdade	Média dos Quadrados	Razão F	P-valor
Modelo	900	1	900	90	< 0,001
Erro	100	10	10		
Total	1.000	11			

A partir das informações e da tabela apresentadas, julgue os itens subsequentes.

A correlação linear entre o número de leitos hospitalares por habitante (y) e o indicador de qualidade de vida (x) foi igual a 0,9.

Comentários:

O coeficiente de correlação linear entre as variáveis X e Y é calculado por meio da seguinte expressão:

$$R = \sqrt{\frac{SQR}{SQT}},$$

em que SQR indica a soma dos quadrados da regressão (modelo) e SQT a soma dos quadrados totais.

Pela tabela, verificamos que $SQT = 1000$ e $SQR = 900$. Substituindo esses valores na equação anterior, teremos:

$$R = \sqrt{\frac{900}{1000}} = \sqrt{0{,}9}$$

Portanto, o coeficiente de determinação R^2 possui valor igual a 0,9, mas o coeficiente de correlação não.

Gabarito: Errado.

(CESPE/EBSERH/2018) Determinado estudo considerou um modelo de regressão linear simples na forma $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, em que y_i representa o número de leitos por habitante existente no município i; x_i representa um indicador de qualidade de vida referente a esse mesmo município i, para $i = 1, \cdots, n$. A componente ε_i representa um erro aleatório com média 0 e variância σ^2. A tabela a seguir mostra a tabela ANOVA resultante do ajuste desse modelo pelo método dos mínimos quadrados ordinários.

Fonte de Variação	Soma dos Quadrados	Graus de Liberdade	Média dos Quadrados	Razão F	P-valor
Modelo	900	1	900	90	< 0,001
Erro	100	10	10		
Total	1.000	11			

A partir das informações e da tabela apresentadas, julgue os itens subsequentes.

A razão F da tabela ANOVA refere-se ao teste de significância estatística do intercepto β_0, em que se testa a hipótese nula $H_0: \beta_0 = 0$ contra a hipótese alternativa $H_A: \beta_0 \neq 0$.

Comentários:

A estatística $F = \frac{QMM}{QMR}$ está relacionada com o teste de hipótese para o coeficiente angular β da reta de regressão, isto é:

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

Se a hipótese H_0 não é rejeitada, significa dizer que não existe uma relação linear significativa entre a variável explicativa (X) e a variável dependente (Y).

Gabarito: Errado.

(CESPE/EBSERH/2018) Determinado estudo considerou um modelo de regressão linear simples na forma $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, em que y_i representa o número de leitos por habitante existente no município i; x_i representa um indicador de qualidade de vida referente a esse mesmo município i, para $i = 1, \cdots, n$. A componente ε_i representa um erro aleatório com média 0 e variância σ^2. A tabela a seguir mostra a tabela ANOVA resultante do ajuste desse modelo pelo método dos mínimos quadrados ordinários.

Fonte de Variação	Soma dos Quadrados	Graus de Liberdade	Média dos Quadrados	Razão F	P-valor
Modelo	900	1	900	90	< 0,001
Erro	100	10	10		
Total	1.000	11			

A partir das informações e da tabela apresentadas, julgue os itens subsequentes.

O desvio padrão amostral do número de leitos por habitante foi superior a 10 leitos por habitante.

Comentários:

A soma dos quadrados totais (SQT) é dada por:

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

A variância amostral é calculada por:

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

Pela tabela, o grau de liberdade do total corresponde a 11, então:

$$n - 1 = 11$$

Logo, a variância amostral é:

$$\frac{\sum_{i=1}^{n} (Y_i - \bar{Y})^2}{n-1} = \frac{SQT}{11} = \frac{1000}{11} = 90{,}90$$

Como a variância amostral é menor que 100, o desvio padrão amostral será:

$$\sqrt{90{,}90} < \sqrt{100}$$

$$\sqrt{90{,}90} < 10$$

Gabarito: Errado.

(CESPE/EBSERH/2018) Determinado estudo considerou um modelo de regressão linear simples na forma $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, em que y_i representa o número de leitos por habitante existente no município i; x_i representa um indicador de qualidade de vida referente a esse mesmo município i, para $i = 1, \cdots, n$. A componente ε_i representa um erro aleatório com média 0 e variância σ^2. A tabela a seguir mostra a tabela ANOVA resultante do ajuste desse modelo pelo método dos mínimos quadrados ordinários.

Fonte de Variação	Soma dos Quadrados	Graus de Liberdade	Média dos Quadrados	Razão F	P-valor
Modelo	900	1	900	90	< 0,001
Erro	100	10	10		
Total	1.000	11			

A partir das informações e da tabela apresentadas, julgue os itens subsequentes.

A estimativa de σ^2 foi igual a 10.

Comentários:

A estimativa de σ^2 equivale ao quadrado médio residual. Logo,

$$\sigma^2 = QMR = 10$$

Gabarito: Certo.

(CESPE/EBSERH/2018) Determinado estudo considerou um modelo de regressão linear simples na forma $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, em que y_i representa o número de leitos por habitante existente no município i; x_i representa um indicador de qualidade de vida referente a esse mesmo município i, para $i = 1, \cdots, n$. A componente ε_i representa um erro aleatório com média 0 e variância σ^2. A tabela a seguir mostra a tabela ANOVA resultante do ajuste desse modelo pelo método dos mínimos quadrados ordinários.

Fonte de Variação	Soma dos Quadrados	Graus de Liberdade	Média dos Quadrados	Razão F	P-valor
Modelo	**900**	**1**	**900**	**90**	**< 0,001**
Erro	**100**	**10**	**10**		
Total	**1.000**	**11**			

A partir das informações e da tabela apresentadas, julgue os itens subsequentes.

O R^2 ajustado (*Adjusted R Square*) foi inferior a 0,9.

Comentários:

O coeficiente de determinação permite avaliar a qualidade do ajuste do modelo, quantificando, basicamente, a capacidade do modelo de explicar os dados coletados. Ele é calculado por meio da expressão:

$$R^2 = \frac{SQM}{SQT} = 1 - \frac{SQR}{SQT},$$

em que SQM = Soma dos quadrados da regressão (modelo), SQR = Soma dos quadrados dos resíduos (erros) e SQT = Soma dos quadrados totais. Além disso, para evitar dificuldades na interpretação de R^2, alguns estatísticos preferem usar o R^2 ajustado, definido para uma equação com 2 coeficientes como

$$\overline{R^2} = 1 - \left(\frac{n-1}{n-2}\right) \times (1 - R^2).$$

Pela tabela temos que $SQT = 1000$ e $SQR = 900$. Substituindo os valores apresentados na tabela nas equações acima teremos:

$$R^2 = \frac{900}{1000} = 0{,}9.$$

Além disso, como temos $n - 1 = 11$ graus de liberdade totais, então

$$\overline{R^2} = 1 - \left(\frac{n-1}{n-2}\right) \times (1 - R^2).$$

$$\overline{R^2} = 1 - \left(\frac{11}{10}\right) \times (1 - 0{,}9).$$

$$\overline{R^2} = 1 - 1{,}1 \times 0{,}1$$

$$\overline{R^2} = 1 - 0{,}11$$

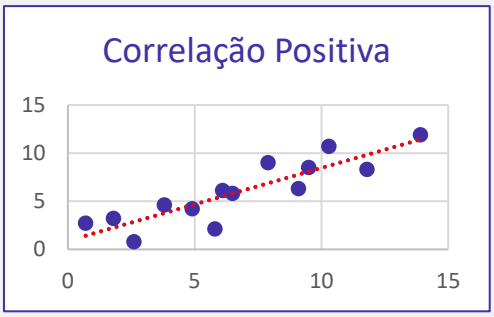
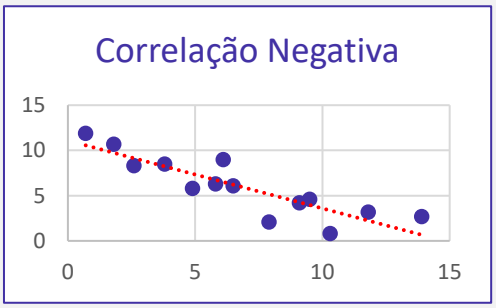
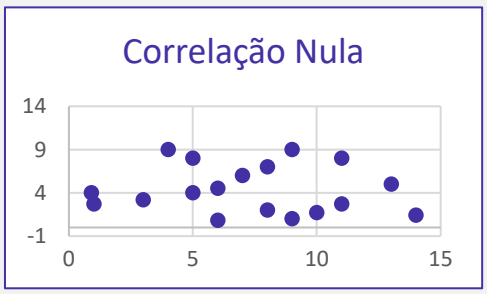
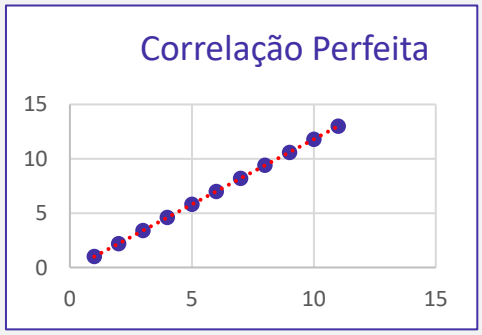
$$\overline{R^2} = 0{,}89.$$

Gabarito: Certo.

RESUMO DA AULA

CORRELAÇÃO LINEAR

A **correlação** é usada para indicar a força que mantém unidos dois conjuntos de valores. A **correlação linear** pode ser:

Gráfico	Definição
	**Direta ou positiva** – quando temos dois fenômenos que variam no mesmo sentido. Se aumentarmos ou diminuirmos um deles, o outro também aumentará ou diminuirá;
	**Inversa ou negativa** – quando temos dois fenômenos que variam em sentido contrário. Se aumentarmos ou diminuirmos um deles, acontecerá o contrário com o outro, no caso, diminuirá ou aumentará;
	**Inexistente ou nula** – quando não existe correlação ou dependência entre os dois fenômenos. Nessa situação, o valor do coeficiente de correlação linear será zero ($r = 0$) ou um valor aproximadamente igual a zero ($r \cong 0$);
	**Perfeita** – quando os fenômenos se ajustam perfeitamente a uma reta.

Coeficiente de Correlação de Pearson

COEFICIENTE DE CORRELAÇÃO LINEAR DE PEARSON

É adotado para medir o quão forte é a RELAÇÃO linear entre duas VARIÁVEIS.

FÓRMULA

$$r = \frac{\sum_{i=1}^{n}[(X_i-\bar{X})\times(Y_i-\bar{Y})]}{\sqrt{\sum_{i=1}^{n}(X_i-\bar{X})^2\times\sum_{i=1}^{n}(Y_i-\bar{Y})^2}}$$

FÓRMULAS ALTERNATIVAS

$$\sum_{i=1}^{n}[(X_i-\bar{X})\times(Y_i-\bar{Y})] = \sum_{i=1}^{n}(X_i\times Y_i) - n\times\bar{X}\times\bar{Y}$$

$$\sum_{i=1}^{n}[(X_i-\bar{X})^2] = \sum_{i=1}^{n}(X_i^2) - n\times\bar{X}^2$$

$$\sum_{i=1}^{n}[(Y_i-\bar{Y})^2] = \sum_{i=1}^{n}(Y_i^2) - n\times\bar{Y}^2$$

Sobre a Coeficiente de Correlação de Pearson, podemos afirmar que:

- I – Pode assumir quaisquer valores entre 1 e -1, ou seja: $-1 \leq r \leq 1$.
- II – Quanto mais próximo r estiver de 0, menor será a relação linear entre as duas variáveis
- III – Quanto mais próximo r estiver de (1 ou -1), maior será a relação linear entre as duas variáveis.

Propriedades do Coeficiente de Correlação

1º Propriedade

- **O coeficiente de correlação não sofre alteração quando uma constante é adicionada a (ou subtraída de) uma variável.**

2º Propriedade

- **O coeficiente de correlação pode não sofrer alteração ou pode ter seu sinal alterado quando uma variável é multiplicada (ou dividida) por uma constante. Caso as constantes tenham o mesmo sinal, o valor do coeficiente de correlação não será alterado. Por outro lado, se as constantes tiverem sinais contrários, o coeficiente mudará de sinal, mas o valor permanecerá inalterado.**

REGRESSÃO LINEAR SIMPLES

REGRESSÃO LINEAR SIMPLES

Calcula a expressão matemática que relaciona Y (variável dependente) em função de X (variável independente).

Trata-se da equação que representa uma reta:

$$y = m \cdot x + b$$

- Propriedades

Sobre a Regressão Linear Simples, podemos afirmar que:

- I – O coeficiente m é conhecido como taxa de variação ou coeficiente angular da reta.
- II – O coeficiente angular é expresso por: $m = \frac{\Delta y}{\Delta x} = \frac{y - y_0}{x - x_0}$
- III – O coeficiente b é conhecido como coeficiente linear da reta e determina o ponto em que a reta intercepta o eixo y.
- IV – Quando a correlação linear não é perfeita, utilizamos a expressão $Y_i = \alpha + \beta X_i + \varepsilon_i$, para determinar a reta de regressão.

Método dos Mínimos Quadrados

MÉTODO DOS MÍNIMOS QUADRADOS

A reta a ser adotada deverá ser aquela que torna mínima a soma dos quadrados das distâncias da reta aos pontos experimentais, medidas no sentido da variação aleatória.

Esse método é empregado na obtenção dos estimadores α e β de um modelo de regressão linear:

$$Y_i = \alpha + \beta X_i + \varepsilon_i.$$

Expressão usada para determinar a reta de regressão é:

$$\hat{Y}_i = a + bX_i$$

Reta Passando pela Origem

MODELO DE REGRESSÃO QUE PASSA OBRIGATORIAMENTE PELA ORIGEM É:

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

$$Y_i = 0 + \beta X_i + \varepsilon_i$$

$$\boxed{Y_i = \beta X_i + \varepsilon_i.}$$

ANÁLISE DE VARIÂNCIA DA REGRESSÃO

ANÁLISE DE VARIÂNCIA DA REGRESSÃO

Estratégia para verificar se compensa ou não utilizar um modelo de regressão linear, $Y_i = \alpha + \beta X_i + \varepsilon_i$

Observar a redução no resíduo (desvio) quando comparado com um modelo aproximadamente uniforme $Y_i = \mu + \varepsilon_i$.

Testar a hipótese:

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

O resultado da **análise de variância** da regressão é uma tabela que resume várias medidas usadas no teste de hipóteses anterior.

Graus de Liberdade

O número de graus de liberdade do modelo de regressão é:

$$GL_{Modelo} = 2 - 1 = 1$$

O número de graus de liberdade dos resíduos é:

$$GL_{Resíduos} = n - 2$$

O número de graus de liberdade total é:

$$GL_{Total} = n - 1$$

Somas de Quadrados

A soma dos quadrados totais é calculada por meio das seguintes fórmulas:

$$SQT = SQM + SQR$$

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

A soma dos quadrados do modelo de regressão é calculada mediante as seguintes fórmulas:

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

$$SQM = b \times \sum_{i=1}^{n} [(X_i - \bar{X}) \times (Y_i - \bar{Y})]$$

$$SQM = b^2 \times \sum_{i=1}^{n} (X_i - \bar{X})^2$$

A soma dos quadrados dos resíduos é calculada pela fórmula:

$$SQR = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2$$

Coeficiente de Determinação

O coeficiente de determinação é calculado pela fórmula:

$$R^2 = \frac{SQM}{SQT}$$

Em que R é o coeficiente de correlação linear, calculado pela expressão:

$$R = \sqrt{\frac{SQM}{SQT}}$$

O coeficiente de determinação também pode ser escrito da seguinte forma:

$$R^2 = \frac{SQM}{SQT} = \frac{SQT - SQR}{SQT} = 1 - \frac{SQR}{SQT}$$

Coeficiente de Determinação Ajustado

É obtida pela **divisão** de SQR e SQT pelos respectivos **graus de liberdade**:

$$\overline{R^2} = 1 - \frac{SQR/(n-2)}{SQT/(n-1)}$$

A **relação** entre o coeficiente de determinação **ajustado** ($\overline{R^2}$) e o coeficiente de determinação **tradicional** (R^2) é dada por:

$$\overline{R^2} = 1 - (1 - R^2) \times \frac{(n-1)}{(n-2)}$$

Quadrados Médios

Quadrado médio do **modelo (QMM)**: $QMM = \frac{SQM}{1}$

Quadrado médio dos **resíduos (QMR)**: $QMR = \frac{SQR}{n-2}$

Quadrado médio **total (QMT)**: $QMT = \frac{SQT}{n-1}$

Estatística F (Razão F)

Estatística F **(ou razão F)**: $F^* = \frac{QMM}{QMR}$

Se $F^* > F_{crítico}$, podemos **rejeitar a hipótese nula**;

Se $F^* < F_{crítico}$, **não** podemos **rejeitar a hipótese nula**.

Tabela de Análise de Variância da Regressão

Fonte de Variação	Graus de Liberdade	Soma dos Quadrados	Quadrados Médios	Estatística F (Razão F)
Modelo	1	SQM	$QMM = \frac{SQM}{1}$	$F^* = \frac{QMM}{QMR}$
Resíduos	$n-2$	SQR	$QMR = \frac{SQR}{n-2}$	
Total	$\boldsymbol{n-1}$	$\boldsymbol{SQT}$	$\boldsymbol{QMT = \frac{SQT}{n-1}}$	

AVISO IMPORTANTE!

Olá, alunos (as)!

Informamos que não temos mais questões da banca, referente ao assunto tratado na aula de hoje, em virtude de baixa cobrança deste tópico ao longo dos anos. No entanto, para complementar o estudo e deixar sua preparação em alto nível, preparamos um caderno de questões inéditas que servirá como treino e aprimoramento do conteúdo.

Em caso de dúvidas, não deixe de nos chamar no Fórum de dúvidas!

Bons estudos!

Estratégia Concursos

QUESTÕES COMENTADAS – INÉDITAS

Correlação Linear

1. (INÉDITA/2022) Em um gráfico de dispersão, observa-se que a maioria dos pontos está situada no primeiro e no terceiro quadrantes. Aqueles que não estão nessa posição, situam-se próximos da origem. Então, a correlação linear entre as variáveis

a) poderá ser fracamente positiva.

b) será necessariamente fortemente positiva.

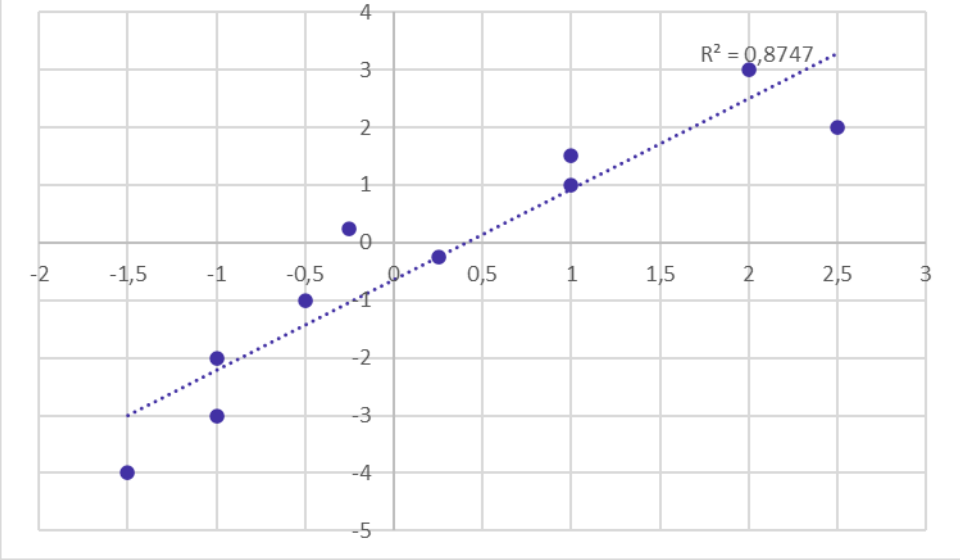
c) poderá ser fracamente negativa.

d) será necessariamente fortemente negativa.

e) será necessariamente nula.

Comentários:

O enunciado afirma que a maioria dos pontos estão no primeiro e no terceiro quadrantes, logo a reta que representa os dados é crescente (correlação linear positiva). Se os pontos não estiverem necessariamente sobre uma reta, a correlação poderá ser fracamente positiva.



Assim, concluímos que o gabarito é a letra A.

Gabarito: A.

2. (INÉDITA/2022) O coeficiente linear de Pearson indica a intensidade da correlação entre duas variáveis e, ainda, o sentido dessa correlação. Assinale a alternativa que indica adequadamente a interpretação de um determinado resultado obtido $r = -0,91$ para o coeficiente de correlação de Pearson:

a) Correlação linear negativa altamente significativa entre as variáveis.

b) Correlação linear positiva altamente significativa entre as variáveis.

c) Correlação linear positiva muito fraca entre as variáveis.

d) Correlação linear positiva relativamente fraca entre as variáveis.

e) Não há correlação linear positiva entre as variáveis.

Comentários:

O coeficiente de correlação varia entre -1 e 1, sendo que quanto mais próximo de zero estiverem os pares ordenados, mais fraca será a correlação; e quanto mais próximo de 1 ou -1 estiverem os pares ordenados, mais forte será a correlação. Na presente questão, temos uma correlação negativa e muito próxima de -1, portanto, altamente significativa.

Gabarito: A.

3. (INÉDITA/2022) A tabela a seguir apresenta as penas de reclusão (P), em anos, cominadas a um grupo de dez réus, e suas respectivas rendas familiares mensais per capitas (R), em número de salários-mínimos.

Réu	P	R	$P \times R$	P^2	R^2
1	15	0,5	7,5	225	0,25
2	12	1,5	18	144	2,25
3	11	2	22	121	4
4	7	1,5	10,5	49	2,25
5	6	1,25	7,5	36	1,5625
6	5	2	10	25	4
7	6	2,5	15	36	6,25
8	2	3	6	4	9
9	2,5	3,5	8,75	6,25	12,25

10	2	4	8	4	16
Totais	**68,5**	**21,75**	**113,25**	**650,25**	**57,8125**

Dados:

$$\mathbf{1810,25}^{\mathbf{1/2}} = \mathbf{42,54}$$

$$\mathbf{105,06}^{\mathbf{1/2}} = \mathbf{10,25}$$

A partir das informações, o coeficiente de correlação linear entre as variáveis R e P é

a) – 0,31.

b) – 0,52.

c) – 0,66.

d) – 0,77.

e) – 0,82.

Comentários:

A fórmula do coeficiente de correlação linear que vamos utilizar é:

$$\rho(x,y) = \frac{n(\sum x_i y_i) - (\sum x_i)(\sum y_i)}{\sqrt{n(\sum x_i^2) - (\sum x_i)^2}\sqrt{n(\sum y_i^2) - (\sum y_i)^2}}$$

Portanto, basta aplicarmos os valores apresentados na tabela. Assim, considerando P como a variável X e R como a variável Y, temos:

$$\rho(x,y) = \frac{10 \times 113,25 - 68,5 \times 21,75}{\sqrt{10 \times 650,25 - 68,5^2}\sqrt{10 \times 57,8125 - 21,75^2}}$$

$$\rho(x,y) = \frac{1132,5 - 1489,88}{\sqrt{6502,5 - 4692,25}\sqrt{578,125 - 473,0625}}$$

$$\rho(x,y) = \frac{-471,5}{\sqrt{1810,25}\sqrt{105,0625}}$$

O enunciado também trouxe alguns dados importantes:

$$1810,25^{1/2} = 42,54$$

$$105,06^{1/2} = 10,25$$

Sabendo disso, podemos utilizar esses valores na fórmula de correlação, ficando assim:

$$\rho(x,y) = \frac{-357,375}{42,547 \times 10,25}$$

$$\rho(x,y) = \frac{-357{,}375}{436{,}1071}$$

$$\rho(x,y) = -0{,}819$$

Gabarito: E.

4. (INÉDITA/2022) A tabela a seguir apresenta as quantidades vendidas de pacotes de leite em pó (P), em unidades, e de pacotes de café (Y), também em unidades.

Mês	X	Y	$X \times Y$	X^2	Y^2
1	10	15	150	100	225
2	8	14	112	64	196
3	12	16	192	144	256
4	5	9	45	25	81
5	3	17	51	9	289
Totais	**38**	**71**	**550**	**342**	**1047**

Dados:

$\mathbf{266}^{1/2} = \mathbf{16,31}$

$\mathbf{194}^{1/2} = \mathbf{13,93}$

A partir das informações, o coeficiente de correlação linear entre as variáveis X e Y é

a) 0,23.

b) 0,31.

c) 0,39.

d) 0,44.

e) 0,49.

Comentários:

A fórmula do coeficiente de correlação linear que vamos utilizar é:

$$\rho(x,y) = \frac{n(\sum x_i y_i) - (\sum x_i)(\sum y_i)}{\sqrt{n(\sum x_i^2) - (\sum x_i)^2}\sqrt{n(\sum y_i^2) - (\sum y_i)^2}}$$

Portanto, basta aplicarmos os valores apresentados na tabela. Assim, temos:

$$\rho(x,y) = \frac{5 \times 550 - 38 \times 71}{\sqrt{5 \times 342 - 38^2}\sqrt{5 \times 1047 - 71^2}}$$

$$\rho(x,y) = \frac{2750 - 2698}{\sqrt{1710 - 1444}\sqrt{5235 - 5041}}$$

$$\rho(x,y) = \frac{52}{\sqrt{266}\sqrt{194}}$$

O enunciado também trouxe alguns dados importantes:

$266^{1/2} = 16{,}31$

$194^{1/2} = 13{,}93$

Sabendo disso, podemos utilizar esses valores na fórmula de correlação, ficando assim:

$$\rho(x,y) = \frac{52}{16{,}31 \times 13{,}93}$$

$$\rho(x,y) = \frac{52}{227{,}1651}$$

$$\rho(x,y) = 0{,}23$$

Gabarito: A.

5. (INÉDITA/2022) A variável x tem desvio padrão igual a 1, enquanto a variável y tem desvio padrão igual a 2. Se a covariância entre x e y é 0,5, então o coeficiente de correlação entre x e y é

a) 0,65.

b) 0,55.

c) 0,45.

d) 0,35.

e) 0,25.

Comentários:

A fórmula do coeficiente de correlação é expressa por:

$$\rho = \frac{Cov(x,y)}{\sigma_x \times \sigma_y}$$

em que $Cov(x,y)$,representa a covariância entre x e y, e σ representa o desvio padrão da variável aleatória.

Substituindo os valores dados no enunciado, temos:

$$\rho(x,y) = \frac{0,5}{1 \times 2}$$

$$\rho(x,y) = 0,25$$

Gabarito: E.

6. (INÉDITA/2022) A variável x tem média 10 e desvio padrão 5, enquanto a variável y tem média 2 e desvio padrão 1. Se a covariância entre x e y é 2, então o coeficiente de correlação entre x e y é

a) 0,4.

b) 0,5.

c) 0,8.

d) –0,4.

e) –0,6.

Comentários:

A fórmula do coeficiente de correlação é expressa por:

$$\rho = \frac{Cov(x,y)}{\sigma_x \times \sigma_y}$$

em que $Cov(x,y)$,representa a covariância entre x e y, e σ representa o desvio padrão da variável aleatória.

Substituindo os valores dados no enunciado, temos:

$$\rho(x,y) = \frac{2}{5 \times 1}$$

$$\rho(x,y) = 0,4$$

Gabarito: A.

7. (INÉDITA/2022) O barista é o profissional especializado em preparar cafés de alta qualidade. Considere que dois baristas foram convidados a avaliar cinco amostras de diferentes cafés.

Amostra	Barista 1	Barista 2
1	7	6
2	6	5
3	9	10

4	10	9
5	3	2

O coeficiente de correlação entre as avaliações atribuídas pelos dois baristas é:

a) 0,57.

b) 0,96.

c) 0,77.

d) 0,89.

e) 0,91.

Comentários:

Primeiro, teremos que calcular as médias de X e Y:

$$\bar{X} = \frac{7+6+9+10+3}{5} = \frac{35}{5} = 7$$

$$\bar{Y} = \frac{6+5+10+9+2}{5} = \frac{32}{5} = 6{,}4$$

Agora, calcularemos os desvios de X e Y em relação às suas médias:

Amostra	X_i	Y_i	$X_i - \bar{X}$	$Y_i - \bar{Y}$
1	7	6	0	-0,4
2	6	5	-1	-1,4
3	9	10	2	3,6
4	10	9	3	2,6
5	3	2	-4	-4,4

Nesse ponto, teremos que calcular o numerador e o denominador do coeficiente de correlação. Para tanto, precisaremos multiplicar os desvios de X pelos desvios de Y, bem como calcular os quadrados dos desvios:

Amostra	X_i	Y_i	$X_i - \bar{X}$	$Y_i - \bar{Y}$	$(X_i - \bar{X}) \times (Y_i - \bar{Y})$	$(X_i - \bar{X})^2$	$(Y_i - \bar{Y})^2$

1	7	6	0	-0,4	0	0	0,16
2	6	5	-1	-1,4	1,4	1	1,96
3	9	10	2	3,6	7,2	4	12,96
4	10	9	3	2,6	7,8	9	6,76
5	3	2	-4	-4,4	17,6	16	19,36
Total					**34**	**30**	**41,2**

Aplicando esses valores na fórmula do coeficiente de correlação linear, temos:

$$r = \frac{\sum_{i=1}^{n}[(X_i - \bar{X}) \times (Y_i - \bar{Y})]}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \times \sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

$$r = \frac{34}{\sqrt{30 \times 41{,}2}}$$

$$r \cong 0{,}967$$

Gabarito: B.

8. (INÉDITA/2022) A quantidade de metros de tecido vendidos por uma loja e o valor das vendas mensais são apresentados na tabela a seguir, em que y representa os valores arrecadados com a venda de tecidos no mês e x a quantidade de metros vendidos.

Mês	x: Venda de Tecidos (a cada 100m)	y : Vendas Mensais (em R$ 1.000,00)
1	1	1
2	2	1
3	3	2
4	4	2
5	5	4

O coeficiente de correlação linear de Pearson entre variáveis x e y é:

a) 0,78

b) 0,97

c) 0,90

d) 0,83

e) 0,86

Comentários:

Primeiro, teremos que calcular as médias de X e Y:

$$\bar{X} = \frac{1+2+3+4+5}{5} = \frac{15}{5} = 3$$

$$\bar{Y} = \frac{1+1+2+2+4}{5} = \frac{10}{5} = 2$$

Agora, calcularemos os desvios de X e Y em relação às suas médias:

Mês	X_i	Y_i	$X_i - \bar{X}$	$Y_i - \bar{Y}$
1	1	1	-2	-1
2	2	1	-1	-1
3	3	2	0	0
4	4	2	1	0
5	5	4	2	2

Nesse ponto, teremos que calcular o numerador e o denominador do coeficiente de correlação. Para tanto, precisaremos multiplicar os desvios de X pelos desvios de Y, bem como calcular os quadrados dos desvios:

Mês	X_i	Y_i	$X_i - \bar{X}$	$Y_i - \bar{Y}$	$(X_i - \bar{X}) \times (Y_i - \bar{Y})$	$(X_i - \bar{X})^2$	$(Y_i - \bar{Y})^2$
1	1	1	-2	-1	2	4	1
2	2	1	-1	-1	1	1	1

3	3	2	0	0	0	0	0
4	4	2	1	0	0	1	0
5	5	4	2	2	4	4	4
Total					7	10	6

Aplicando esses valores na fórmula do coeficiente de correlação linear, temos:

$$r = \frac{\sum_{i=1}^{n}[(X_i - \bar{X}) \times (Y_i - \bar{Y})]}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \times \sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

$$r = \frac{6}{\sqrt{7 \times 10}}$$

$$r \cong 0{,}904$$

Gabarito: C.

9. (INÉDITA/2022) Uma empresa está aplicando métodos motivacionais para melhorar os resultados de seus funcionários com as vendas. Sejam y os valores relativos à quantidade vendida de um produto e x as despesas relativas aos valores investidos em métodos motivacionais apresentados na tabela a seguir.

x	2	6	10	3	7	5
y	100	120	240	120	200	150

O coeficiente de correlação linear entre as quantidades vendidas e os valores empregados em métodos motivacionais:

a) 0,91.

b) 0,87.

c) 0,83.

d) 0,79.

e) 0,75.

Comentários:

Primeiro, teremos que calcular as médias de X e Y:

$$\bar{X} = \frac{2+6+10+8+3+7+5}{6} = 5{,}5$$

$$\bar{Y} = \frac{100+120+240+120+200+150}{6} = 155$$

Agora, calcularemos os desvios de X e Y em relação às suas médias:

X_i	Y_i	$X_i - \bar{X}$	$Y_i - \bar{Y}$
2	100	-3,5	-55
6	120	0,5	-35
10	240	4,5	85
3	120	-2,5	-35
7	200	1,5	45
5	150	-0,5	-5

Nesse ponto, temos que calcular o numerador e o denominador do coeficiente de correlação. Para tanto, precisaremos multiplicar os desvios de X pelos desvios de Y, bem como calcular os quadrados dos desvios:

X_i	Y_i	$X_i - \bar{X}$	$Y_i - \bar{Y}$	$(X_i - \bar{X}) \times (Y_i - \bar{Y})$	$(X_i - \bar{X})^2$	$(Y_i - \bar{Y})^2$
2	100	-3,5	-55	192,5	12,25	3025
6	120	0,5	-35	-17,5	0,25	1225
10	240	4,5	85	382,5	20,25	7225
3	120	-2,5	-35	87,5	6,25	1225
7	200	1,5	45	67,5	2,25	2025
5	150	-0,5	-5	2,5	0,25	25
Total				**715**	**41,5**	**14750**

Aplicando esses valores na fórmula do coeficiente de correlação linear, temos:

$$r = \frac{\sum_{i=1}^{n}[(X_i - \bar{X}) \times (Y_i - \bar{Y})]}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \times \sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

$$r = \frac{715}{\sqrt{41{,}5 \times 14750}}$$

$$r \cong 0{,}914$$

Gabarito: A.

10. (INÉDITA/2022) Os autores de um artigo apresentaram uma análise de correlação para investigar a relação entre o nível máximo de lactato x e a resistência muscular y. Os dados a seguir foram obtidos

x	400	750	700	800	850	1000
y	3,80	4,00	5,00	5,20	4,00	3,50

O coeficiente de correlação linear entre o nível máximo de lactato (x) e a resistência (y) é:

a) 0,11.

b) 0,23.

c) -0,13.

d) -0,06.

e) 0,16.

Comentários:

Primeiro, teremos que calcular as médias de X e Y:

$$\bar{X} = \frac{400 + 750 + 700 + 800 + 850 + 1000}{6} = \frac{4500}{6} = 750$$

$$\bar{Y} = \frac{3{,}8 + 4{,}0 + 5{,}0 + 5{,}2 + 4{,}0 + 3{,}5}{6} = \frac{25{,}5}{6} = 4{,}25$$

Agora, calcularemos os desvios de X e Y em relação às suas médias:

X_i	Y_i	$X_i - \bar{X}$	$Y_i - \bar{Y}$

400	3,8	-350	-0,45
750	4	0	-0,25
700	5	-50	0,75
800	5,2	50	0,95
850	4	100	-0,25
1000	3,5	250	-0,75

Nesse ponto, temos que calcular o numerador e o denominador do coeficiente de correlação. Para tanto, precisaremos multiplicar os desvios de X pelos desvios de Y, bem como calcular os quadrados dos desvios:

X_i	Y_i	$X_i - \bar{X}$	$Y_i - \bar{Y}$	$(X_i - \bar{X}) \times (Y_i - \bar{Y})$	$(X_i - \bar{X})^2$	$(Y_i - \bar{Y})^2$
400	3,8	-350	-0,45	157,5	122500	0,2025
750	4	0	-0,25	0	0	0,0625
700	5	-50	0,75	-37,5	2500	0,5625
800	5,2	50	0,95	47,5	2500	0,9025
850	4	100	-0,25	-25	10000	0,0625
1000	3,5	250	-0,75	-187,5	62500	0,5625
Total				-45	**200000**	**2,355**

Aplicando esses valores na fórmula do coeficiente de correlação linear, temos:

$$r = \frac{\sum_{i=1}^{n}[(X_i - \bar{X}) \times (Y_i - \bar{Y})]}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \times \sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

$$r = \frac{-45}{\sqrt{200000 \times 2{,}355}}$$

$$r \cong -0{,}065$$

Gabarito: D.

QUESTÕES COMENTADAS – INÉDITAS

Regressão Linear Simples

1. (INÉDITA/2022) A nota final de uma disciplina é calculada com base no valor médio obtido pelo método dos mínimos quadrados. Se um aluno obtiver as notas 3,5; 5,5; 7; 8; a sua nota final deverá ser:

a) 5.

b) 5,5.

c) 6.

d) 6,5.

e) 7.

Comentários:

Vamos montar uma tabela com os dados fornecidos:

X	Y	$(X-\bar{X})$	$(Y-\bar{Y})$	$(X-\bar{X})(Y-\bar{Y})$	$(X-\bar{X})^2$
1	3,5	-1,5	-2,5	3,75	2,25
2	5,5	-0,5	-0,5	0,25	0,25
3	7	0,5	1,0	0,50	0,25
4	8	1,5	2,0	3,00	2,25
$\bar{X}=2,5$	$\bar{Y}=6$	**Total**		**7,5**	**5**

Pelo método dos mínimos quadrados, temos:

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

$$b = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2}$$

$$b = \frac{7,5}{5} = 1,5$$

$$a = \bar{Y} - b\bar{X}$$

$$a = 6 - 1,5 \times 2,5$$

$$a = 2,25$$

Nossa reta de regressão será:

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

$$\hat{Y} = 2{,}25 + 1{,}5X$$

Sabendo que a reta de regressão passa pelo ponto $(\bar{X}, \bar{Y})$, então:

$$\hat{Y} = 2{,}25 + 1{,}5 \times 2{,}5$$

$$\hat{Y} = 6$$

Gabarito: C.

2. (INÉDITA/2022) Um pesquisador, ao avaliar o efeito de uma variável X sobre outra Y por meio de uma regressão linear da forma $\hat{Y} = \hat{\alpha} + \hat{\beta}X$, usando o método dos mínimos quadrados, encontrou, a partir de 10 amostras, os seguintes somatórios:

$$\boldsymbol{\sum X = 150; \sum Y = 200; \sum X^2 = 2.750;\ e\ \sum (XY) = 4.500}$$

A partir desses resultados, julgue o item a seguir.

a) para $X = 10$, a estimativa de Y é $\hat{Y} = 2$.

b) para $X = 8$, a estimativa de Y é $\hat{Y} = 5$.

c) para $X = 8$, a estimativa de Y é $\hat{Y} = -1$.

d) para $X = 9$, a estimativa de Y é $\hat{Y} = 5$.

e) para $X = 10$, a estimativa de Y é $\hat{Y} = -1$.

Comentários:

Inicialmente, vamos calcular os valores de $\bar{Y}$ e de $\bar{X}$:

$$\bar{Y} = \frac{\sum y}{n} = \frac{200}{10} = 20$$

$$\bar{X} = \frac{\sum x}{n} = \frac{150}{10} = 15$$

Agora, utilizaremos o método dos mínimos quadrados para determinar $\hat{\beta}$:

$$\hat{\beta} = \frac{\sum X_i Y_i - n\bar{X}\bar{Y}}{\sum X_i^2 - n\bar{X}^2}$$

$$\hat{\beta} = \frac{4500 - 10 \times 15 \times 20}{2750 - 10 \times 15^2}$$

$$\hat{\beta} = \frac{1500}{500} = 3$$

Conhecendo $\hat{\beta}$, podemos determinar o valor de $\hat{\alpha}$:

$$\hat{\alpha} = \frac{\sum Y_i - \hat{\beta} \sum X_i}{n}$$

$$\hat{\alpha} = 20 - 3 \times 15 = -25$$

Assim, o modelo de regressão é dado por:

$$\hat{Y} = -25 + 3 \times X$$

Para $X = 10$, temos o seguinte valor de $\hat{Y}$:

$$\hat{Y} = -25 + 3 \times 10$$

$$\hat{Y} = 5$$

Para $X = 9$, temos o seguinte valor de $\hat{Y}$:

$$\hat{Y} = -25 + 3 \times 9$$

$$\hat{Y} = 2$$

Para $X = 8$, temos o seguinte valor de $\hat{Y}$:

$$\hat{Y} = -25 + 3 \times 8$$

$$\hat{Y} = -1$$

Gabarito: C.

3. (INÉDITA/2022) A respeito de duas variáveis, X e Y, sabe-se que:

$$\boldsymbol{\sum X = 20}$$

$$\boldsymbol{\sum Y = 12}$$

$$\boldsymbol{\sum XY = 48}$$

$$\boldsymbol{\sum X^2 = 52}$$

$$\boldsymbol{(\sum X)^2 = 169}$$

Se o tamanho da amostra é igual a 4, então o valor de "b" na regressão simples $\boldsymbol{Y = a + bX}$ é:

a) 1/4.

b) 3/7.

c) 1/3.

d) 3/4.

e) 2/5.

Comentários:

Primeiro calculamos as médias de X e de Y:

$$\bar{X} = \frac{\sum X}{n} = \frac{20}{4} = 5$$

$$\bar{Y} = \frac{\sum Y}{n} = \frac{12}{4} = 3$$

O valor de b é calculado pela seguinte relação:

$$b = \frac{\sum XY - n \times \bar{X} \times \bar{Y}}{\sum X^2 - n \times (\bar{X})^2}$$

$$b = \frac{48 - 4 \times 5 \times 3}{52 - 4 \times 5^2}$$

$$b = \frac{48 - 60}{52 - 100}$$

$$b = \frac{-12}{-48}$$

$$b = \frac{1}{4}$$

Gabarito: A.

QUESTÕES COMENTADAS – INÉDITAS

Análise de Variância da Regressão

1. (INÉDITA/2022) Em uma regressão linear simples em que foi utilizada uma amostra com 42 observações, a soma dos quadrados totais é de 40 e a soma dos quadrados dos resíduos é de 10. O coeficiente de determinação e a estatística F dessa regressão são, respectivamente:

a) 0,6 e 225.

b) 0,6 e 120.

c) 0,75 e 120.

d) 0,75 e 225.

e) 0,8 e 75.

Comentários:

O coeficiente de determinação da regressão linear é expresso por:

$$R^2 = \frac{SQM}{SQT} = 1 - \frac{SQR}{SQT}$$

Substituindo os valores, temos:

$$R^2 = 1 - \frac{10}{40}$$

$$R^2 = 1 - 0{,}25$$

$$R^2 = 0{,}75$$

O coeficiente de determinação e a estatística F:

$$F = \frac{QMM}{QMR} = \frac{\frac{SQM}{1}}{\frac{SQR}{(n-2)}} = \frac{SQM \times (n-2)}{SQR}$$

Temos 42 observações, logo:

$$F = \frac{30 \times (42-2)}{10}$$

$$F = 3 \times 40$$

$$F = 120$$

Gabarito: C.

2. (INÉDITA/2022) Um pesquisador aplicou um modelo de regressão linear simples na forma $Y = bX + a + \varepsilon$, em que b representa o coeficiente angular, a é o intercepto do modelo e ε denota o erro aleatório com média zero e variância σ^2. A tabela a seguir representa a análise de variância (ANOVA) proporcionada por esse modelo.

Fonte de variação	Graus de Liberdade	Soma de Quadrados
Modelo	1	275
Erro	89	125
Total	90	400

A correlação linear de Pearson entre a variável resposta Y e a variável regressora X é igual:

a) 0,83

b) 0,81

c) 0,79

d) 0,74

e) 0,70

Comentários:

O coeficiente de determinação da regressão linear é dado por:

$$R^2 = \frac{SQM}{SQT}$$

$SQM \rightarrow Soma\ dos\ quadrados\ do\ modelo\ de\ regressão$

$SQT \rightarrow Soma\ dos\ quadrados\ total\ (SQT = SQM + SQR)$

Substituindo pelos valores da tabela, temos:

$$R^2 = \frac{275}{400}$$

$$R^2 = 0{,}6875$$

$$R = \sqrt{0{,}6875}$$

$$R = 0{,}83$$

Gabarito: A.

3. (INÉDITA/2022) Um pesquisador aplicou um modelo de regressão linear simples na forma $Y = bX + a + \varepsilon$, em que b representa o coeficiente angular, a é o intercepto do modelo e ε denota o erro aleatório com média zero e variância σ^2. A tabela a seguir representa a análise de variância (ANOVA) proporcionada por esse modelo.

Fonte	S. Quadrados	G.L.	Q. Médio
Modelo	500	1	500
Erro	250	10	25
Total	750	11	

A partir das informações e da tabela apresentadas, a correlação linear de Pearson entre a variável resposta Y e a variável regressora X é igual:

a) 0,416

b) 0,516

c) 0,616

d) 0,716

e) 0,816

Comentários:

O coeficiente de correlação linear entre as variáveis X e Y é calculado por meio da seguinte expressão:

$$R = \sqrt{\frac{SQM}{SQT}},$$

em que SQM indica a soma dos quadrados da regressão (modelo) e SQT a soma dos quadrados totais.

Pela tabela, verificamos que $SQT = 750$ e $SQM = 500$. Substituindo esses valores na equação anterior, teremos:

$$R = \sqrt{\frac{500}{750}} = \sqrt{0{,}666} = 0{,}816$$

Gabarito: E.

4. (INÉDITA/2022) Um pesquisador aplicou um modelo de regressão linear simples na forma $Y = bX + a + \varepsilon$, em que b representa o coeficiente angular, a é o intercepto do modelo e ε denota o erro aleatório

com média zero e variância σ^2. A tabela a seguir representa a análise de variância (ANOVA) proporcionada por esse modelo.

Fonte	S. Quadrados	G.L.	Q. Médio
Modelo	500	1	500
Erro	250	10	25
Total	750	11	

A partir das informações e da tabela apresentadas, pode-se constatar que a variância amostral é aproximadamente igual a:

a) 64,18

b) 66,18

c) 68,18

d) 70,18

e) 72,18

Comentários:

A soma dos quadrados totais (SQT) é dada por:

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

A variância amostral é calculada por:

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

Pela tabela, o grau de liberdade do total corresponde a 11, então:

$$n - 1 = 11$$

Logo, a variância amostral é:

$$\frac{\sum_{i=1}^{n} (Y_i - \bar{Y})^2}{n-1} = \frac{750}{11} = 68{,}18$$

Gabarito: C.

5. (INÉDITA/2022) Um pesquisador aplicou um modelo de regressão linear simples na forma $Y = bX + a + \varepsilon$, em que b representa o coeficiente angular, a é o intercepto do modelo e ε denota o erro aleatório

com média zero e variância σ^2. A tabela a seguir representa a análise de variância (ANOVA) proporcionada por esse modelo.

Fonte	S. Quadrados	G.L.	Q. Médio
Modelo	500	1	500
Erro	250	10	25
Total	750	11	

A partir das informações e da tabela apresentadas, pode-se constatar que o R^2 ajustado é aproximadamente igual a:

a) 0, 733

b) 0,633

c) 0,583

d) 0,553

e) 0,513

Comentários:

O coeficiente de determinação é calculado por meio da expressão:

$$R^2 = \frac{SQM}{SQT} = 1 - \frac{SQR}{SQT},$$

em que SQM = soma dos quadrados da regressão (modelo), SQR = soma dos quadrados dos resíduos e SQT = soma dos quadrados totais.

O R^2 ajustado é definido para uma equação com 2 coeficientes como

$$\overline{R^2} = 1 - \left(\frac{n-1}{n-2}\right) \times (1 - R^2).$$

Pela tabela temos que $SQT = 750$ e $SQM = 500$. Substituindo os valores apresentados na tabela nas equações acima, teremos:

$$R^2 = \frac{500}{750} = 0{,}666.$$

Além disso, como temos $n - 1 = 11$ graus de liberdade totais, então

$$\overline{R^2} = 1 - \left(\frac{n-1}{n-2}\right) \times (1 - R^2).$$

$$\overline{R^2} = 1 - \left(\frac{11}{10}\right) \times (1 - 0{,}666).$$

$$\overline{R^2} = 1 - 1{,}1 \times 0{,}333$$

$$\overline{R^2} = 1 - 0{,}3666$$

$$\overline{R^2} = 0{,}6333.$$

Gabarito: B.

6. (INÉDITA/2022) Após estimar um modelo de regressão simples, um estatístico trabalha nos resultados, obtendo a tabela a seguir:

Fonte	S. Quadrados	G.L.	Q. Médio	p-value
Equação	600	1	600	0,15%
Resíduos	200	8	25	
Total	800	9	88,88	

A partir desses números, é correto concluir que:

a) a estimativa do coeficiente angular da equação é igual a 5.

b) a variância estimada do erro aleatório é inferior a 20;

c) apesar de um poder de explicação de 75%, o modelo não passa no teste de significância da estatística F;

d) o modelo é capaz de explicar 75% da variação total;

e) a variância da variável explicativa do modelo é igual a 200;

Comentários:

Analisando-se cada alternativa, temos:

Letra A: **Errado**. Não é possível estimar o coeficiente angular, pois não há informações acerca de $\sum(X_i - \bar{X})^2$. Logo, não é possível afirmar que tal estimativa valha 5.

Letra B: **Errado**. Um estimador não viciado para a variância do erro σ^2 é dado por

$$\sigma^2 = QM_{Res} = 25$$

Letra C: **Errado**. O p-valor próximo de zero significa que o modelo proposto é adequado, pois fornece evidências contra a hipótese nula. Contudo, com base nas informações dadas na questão, não podemos afirmar que o modelo passa no teste de significância da estatística F.

Letra D: **Correto**. O coeficiente de determinação R^2 estima a proporção da variabilidade da variável dependente que é explicada pelo conjunto das k variáveis independentes do modelo de regressão. Devemos lembrar que a definição do coeficiente é

$$R^2 = \frac{SQM}{SQT}.$$

Com isso, teremos:

$$R^2 = \frac{600}{800} = 0{,}75 = 75\%.$$

Letra E: **Errado**. A variância da variável explicativa do modelo é dada pela soma dos quadrados do modelo. Representa as distâncias quadráticas dos valores ajustados pelo modelo em relação à média aritmética. Nessa questão, esse valor é igual a 600.

Gabarito: D.

7. (INÉDITA/2022) Um perito foi convocado para verificar se o valor de Y, número de pastilhas descoladas de uma fachada, estava relacionado ao valor de X, tempo de exposição ao sol.

Fonte de variação	Graus de Liberdade	Soma de Quadrados
Modelo	1	4,80
Erro	5	1,20
Total	6	6,000

Considerando que $\sum(x_i - \bar{x})^2 = 20; e\ 0,06^{1/2} = 0,244$, o valor do coeficiente angular é aproximadamente igual a:

a) 0,244

b) 0,488

c) 0,366

d) 0,566

e) 0,433

Comentários:

Sabemos que:

$$SQM = b^2 \times \sum (X_i - \bar{X})^2$$

em que b^2 é a estimativa do coeficiente angular da reta de regressão. Substituindo, temos:

$$4{,}80 = b^2 \times 20$$

$$b^2 = 0{,}24$$

$$b^2 = \sqrt{0{,}24} = 2 \times \sqrt{0{,}06} = 2 \times 0{,}244 = 0{,}488$$

Gabarito: B.

8. (INÉDITA/2022) Em um modelo de regressão linear simples, obteve-se um coeficiente de correlação igual a 0,9. O coeficiente de determinação é aproximadamente igual a

a) 0,36.

b) 0,68.

c) 0,70.

d) 0,81.

e) 0,89.

Comentários:

O coeficiente de determinação é o quadrado do coeficiente de correlação:

$$R^2 = 0{,}9^2 = 0{,}81$$

Gabarito: D.

9. (INÉDITA/2022) Considere que o coeficiente de determinação tenha sido de 0,49, no ajuste de uma reta de regressão linear simples de uma variável Y em uma variável X. Então, o módulo do coeficiente de correlação amostral entre X e Y é igual a:

a) 0,25

b) 0,35

c) 0,50

d) 0,65

e) 0,70

Comentários:

Seja R^2 o coeficiente de determinação e r o coeficiente de correlação.

$$R^2 = 0{,}49$$

Logo,

$$R = \pm\sqrt{0{,}49}$$

$$R = \pm 0{,}7$$

O módulo de R vale 0,7.

Gabarito: E.

LISTA DE QUESTÕES – INÉDITAS

Correlação Linear

1. (INÉDITA/2022) Em um gráfico de dispersão, observa-se que a maioria dos pontos está situada no primeiro e no terceiro quadrantes. Aqueles que não estão nessa posição, situam-se próximos da origem. Então, a correlação linear entre as variáveis

a) poderá ser fracamente positiva.

b) será necessariamente fortemente positiva.

c) poderá ser fracamente negativa.

d) será necessariamente fortemente negativa.

e) será necessariamente nula.

2. (INÉDITA/2022) O coeficiente linear de Pearson indica a intensidade da correlação entre duas variáveis e, ainda, o sentido dessa correlação. Assinale a alternativa que indica adequadamente a interpretação de um determinado resultado obtido $r = -0,91$ para o coeficiente de correlação de Pearson:

a) Correlação linear negativa altamente significativa entre as variáveis.

b) Correlação linear positiva altamente significativa entre as variáveis.

c) Correlação linear positiva muito fraca entre as variáveis.

d) Correlação linear positiva relativamente fraca entre as variáveis.

e) Não há correlação linear positiva entre as variáveis.

3. (INÉDITA/2022) A tabela a seguir apresenta as penas de reclusão (P), em anos, cominadas a um grupo de dez réus, e suas respectivas rendas familiares mensais per capitas (R), em número de salários-mínimos.

Réu	P	R	$P \times R$	P^2	R^2
1	15	0,5	7,5	225	0,25
2	12	1,5	18	144	2,25
3	11	2	22	121	4
4	7	1,5	10,5	49	2,25

5	6	1,25	7,5	36	1,5625
6	5	2	10	25	4
7	6	2,5	15	36	6,25
8	2	3	6	4	9
9	2,5	3,5	8,75	6,25	12,25
10	2	4	8	4	16
Totais	**68,5**	**21,75**	**113,25**	**650,25**	**57,8125**

Dados:

$$1810,25^{1/2} = 42,54$$

$$105,06^{1/2} = 10,25$$

A partir das informações, o coeficiente de correlação linear entre as variáveis R e P é

a) – 0,31.

b) – 0,52.

c) – 0,66.

d) – 0,77.

e) – 0,82.

4. (INÉDITA/2022) A tabela a seguir apresenta as quantidades vendidas de pacotes de leite em pó (P), em unidades, e de pacotes de café (Y), também em unidades.

Mês	X	Y	$X \times Y$	X^2	Y^2
1	10	15	150	100	225
2	8	14	112	64	196
3	12	16	192	144	256
4	5	9	45	25	81
5	3	17	51	9	289

Totais	38	71	550	342	1047

Dados:

$266^{1/2} = 16,31$

$194^{1/2} = 13,93$

A partir das informações, o coeficiente de correlação linear entre as variáveis X e Y é

a) 0,23.

b) 0,31.

c) 0,39.

d) 0,44.

e) 0,49.

5. (INÉDITA/2022) A variável x tem desvio padrão igual a 1, enquanto a variável y tem desvio padrão igual a 2. Se a covariância entre x e y é 0,5, então o coeficiente de correlação entre x e y é

a) 0,65.

b) 0,55.

c) 0,45.

d) 0,35.

e) 0,25.

6. (INÉDITA/2022) A variável x tem média 10 e desvio padrão 5, enquanto a variável y tem média 2 e desvio padrão 1. Se a covariância entre x e y é 2, então o coeficiente de correlação entre x e y é

a) 0,4.

b) 0,5.

c) 0,8.

d) –0,4.

e) –0,6.

7. (INÉDITA/2022) O barista é o profissional especializado em preparar cafés de alta qualidade. Considere que dois baristas foram convidados a avaliar cinco amostras de diferentes cafés.

Amostra	Barista 1	Barista 2

1	7	6
2	6	5
3	9	10
4	10	9
5	3	2

O coeficiente de correlação entre as avaliações atribuídas pelos dois baristas é:

a) 0,57.

b) 0,96.

c) 0,77.

d) 0,89.

e) 0,91.

8. (INÉDITA/2022) A quantidade de metros de tecido vendidos por uma loja e o valor das vendas mensais são apresentados na tabela a seguir, em que y representa os valores arrecadados com a venda de tecidos no mês e x a quantidade de metros vendidos.

Mês	x: Venda de Tecidos (a cada 100m)	y : Vendas Mensais (em R$ 1.000,00)
1	1	1
2	2	1
3	3	2
4	4	2
5	5	4

O coeficiente de correlação linear de Pearson entre variáveis x e y é:

a) 0,78

b) 0,97

c) 0,90

d) 0,83

e) 0,86

9. (INÉDITA/2022) Uma empresa está aplicando métodos motivacionais para melhorar os resultados de seus funcionários com as vendas. Sejam y os valores relativos à quantidade vendida de um produto e x as despesas relativas aos valores investidos em métodos motivacionais apresentados na tabela a seguir.

x	2	6	10	3	7	5
y	100	120	240	120	200	150

O coeficiente de correlação linear entre as quantidades vendidas e os valores empregados em métodos motivacionais:

a) 0,91.

b) 0,87.

c) 0,83.

d) 0,79.

e) 0,75.

10. (INÉDITA/2022) Os autores de um artigo apresentaram uma análise de correlação para investigar a relação entre o nível máximo de lactato x e a resistência muscular y. Os dados a seguir foram obtidos

x	400	750	700	800	850	1000
y	3,80	4,00	5,00	5,20	4,00	3,50

O coeficiente de correlação linear entre o nível máximo de lactato (x) e a resistência (y) é:

a) 0,11.

b) 0,23.

c) -0,13.

d) -0,06.

e) 0,16.

GABARITO – INÉDITAS

Correlação Linear

1. LETRA A
2. LETRA A
3. LETRA E
4. LETRA A
5. LETRA E
6. LETRA A
7. LETRA B
8. LETRA C
9. LETRA A
10. LETRA D

LISTA DE QUESTÕES – INÉDITAS

Regressão Linear Simples

1. (INÉDITA/2022) A nota final de uma disciplina é calculada com base no valor médio obtido pelo método dos mínimos quadrados. Se um aluno obtiver as notas 3,5; 5,5; 7; 8; a sua nota final deverá ser:

a) 5.

b) 5,5.

c) 6.

d) 6,5.

e) 7.

2. (INÉDITA/2022) Um pesquisador, ao avaliar o efeito de uma variável X sobre outra Y por meio de uma regressão linear da forma $\widehat{Y} = \widehat{\alpha} + \widehat{\beta}X$, usando o método dos mínimos quadrados, encontrou, a partir de 10 amostras, os seguintes somatórios:

$$\sum X = 150; \sum Y = 200; \sum X^2 = 2.750;\ e\ \sum(XY) = 4.500$$

A partir desses resultados, julgue o item a seguir.

a) para $X = 10$, a estimativa de Y é $\hat{Y} = 2$.

b) para $X = 8$, a estimativa de Y é $\hat{Y} = 5$.

c) para $X = 8$, a estimativa de Y é $\hat{Y} = -1$.

d) para $X = 9$, a estimativa de Y é $\hat{Y} = 5$.

e) para $X = 10$, a estimativa de Y é $\hat{Y} = -1$.

3. (INÉDITA/2022) A respeito de duas variáveis, X e Y, sabe-se que:

$$\sum X = 20$$

$$\sum Y = 12$$

$$\sum XY = 48$$

$$\sum X^2 = 52$$

$$(\sum X)^2 = 169$$

Se o tamanho da amostra é igual a 4, então o valor de "b" na regressão simples $Y = a + bX$ é:

a) 1/4.

b) 3/7.

c) 1/3.

d) 3/4.

e) 2/5.

GABARITO – INÉDITAS

Regressão Linear Simples

1. LETRA C

2. LETRA C

3. LETRA A

LISTA DE QUESTÕES – INÉDITAS

Análise de Variância da Regressão

1. (INÉDITA/2022) Em uma regressão linear simples em que foi utilizada uma amostra com 42 observações, a soma dos quadrados totais é de 40 e a soma dos quadrados dos resíduos é de 10. O coeficiente de determinação e a estatística F dessa regressão são, respectivamente:

a) 0,6 e 225.

b) 0,6 e 120.

c) 0,75 e 120.

d) 0,75 e 225.

e) 0,8 e 75.

2. (INÉDITA/2022) Um pesquisador aplicou um modelo de regressão linear simples na forma $Y = bX + a + \varepsilon$, em que b representa o coeficiente angular, a é o intercepto do modelo e ε denota o erro aleatório com média zero e variância σ^2. A tabela a seguir representa a análise de variância (ANOVA) proporcionada por esse modelo.

Fonte de variação	Graus de Liberdade	Soma de Quadrados
Modelo	1	275
Erro	89	125
Total	90	400

A correlação linear de Pearson entre a variável resposta Y e a variável regressora X é igual:

a) 0,83

b) 0,81

c) 0,79

d) 0,74

e) 0,70

3. (INÉDITA/2022) Um pesquisador aplicou um modelo de regressão linear simples na forma $Y = bX + a + \varepsilon$, em que b representa o coeficiente angular, a é o intercepto do modelo e ε denota o erro aleatório com média zero e variância σ^2. A tabela a seguir representa a análise de variância (ANOVA) proporcionada por esse modelo.

Fonte	S. Quadrados	G.L.	Q. Médio
Modelo	500	1	500
Erro	250	10	25
Total	750	11	

A partir das informações e da tabela apresentadas, a correlação linear de Pearson entre a variável resposta Y e a variável regressora X é igual:

a) 0,416

b) 0,516

c) 0,616

d) 0,716

e) 0,816

4. (INÉDITA/2022) Um pesquisador aplicou um modelo de regressão linear simples na forma $Y = bX + a + \varepsilon$, em que b representa o coeficiente angular, a é o intercepto do modelo e ε denota o erro aleatório com média zero e variância σ^2. A tabela a seguir representa a análise de variância (ANOVA) proporcionada por esse modelo.

Fonte	S. Quadrados	G.L.	Q. Médio
Modelo	500	1	500
Erro	250	10	25
Total	750	11	

A partir das informações e da tabela apresentadas, pode-se constatar que a variância amostral é aproximadamente igual a:

a) 64,18

b) 66,18

c) 68,18

d) 70,18

e) 72,18

5. (INÉDITA/2022) Um pesquisador aplicou um modelo de regressão linear simples na forma $Y = bX + a + \varepsilon$, em que b representa o coeficiente angular, a é o intercepto do modelo e ε denota o erro aleatório com média zero e variância σ^2. A tabela a seguir representa a análise de variância (ANOVA) proporcionada por esse modelo.

Fonte	S. Quadrados	G.L.	Q. Médio
Modelo	500	1	500
Erro	250	10	25
Total	750	11	

A partir das informações e da tabela apresentadas, pode-se constatar que o R^2 ajustado é aproximadamente igual a:

a) 0, 733

b) 0,633

c) 0,583

d) 0,553

e) 0,513

6. (INÉDITA/2022) Após estimar um modelo de regressão simples, um estatístico trabalha nos resultados, obtendo a tabela a seguir:

Fonte	S. Quadrados	G.L.	Q. Médio	p-value
Equação	600	1	600	0,15%
Resíduos	200	8	25	
Total	800	9	88,88	

A partir desses números, é correto concluir que:

a) a estimativa do coeficiente angular da equação é igual a 5.

b) a variância estimada do erro aleatório é inferior a 20;

c) apesar de um poder de explicação de 75%, o modelo não passa no teste de significância da estatística F;

d) o modelo é capaz de explicar 75% da variação total;

e) a variância da variável explicativa do modelo é igual a 200;

7. (INÉDITA/2022) Um perito foi convocado para verificar se o valor de Y, número de pastilhas descoladas de uma fachada, estava relacionado ao valor de X, tempo de exposição ao sol.

Fonte de variação	Graus de Liberdade	Soma de Quadrados
Modelo	1	4,80
Erro	5	1,20
Total	6	6,000

Considerando que $\sum(x_i - \bar{x})^2 = 20; e\ 0,06^{1/2} = 0,244$, o valor do coeficiente angular é aproximadamente igual a:

a) 0,244

b) 0,488

c) 0,366

d) 0,566

e) 0,433

8. (INÉDITA/2022) Em um modelo de regressão linear simples, obteve-se um coeficiente de correlação igual a 0,9. O coeficiente de determinação é aproximadamente igual a

a) 0,36.

b) 0,68.

c) 0,70.

d) 0,81.

e) 0,89.

9. (INÉDITA/2022) Considere que o coeficiente de determinação tenha sido de 0,49, no ajuste de uma reta de regressão linear simples de uma variável Y em uma variável X. Então, o módulo do coeficiente de correlação amostral entre X e Y é igual a:

a) 0,25

b) 0,35

c) 0,50

d) 0,65

e) 0,70

GABARITO – INÉDITAS

Análise de Variância da Regressão

1. LETRA C
2. LETRA A
3. LETRA E
4. LETRA C
5. LETRA B
6. LETRA D
7. LETRA B
8. LETRA D
9. LETRA E

ESSA LEI TODO MUNDO CONHECE:
PIRATARIA É CRIME.

Mas é sempre bom revisar o porquê e como você pode ser prejudicado com essa prática.

1. 

Professor investe seu tempo para elaborar os cursos e o site os coloca à venda.

2. 

Pirata divulga ilicitamente (grupos de rateio), utilizando-se do anonimato, nomes falsos ou laranjas (geralmente o pirata se anuncia como formador de "grupos solidários" de rateio que não visam lucro).

3. 

Pirata cria alunos fake praticando falsidade ideológica, comprando cursos do site em nome de pessoas aleatórias (usando nome, CPF, endereço e telefone de terceiros sem autorização).

4. 

Pirata compra, muitas vezes, clonando cartões de crédito (por vezes o sistema anti-fraude não consegue identificar o golpe a tempo).

5. 

Pirata fere os Termos de Uso, adultera as aulas e retira a identificação dos arquivos PDF (justamente porque a atividade é ilegal e ele não quer que seus fakes sejam identificados).

6. 

Pirata revende as aulas protegidas por direitos autorais, praticando concorrência desleal e em flagrante desrespeito à Lei de Direitos Autorais (Lei 9.610/98).

7. 

Concurseiro(a) desinformado participa de rateio, achando que nada disso está acontecendo e esperando se tornar servidor público para exigir o cumprimento das leis.

8. 

O professor que elaborou o curso não ganha nada, o site não recebe nada, e a pessoa que praticou todos os ilícitos anteriores (pirata) fica com o lucro.

Deixando de lado esse mar de sujeira, aproveitamos para agradecer a todos que adquirem os cursos honestamente e permitem que o site continue existindo.