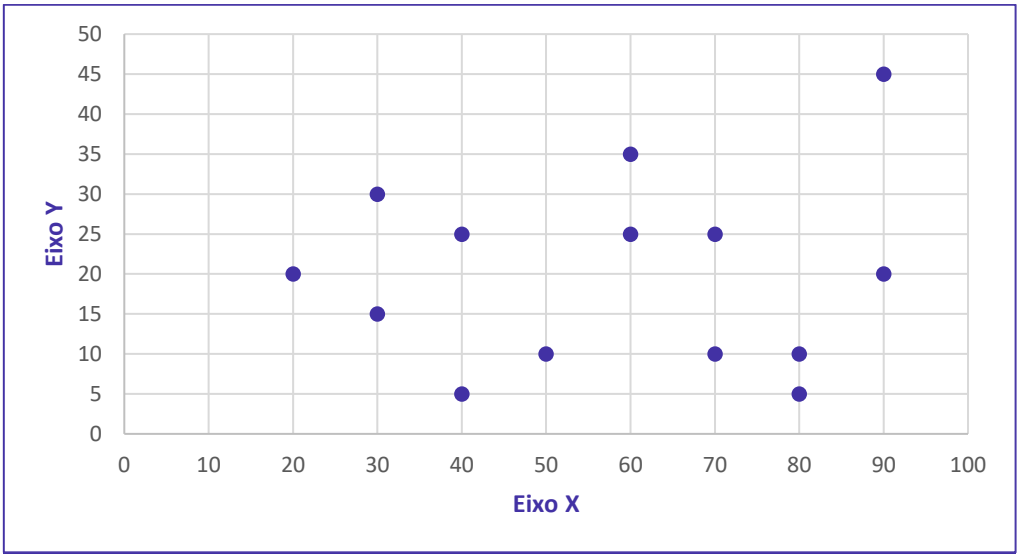
By @kakashi_copiador

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).

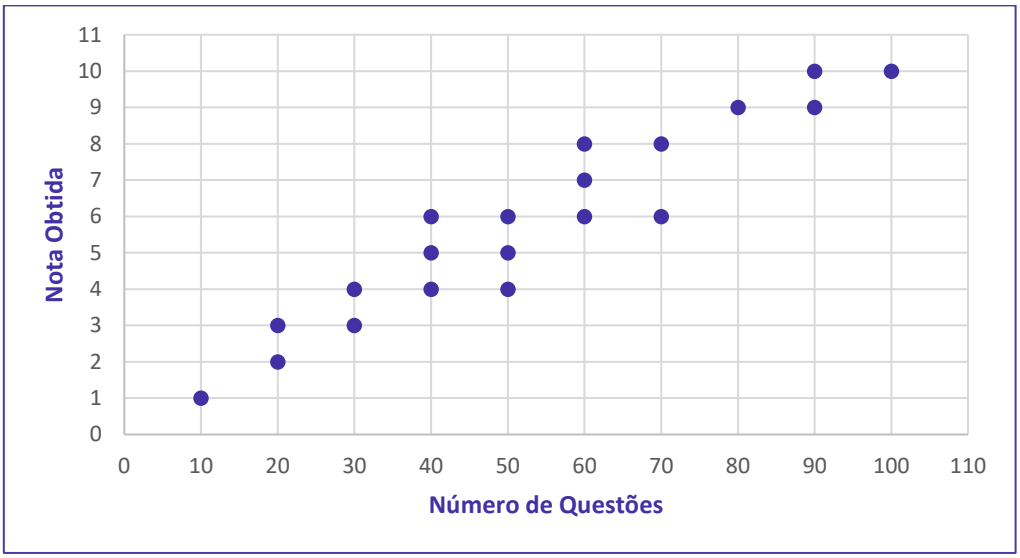


A título exemplificativo, as informações da tabela a seguir referem-se ao número de questões resolvidas por um determinado aluno e a nota obtida por ele em uma avaliação. Observe que quanto maior o número de questões resolvidas, maior é a nota obtida na avaliação.

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

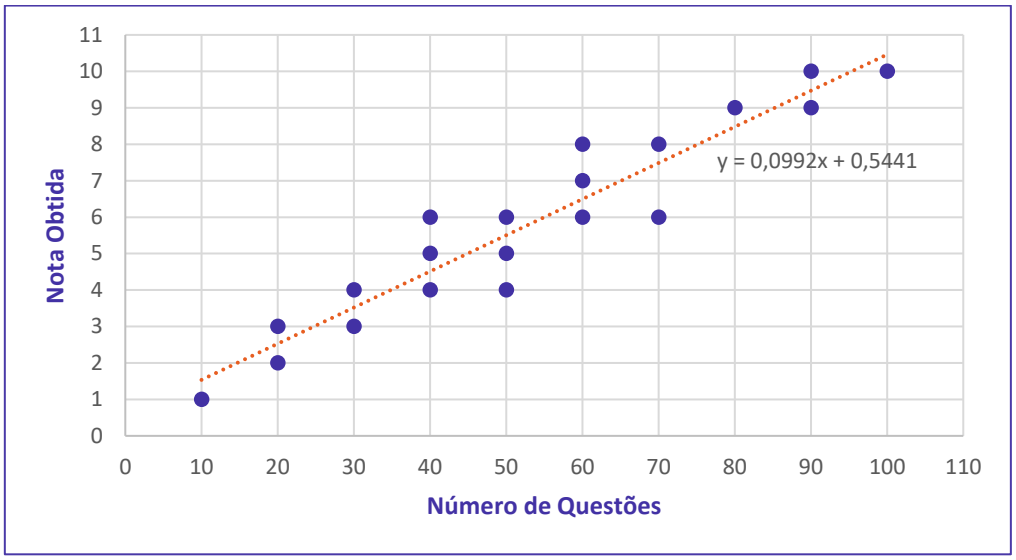
10	30	4
11	10	1
12	60	6
13	50	5
14	70	6
15	90	9
16	100	10
17	20	3
18	40	5
19	60	7
20	50	4

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:

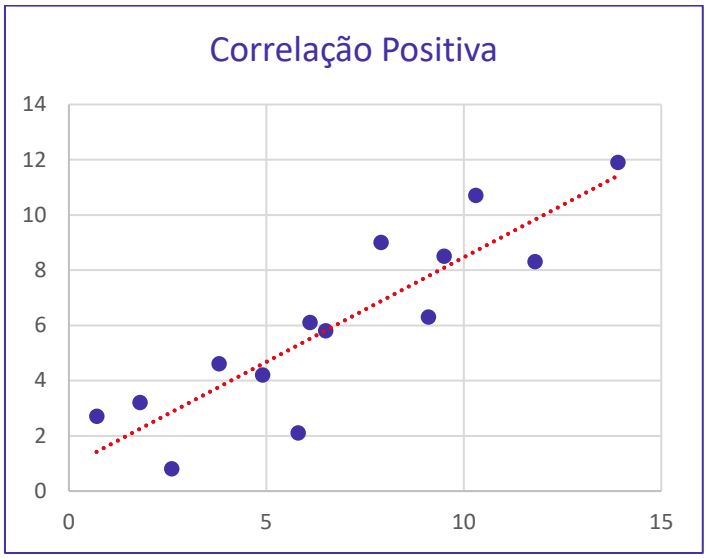
a) direta ou positiva – quando temos dois fenômenos que variam no mesmo sentido. Se aumentarmos ou diminuirmos um deles, o outro também aumentará ou diminuirá;

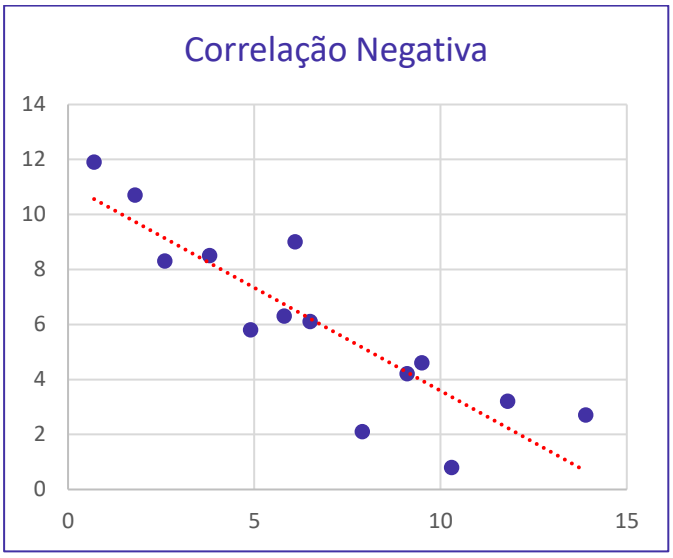
b) inversa ou negativa – quando temos dois fenômenos que variam em sentido contrário. Se aumentarmos ou diminuirmos um deles, acontecerá o contrário com o outro, no caso, diminuirá ou aumentará;

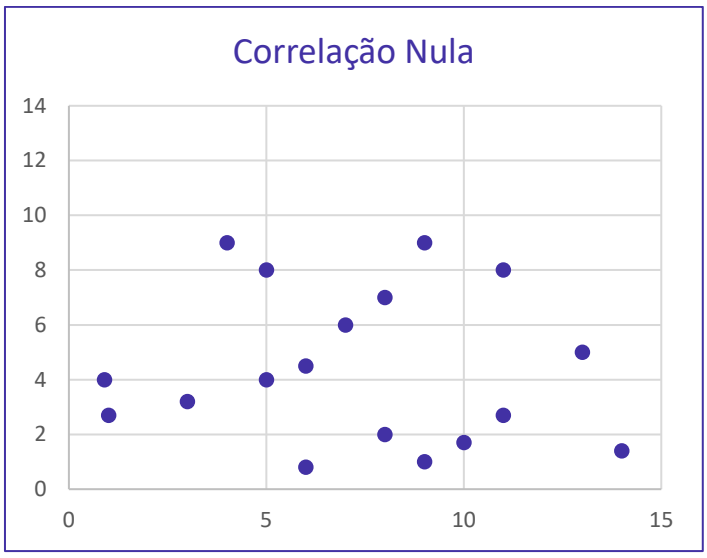
c) inexistente ou nula – quando não existe correlação ou dependência entre os dois fenômenos. Nessa situação, o valor do coeficiente de correlação linear será zero ($r = 0$) ou um valor aproximadamente igual a zero ($r \cong 0$); e

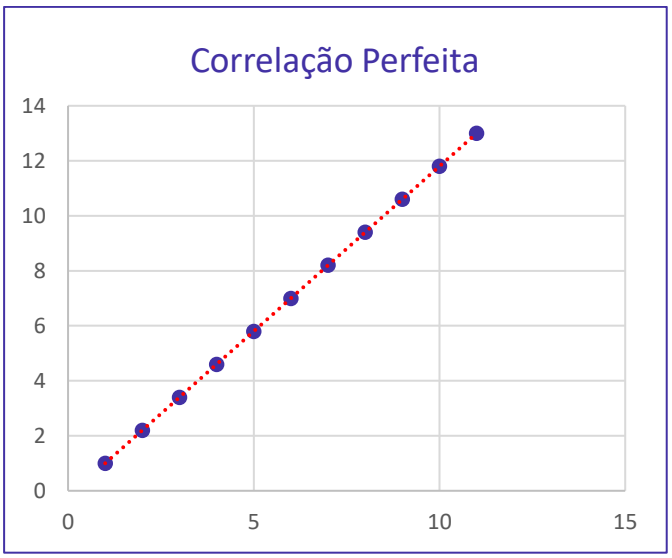
d) perfeita – quando os fenômenos se ajustam perfeitamente a uma reta.

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:

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

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

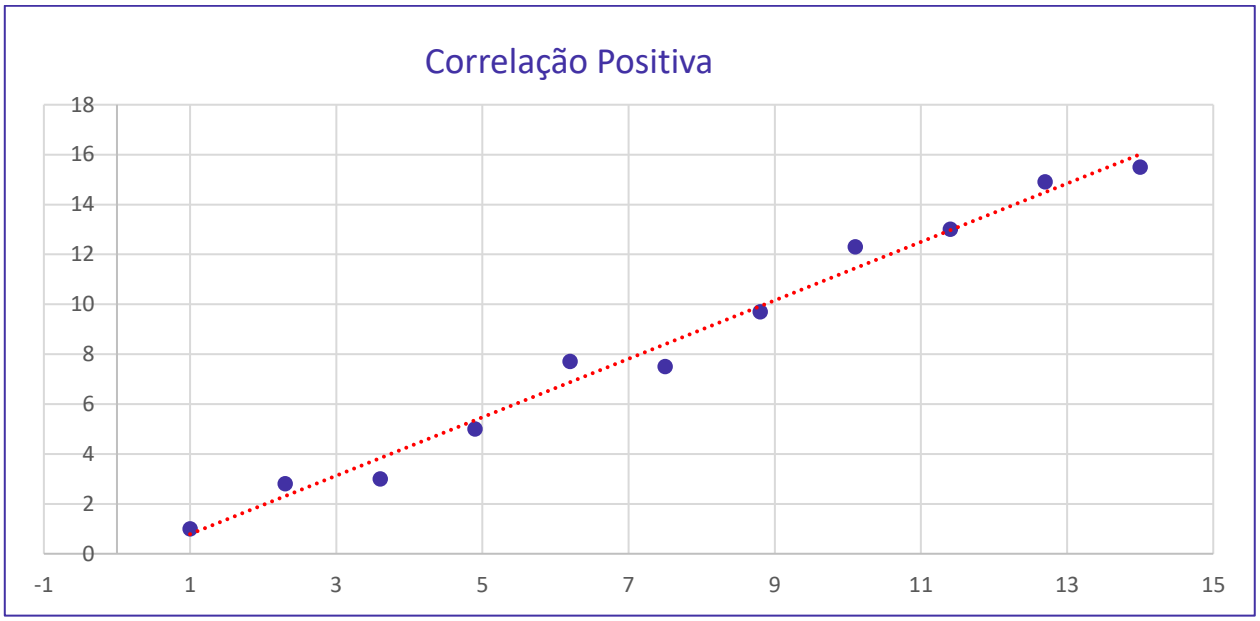
As últimas duas fórmulas são formas alternativas que podem ser empregadas no cálculo do denominador do coeficiente de correlação.

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

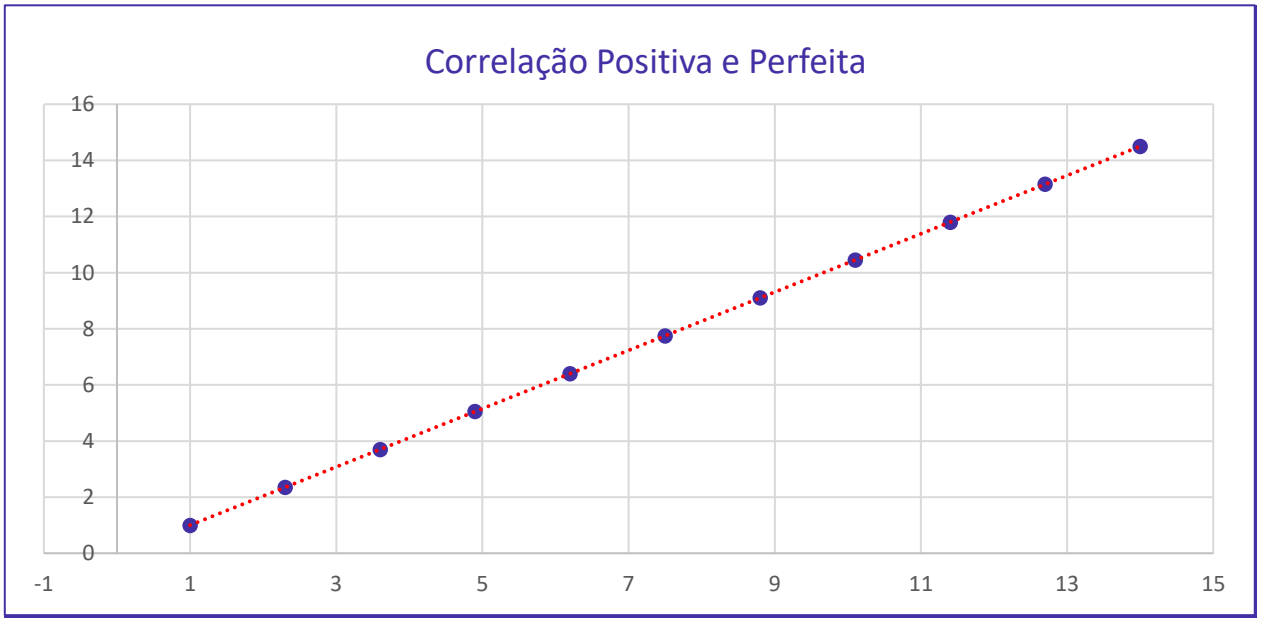
$$\boldsymbol{-1 \leq r \leq 1}$$

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.

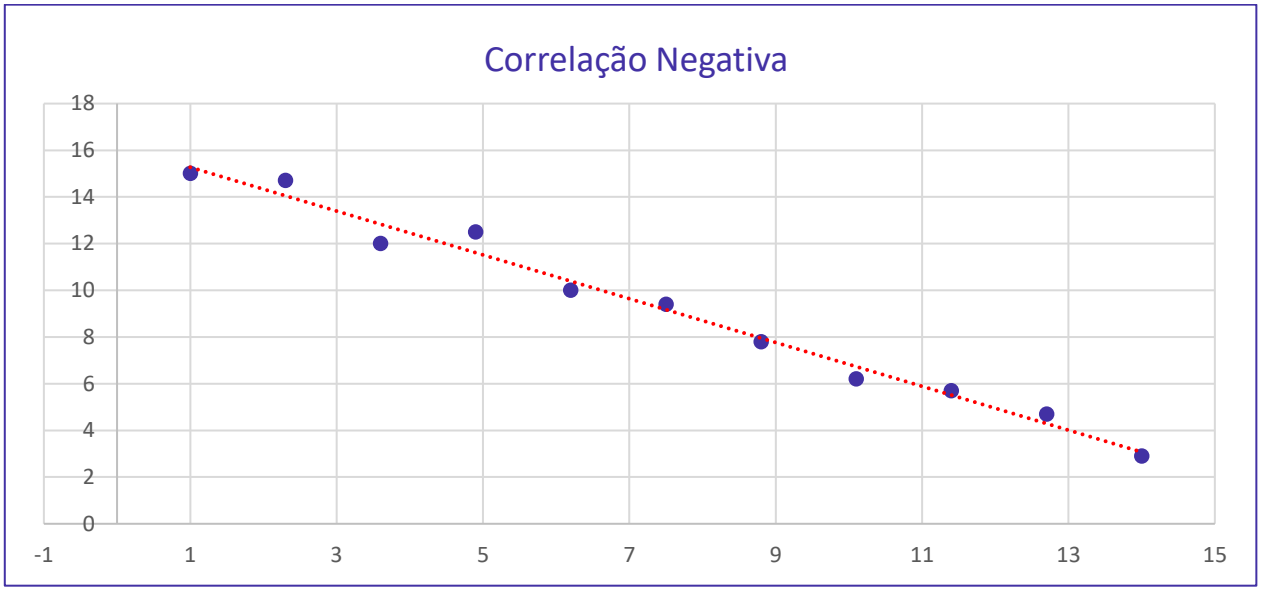


Se a correlação for positiva e todos os pontos estiverem sobre uma mesma reta, o valor de r será exatamente 1. Nesse caso, dizemos que a correlação é **positiva e perfeita**.

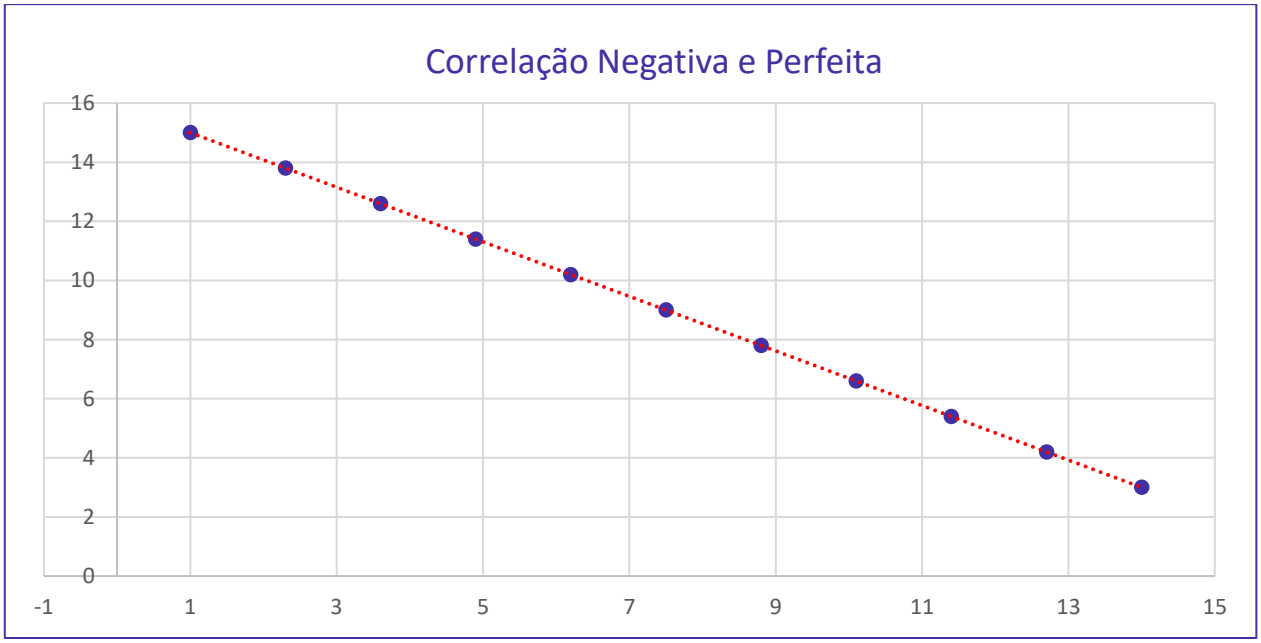


O valor de r é negativo quando a variável Y tende a diminuir ou aumentar quando X aumentar ou diminuir, respectivamente. Nessa situação, dizemos que as variáveis estão negativamente correlacionadas. No

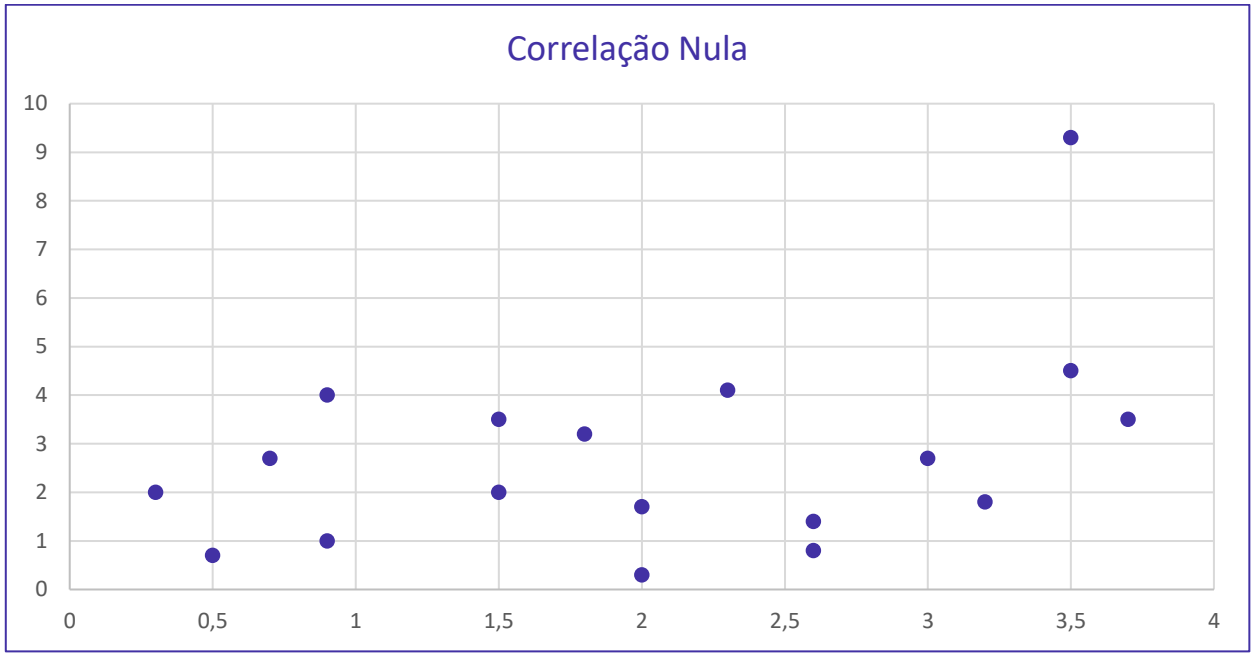
exemplo a seguir, os dados estão praticamente em cima de uma reta, indicando a existência de uma correlação negativa forte, isto é, r muito próximo de -1. No caso, o coeficiente de correlação foi de -0,9918.



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.

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

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

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

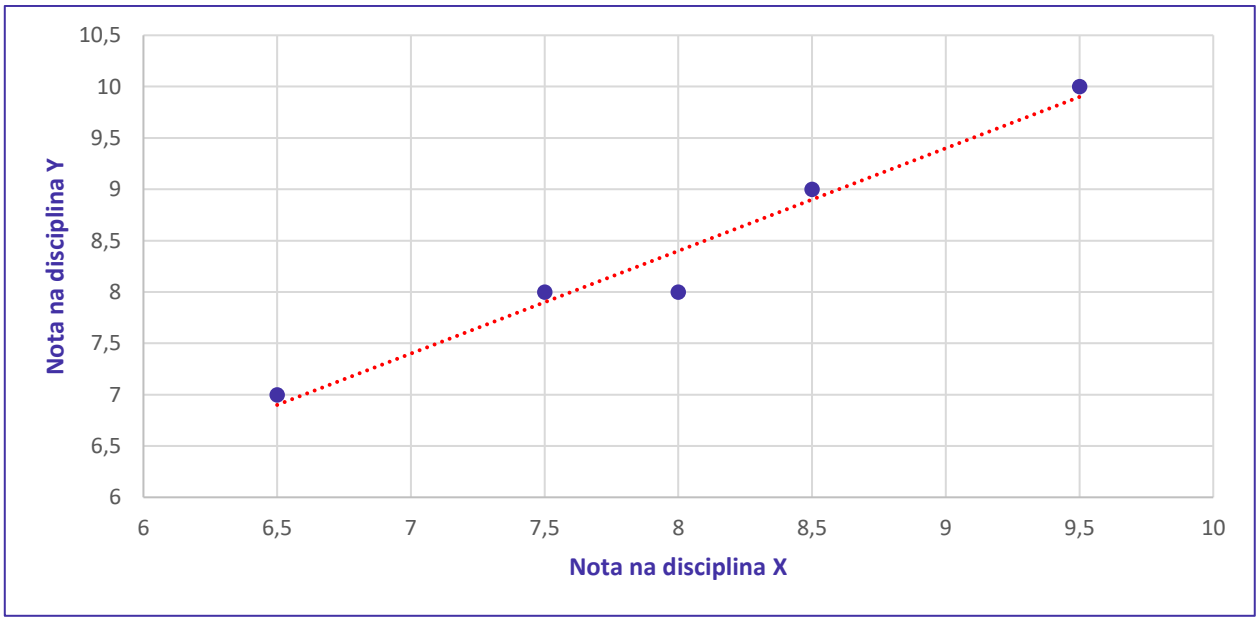
Aplicando esses valores na fórmula do coeficiente de correlação linear, temos:

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

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

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

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.

EXEMPLIFICANDO

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.

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

EXEMPLIFICANDO

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:

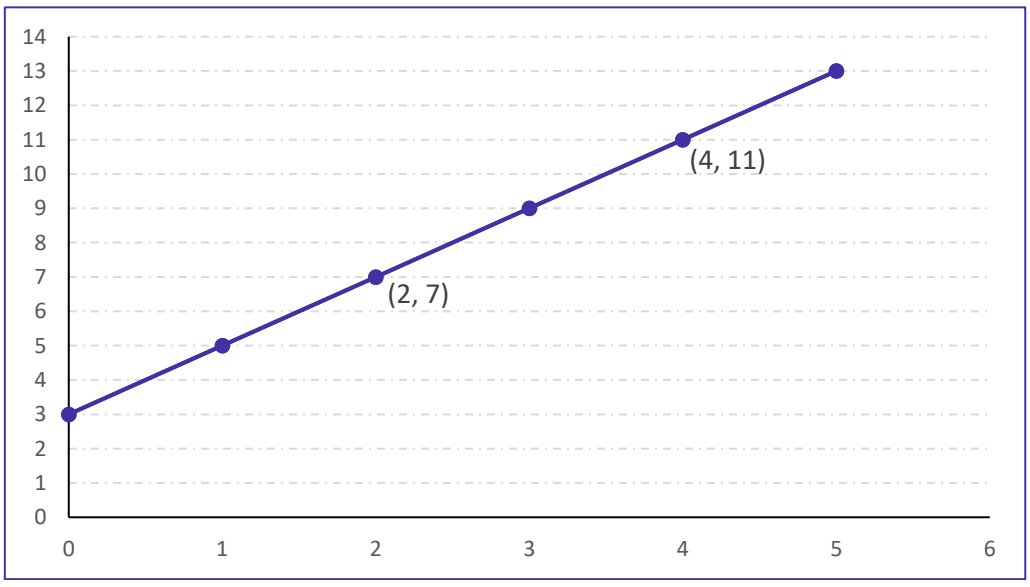
$$y = m \cdot x + b$$

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).

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

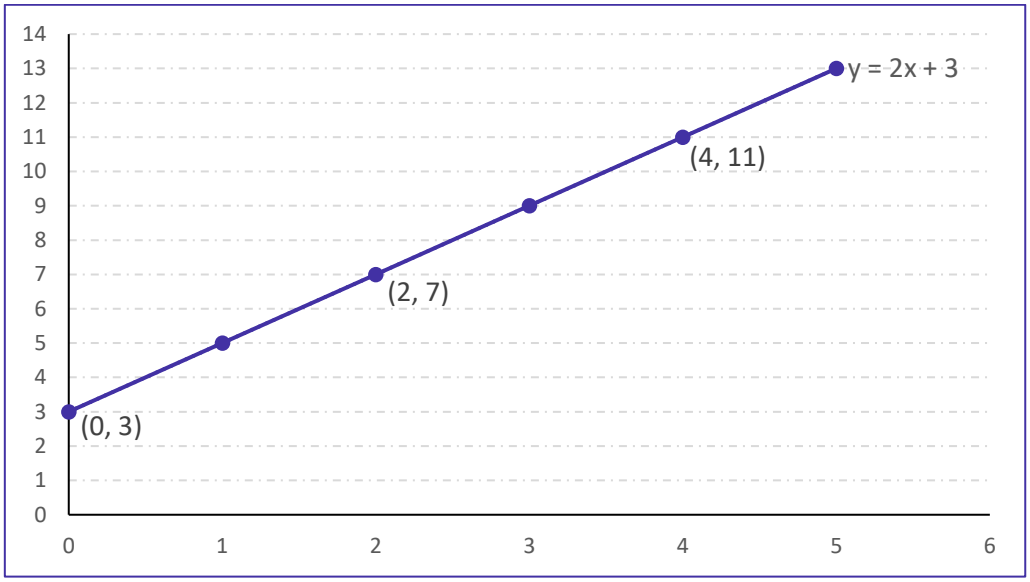
$$7 = 4 + b$$

$$b = 3$$

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:

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

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

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

$$a = 10{,}499$$

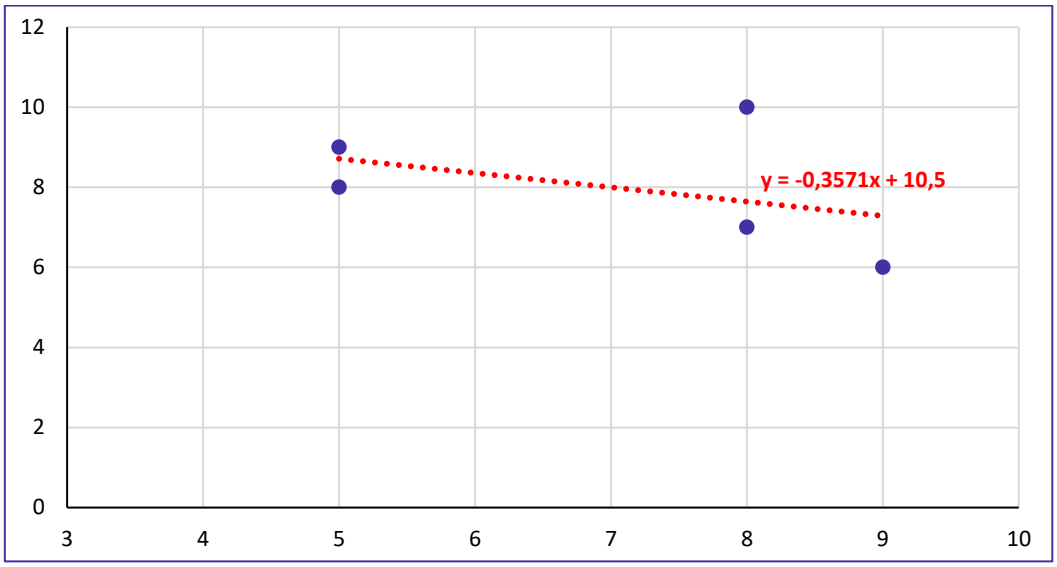
Assim, a reta de regressão estimada é:

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

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

(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 $\widehat{Y} = \widehat{\alpha} + \widehat{\beta}X$, 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:

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

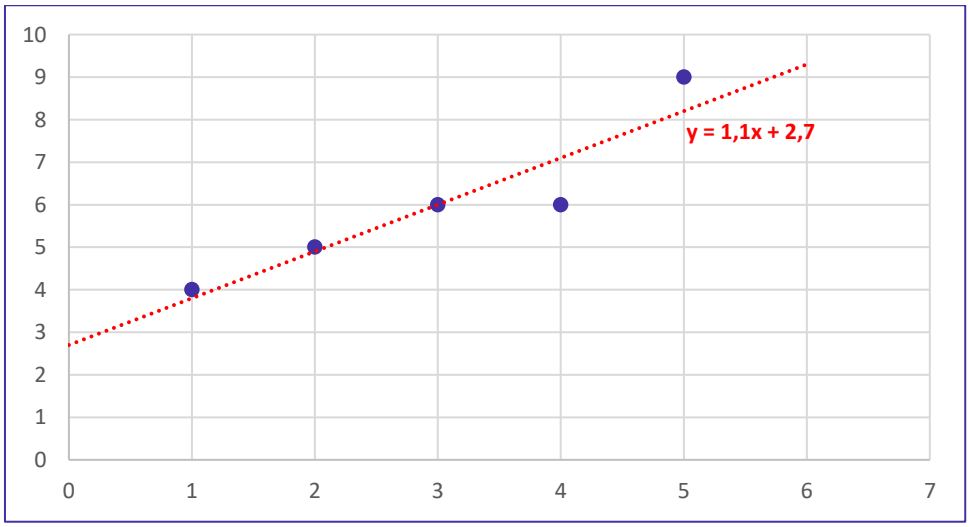
Nesse caso, não há garantia de que o somatório dos resíduos seja igual a zero.

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:

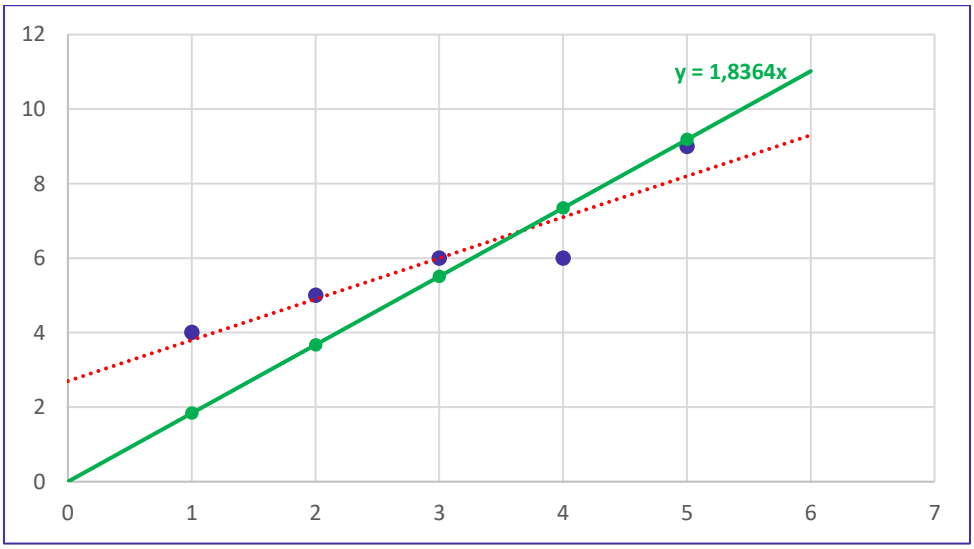
$$b = \frac{\sum X_i \times Y_i}{\sum {X_i}^2} = \frac{101}{55} \cong 1{,}836$$

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:

$$n - 1 = 1 + GL_{Resíduos}$$

$$GL_{Resíduos} = n - 2$$

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,

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

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 $\bar{Y}$.

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

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 $\sum \left(\hat{Y}_i - \bar{Y}\right)^2$:

$$SQT = SQM + SQR$$

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:

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

Podemos demonstrar que:

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

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 $\widehat{Y}_i$. Assim, podemos definir a **soma dos quadrados dos erros (resíduos).**

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

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:

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

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:

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

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

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

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

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

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

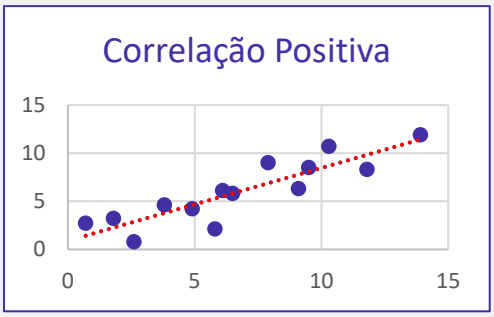
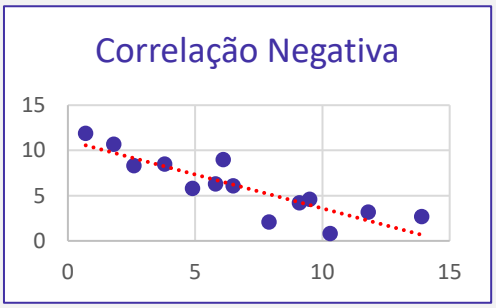
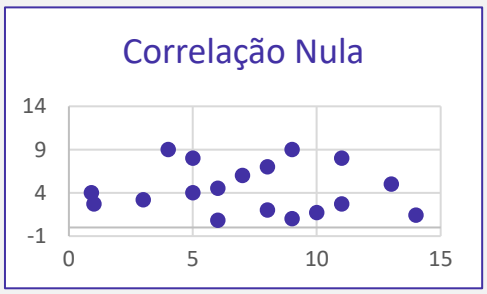
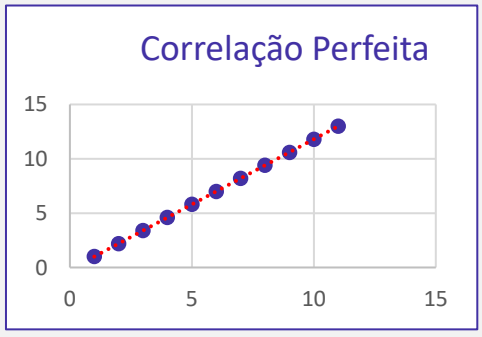
$$\overline{R^2} = 0{,}89.$$

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 $\boldsymbol{SQR}$ e $\boldsymbol{SQT}$ pelos respectivos **graus de liberdade**:

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

A **relação** entre o coeficiente de determinação **ajustado** ($\boldsymbol{\overline{R^2}}$) e o coeficiente de determinação **tradicional** ($\boldsymbol{R^2}$) 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 $\boldsymbol{F}$ **(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	$\boldsymbol{n-1}$	$\boldsymbol{SQT}$	$\boldsymbol{QMT = \frac{SQT}{n-1}}$	

QUESTÕES COMENTADAS

Correlação Linear

1. (CESPE/TJ-PA/2020) Em um gráfico de dispersão, por meio de transformações convenientes, a origem foi colocada no centro da nuvem de dispersão e as variáveis foram reduzidas a uma mesma escala. Se, nesse gráfico, for observado que a grande maioria dos pontos está situada no segundo e no quarto quadrantes, e que aqueles que não estão nessa posição situam-se próximos da origem, então a correlação linear entre as variáveis

a) Será necessariamente fortemente positiva.

b) Poderá ser fracamente positiva.

c) Será necessariamente nula.

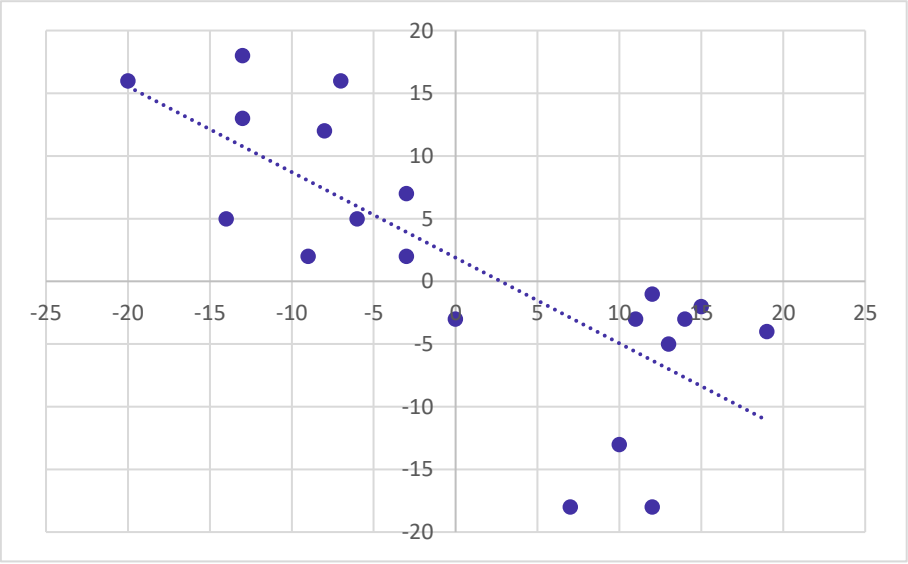
d) Poderá ser fracamente negativa.

e) Será necessariamente fortemente negativa.

Comentários:

Sabemos que o coeficiente de correlação varia entre -1 e 1. Também sabemos 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.

Temos do enunciado que a maioria dos pontos estão no segundo e no quarto quadrantes, logo a reta que representa os dados é decrescente (correlação linear negativa). Sabemos também que os pontos no primeiro e terceiro quadrantes estão mais próximos de zero (correlação fraca).



Assim, concluímos que o gabarito é a letra D.

Gabarito: D.

2. (CESPE/TJ-PA/2020) Texto 7A3-I. O coeficiente de correlação linear de Pearson entre duas variáveis aleatórias discretas X e Y definidas sobre um mesmo espaço amostral é dado por

$$CORR(X,Y) = \frac{n(\sum_{i=1}^{n} x_i y_i) - (\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{\sqrt{n(\sum_{i=1}^{n} x_i^2) - (\sum_{i=1}^{n} x_i)^2}\sqrt{n(\sum_{i=1}^{n} y_i^2) - (\sum_{i=1}^{n} y_i)^2}}.$$

Já na reta de melhor ajuste $Y = aX + b$, determinada pelo método dos mínimos quadrados, os coeficientes são dados por

$$a = \frac{n(\sum_{i=1}^{n} x_i y_i) - (\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n(\sum_{i=1}^{n} x_i^2) - (\sum_{i=1}^{n} x_i)^2}$$

$$b = \frac{\sum_{i=1}^{n} y_i - a\sum_{i=1}^{n} x_i}{n}$$

Uma forma de avaliar a precisão do modelo consiste em comparar o estimador não viesado da variância residual, obtido das diferenças entre os valores observados e os previstos pelo modelo, $\widehat{S}_e = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2$, com o estimador não viesado da variância dos valores observados, $\widehat{S}_e = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2$.

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, em que a última coluna foi obtida usando a reta ajustada pelo método dos mínimos quadrados.

Réu	P	R	$P \times R$	P^2	R^2	$(R - \overline{R})^2$	$(R - \widehat{R})^2$
1	14	0,25	3,5	196	0,0625	3,0625	0,0547560
2	12	0,5	6	144	0,25	2,25	0,0001440
3	10,9	1	10,9	118,81	1	1	0,0463110
4	6	1,5	9	36	2,25	0,25	0,2500000
5	5	1,75	8,75	25	3,0625	0,0625	0,2480040
6	3	2	6	9	4	0	0,5535360
7	3	2,5	7,5	9	6,25	0,25	0,0595360
8	2,3	3	6,9	5,29	9	1	0,0067898
9	1,8	3,5	6,3	3,24	12,25	2,25	0,2101306

10	2	4	8	4	16	4	1,0160640
Totais	**60**	**20**	**72,85**	**550,34**	**54,125**	**14,125**	**2,4452714**

Dados:

$\mathbf{1903,4^{1/2} = 43,63}$

$\mathbf{141,25^{1/2} = 11,88}$

A partir das informações do texto 7A3-I, o coeficiente de correlação linear entre as variáveis R e P é

a) – 0,33.

b) – 0,51.

c) – 0,67.

d) – 0,82.

e) – 0,91.

Comentários:

Nessa questão, a fórmula do coeficiente de correlação linear veio no próprio enunciado, restando apenas a aplicação dos valores apresentados na tabela. Assim, considerando P como a variável X e R como a variável Y, temos:

$$r(X,Y) = \frac{n(\sum x_i y_i) - (\sum x_i)(\sum y_i)}{\sqrt{n(\sum x_i^2) - (\sum x_i)^2}\sqrt{n(\sum y_i^2) - (\sum y_i)^2}}$$

$$r(X,Y) = \frac{10 \times 72{,}85 - 60 \times 20}{\sqrt{10 \times 550{,}34 - 60^2}\sqrt{10 \times 54{,}125 - 20^2}}$$

$$r(X,Y) = \frac{728{,}5 - 1200}{\sqrt{5503{,}4 - 3600}\sqrt{541{,}25 - 400}}$$

$$r(X,Y) = \frac{-471{,}5}{\sqrt{1903{,}4}\sqrt{141{,}25}}$$

Seria praticamente impossível passarmos desse ponto sem a ajuda de uma calculadora. Contudo, o enunciado também trouxe alguns dados importantes, que nos ajudam a superar esta etapa:

$1903{,}4^{1/2} = 43{,}63$

$141{,}25^{1/2} = 11{,}88$

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

$$r(X,Y) = \frac{-471,5}{43,63 \times 11,88}$$

$$r(X,Y) = \frac{-471,5}{518,32}$$

$$r(X,Y) = -0,91$$

Gabarito: E.

3. (CESPE/TJ-PA/2020) Texto 7A3-I. O coeficiente de correlação linear de Pearson entre duas variáveis aleatórias discretas X e Y definidas sobre um mesmo espaço amostral é dado por

$$CORR(X,Y) = \frac{n(\sum_{i=1}^{n} x_i y_i) - (\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{\sqrt{n(\sum_{i=1}^{n} x_i^2) - (\sum_{i=1}^{n} x_i)^2}\sqrt{n(\sum_{i=1}^{n} y_i^2) - (\sum_{i=1}^{n} y_i)^2}}.$$

Já na reta de melhor ajuste $Y = aX + b$, determinada pelo método dos mínimos quadrados, os coeficientes são dados por

$$a = \frac{n(\sum_{i=1}^{n} x_i y_i) - (\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n(\sum_{i=1}^{n} x_i^2) - (\sum_{i=1}^{n} x_i)^2}$$

$$b = \frac{\sum_{i=1}^{n} y_i - a\sum_{i=1}^{n} x_i}{n}$$

Uma forma de avaliar a precisão do modelo consiste em comparar o estimador não viesado da variância residual, obtido das diferenças entre os valores observados e os previstos pelo modelo, $\widehat{S}_e = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$, com o estimador não viesado da variância dos valores observados, $\widehat{S}_e = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2$.

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, em que a última coluna foi obtida usando a reta ajustada pelo método dos mínimos quadrados.

Réu	P	R	$P \times R$	P^2	R^2	$(R - \overline{R})^2$	$(R - \widehat{R})^2$
1	14	0,25	3,5	196	0,0625	3,0625	0,0547560
2	12	0,5	6	144	0,25	2,25	0,0001440
3	10,9	1	10,9	118,81	1	1	0,0463110
4	6	1,5	9	36	2,25	0,25	0,2500000
5	5	1,75	8,75	25	3,0625	0,0625	0,2480040

6	3	2	6	9	4	0	0,5535360
7	3	2,5	7,5	9	6,25	0,25	0,0595360
8	2,3	3	6,9	5,29	9	1	0,0067898
9	1,8	3,5	6,3	3,24	12,25	2,25	0,2101306
10	2	4	8	4	16	4	1,0160640
Totais	**60**	**20**	**72,85**	**550,34**	**54,125**	**14,125**	**2,4452714**

Dados:

$\mathbf{1903,4^{1/2} = 43,63}$

$\mathbf{141,25^{1/2} = 11,88}$

Considerando-se o texto 7A3-I, a relação entre o coeficiente de correlação linear entre as variáveis X e Y e o coeficiente angular, da reta de melhor ajuste aos dados determinada pelo método dos mínimos quadrados pode ser expressa por

a) $a = CORR(X, Y)$.

b) $b = CORR(X, Y)$.

c) $a \times \sqrt{n(\sum_{i=1}^{n} x_i^2) - (\sum_{i=1}^{n} x_i)^2} = CORR(X, Y) \times \sqrt{n(\sum_{i=1}^{n} y_i^2) - (\sum_{i=1}^{n} y_i)^2}$.

d) $b \times \sqrt{n(\sum_{i=1}^{n} x_i^2) - (\sum_{i=1}^{n} x_i)^2} = CORR(X, Y) \times \sqrt{n(\sum_{i=1}^{n} y_i^2) - (\sum_{i=1}^{n} y_i)^2}$.

e) $a = \frac{1}{CORR(X,Y)}$.

Comentários:

O enunciado da questão nos deu algumas fórmulas importantes. Temos que o coeficiente de correlação linear é dado por:

$$r(X, Y) = \frac{n(\sum x_i y_i) - (\sum x_i)(\sum y_i)}{\sqrt{n(\sum x_i^2) - (\sum x_i)^2}\sqrt{n(\sum y_i^2) - (\sum y_i)^2}} \qquad (Eq. 1)$$

Para facilitar a resolução podemos substituir as expressões por letras. Assim, podemos dizer que:

$$P = n\left(\sum_{i=1}^{n} x_i y_i\right) - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)$$

$$Q = n\left(\sum_{i=1}^{n} x_i^2\right) - \left(\sum_{i=1}^{n} x_i\right)^2$$

Assim, temos:

$$a = \frac{P}{Q} \qquad (Eq.2)$$

Agora, vamos considerar que:

$$R = n\left(\sum_{i=1}^{n} y_i^2\right) - \left(\sum_{i=1}^{n} y_i\right)^2$$

Dessa forma, temos:

$$r(X,Y) = \frac{P}{\sqrt{Q}\sqrt{R}} \ (Eq.3)$$

Reparem que esta última expressão corresponde a uma simplificação da Eq. 1.

Se pegarmos a Eq. 2 e isolarmos o P, teremos o seguinte:

$$P = a \times Q$$

Substituindo P na Eq. 3:

$$r(X,Y) = \frac{a \times Q}{\sqrt{Q}\sqrt{R}}$$

Simplificando por $\sqrt{Q}$:

$$r(X,Y) = \frac{a \times Q \times \sqrt{Q}}{\sqrt{Q}\sqrt{Q}\sqrt{R}}$$

$$r(X,Y) = \frac{a \times Q \times \sqrt{Q}}{Q\sqrt{R}}$$

$$r(X,Y) = \frac{a \times \sqrt{Q}}{\sqrt{R}}$$

Se colocarmos $\sqrt{R}$ multiplicando teremos:

$$a \times \sqrt{Q} = r(X,Y) \times \sqrt{R}$$

Já podemos determinar a alternativa correta, mas vamos substituir as letras pelas devias expressões:

$$a \times \sqrt{n\left(\sum x_i^2\right) - \left(\sum x_i\right)^2} = CORR(X,Y) \times \sqrt{n\left(\sum y_i^2\right) - \left(\sum y_i\right)^2}$$

Gabarito: C.

4. (CESPE/TJ-PA/2020) Texto 7A3-I. O coeficiente de correlação linear de Pearson entre duas variáveis aleatórias discretas X e Y definidas sobre um mesmo espaço amostral é dado por

$$CORR(X,Y) = \frac{n(\sum_{i=1}^{n} x_i y_i) - (\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{\sqrt{n(\sum_{i=1}^{n} x_i^2) - (\sum_{i=1}^{n} x_i)^2}\sqrt{n(\sum_{i=1}^{n} y_i^2) - (\sum_{i=1}^{n} y_i)^2}}.$$

Já na reta de melhor ajuste $Y = aX + b$, determinada pelo método dos mínimos quadrados, os coeficientes são dados por

$$a = \frac{n(\sum_{i=1}^{n} x_i y_i) - (\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n(\sum_{i=1}^{n} x_i^2) - (\sum_{i=1}^{n} x_i)^2}$$

$$b = \frac{\sum_{i=1}^{n} y_i - a\sum_{i=1}^{n} x_i}{n}$$

Uma forma de avaliar a precisão do modelo consiste em comparar o estimador não viesado da variância residual, obtido das diferenças entre os valores observados e os previstos pelo modelo, $\widehat{S}_e = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$, com o estimador não viesado da variância dos valores observados, $\widehat{S}_e = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$.

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, em que a última coluna foi obtida usando a reta ajustada pelo método dos mínimos quadrados.

Réu	P	R	$P \times R$	P^2	R^2	$(R - \bar{R})^2$	$(R - \widehat{R})^2$
1	14	0,25	3,5	196	0,0625	3,0625	0,0547560
2	12	0,5	6	144	0,25	2,25	0,0001440
3	10,9	1	10,9	118,81	1	1	0,0463110
4	6	1,5	9	36	2,25	0,25	0,2500000
5	5	1,75	8,75	25	3,0625	0,0625	0,2480040
6	3	2	6	9	4	0	0,5535360
7	3	2,5	7,5	9	6,25	0,25	0,0595360
8	2,3	3	6,9	5,29	9	1	0,0067898
9	1,8	3,5	6,3	3,24	12,25	2,25	0,2101306
10	2	4	8	4	16	4	1,0160640

Totais	**60**	**20**	**72,85**	**550,34**	**54,125**	**14,125**	**2,4452714**

Dados:

$\mathbf{1903,4^{1/2} = 43,63}$

$\mathbf{141,25^{1/2} = 11,88}$

Com base no texto 7A3-I, a renda familiar per capita esperada X, em número de salários-mínimos, obtida aplicando-se a reta de melhor ajuste aos dados determinada pelo método dos mínimos quadrados para um réu ao qual tenha sido cominada uma pena de 4 anos de reclusão é

a) $2{,}3 < X < 2{,}6$.

b) $2{,}1 < X < 2{,}3$.

c) $1{,}9 < X < 2{,}1$.

d) $1{,}2 < X < 1{,}9$.

e) $1{,}0 < X < 1{,}2$.

Comentários:

O enunciado da questão nos forneceu a fórmula para calcularmos o coeficiente de correlação linear. Assim, basta aplicamos os valores da tabela à fórmula.

Tomemos P para x e R para y. Assim, temos:

$$a = \frac{n(\sum x_i y_i) - (\sum x_i)(\sum y_i)}{n(\sum x_i^2) - (\sum x_i)^2}$$

$$b = \frac{\sum_{i=1}^{n} y_i - a\sum_{i=1}^{n} x_i}{n}$$

Substituindo os valores da tabela, temos:

$$a = \frac{10 \times 72{,}85 - 60 \times 20}{10 \times 550{,}34 - 60^2} =$$

$$a = \frac{728{,}5 - 1200}{5503{,}4 - 3600} =$$

$$a = \frac{-471{,}8}{1903{,}4} =$$

$$a = -0{,}24$$

Calculando b:

$$b = \frac{\sum_{i=1}^{n} y_i - a\sum_{i=1}^{n} x_i}{n}$$

$$b = \frac{20 - (-0{,}24 \times 60)}{10} =$$

$$b = \frac{20 - (-14{,}4)}{10}$$

$$b = \frac{34{,}4}{10}$$

$$b = 3{,}44$$

Agora, podemos calcular a reta para um réu ao qual tenha sido cominada uma pena de 4 anos de reclusão, tomando $X = 4$

$$Y = aX + b$$

$$Y = -0{,}24 \times 4 + 3{,}44$$

$$Y = 2{,}48$$

Gabarito: A.

5. (CESPE/TJ-PA/2020) Texto 7A3-I. O coeficiente de correlação linear de Pearson entre duas variáveis aleatórias discretas X e Y definidas sobre um mesmo espaço amostral é dado por

$$CORR(X,Y) = \frac{n(\sum_{i=1}^{n} x_i y_i) - (\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{\sqrt{n(\sum_{i=1}^{n} x_i^2) - (\sum_{i=1}^{n} x_i)^2}\sqrt{n(\sum_{i=1}^{n} y_i^2) - (\sum_{i=1}^{n} y_i)^2}}.$$

Já na reta de melhor ajuste $Y = aX + b$, determinada pelo método dos mínimos quadrados, os coeficientes são dados por

$$a = \frac{n(\sum_{i=1}^{n} x_i y_i) - (\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n(\sum_{i=1}^{n} x_i^2) - (\sum_{i=1}^{n} x_i)^2}$$

$$b = \frac{\sum_{i=1}^{n} y_i - a\sum_{i=1}^{n} x_i}{n}$$

Uma forma de avaliar a precisão do modelo consiste em comparar o estimador não viesado da variância residual, obtido das diferenças entre os valores observados e os previstos pelo modelo, $\widehat{S}_e = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$, com o estimador não viesado da variância dos valores observados, $\widehat{S}_e = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2$.

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, em que a última coluna foi obtida usando a reta ajustada pelo método dos mínimos quadrados.

Réu	P	R	$P \times R$	P^2	R^2	$(R - \overline{R})^2$	$(R - \widehat{R})^2$
1	14	0,25	3,5	196	0,0625	3,0625	0,0547560
2	12	0,5	6	144	0,25	2,25	0,0001440

3	10,9	1	10,9	118,81	1	1	0,0463110
4	6	1,5	9	36	2,25	0,25	0,2500000
5	5	1,75	8,75	25	3,0625	0,0625	0,2480040
6	3	2	6	9	4	0	0,5535360
7	3	2,5	7,5	9	6,25	0,25	0,0595360
8	2,3	3	6,9	5,29	9	1	0,0067898
9	1,8	3,5	6,3	3,24	12,25	2,25	0,2101306
10	2	4	8	4	16	4	1,0160640
Totais	**60**	**20**	**72,85**	**550,34**	**54,125**	**14,125**	**2,4452714**

Dados:

$\mathbf{1903,4^{1/2} = 43,63}$

$\mathbf{141,25^{1/2} = 11,88}$

Levando-se em consideração o texto 7A3-I, a discrepância na renda familiar per capita X, em número de salários-mínimos, obtida entre o valor observado e aquele em que se aplica a reta de melhor ajuste aos dados determinada pelo método dos mínimos quadrados para o nono réu é

a) $0{,}47 < X < 0{,}50$.

b) $0{,}44 < X < 0{,}47$.

c) $0{,}42 < X < 0{,}44$.

d) $0{,}39 < X < 0{,}42$.

e) $0{,}38 < X < 0{,}39$.

Comentários:

A reta de regressão é dada por $Y = aX + b$, como a questão não apresenta uma correlação perfeita, existem desvios ou "erros" de cada ponto em relação à reta de regressão. Assim, os valores desses "erros" são calculados pela diferença entre os valores observados e os valores obtidos pela reta de regressão.

A questão traz na tabela os quadrados das diferenças entre cada valor R_i observado e o valor $\hat{R}_i$ da reta de regressão. Portanto, para acharmos a diferença para o nono réu, basta calcularmos a raiz de $(R_i - \hat{R}_i)^2$, dado na tabela.

Para o réu 9:

$$(R_i - \hat{R}_i)^2 = 0{,}2101306$$

$$(R_i - \hat{R}_i) \cong \sqrt{0{,}21}$$

$$(R_i - \hat{R}_i) \cong 0{,}45$$

Gabarito: B.

6. (VUNESP/EBSERH/2020) Dados para responder à questão.

A variável x tem média 4 e desvio padrão 2, enquanto a variável y tem média 3 e desvio padrão 1. A covariância entre x e y é –1.

O coeficiente de correlação entre x e y é

a) 0,5.

b) –0,5.

c) 1.

d) –1.

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:

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

$$\rho(x,y) = -0{,}5$$

Gabarito: B.

QUESTÕES COMENTADAS

Regressão Linear Simples

1. (CESPE/SEFAZ-SE/2022) Para a obtenção de projeções de resultados financeiros de empresas de determinado ramo de negócios, será ajustado um modelo de regressão linear simples na forma $y = ax + b + \epsilon$, no qual x representa o grau de endividamento; y denota um índice contábil; o termo ϵ é o erro aleatório, que segue uma distribuição com média nula e variância σ^2; e a e b são os coeficientes do modelo, com $b \neq 0$. A correlação linear entre as variáveis x e y é positiva e algumas medidas descritivas referentes às variáveis x e y se encontram na tabela a seguir.

	y	x
Média Amostral	2	4
Desvio Padrão Amostral	0,4	8

Com base nessa situação hipotética e considerando que o coeficiente de determinação proporcionado pelo modelo em tela seja $R^2 = 0,81$, assinale a opção em que é apresentada a reta ajustada pelo critério de mínimos quadrados ordinários.

a) $\hat{y} = 0,045x + 1,82$

b) $\hat{y} = 0,5x$

c) $\hat{y} = 0,4x + 0,4 + \epsilon$

d) $\hat{y} = 18x - 70 + \epsilon$

e) $\hat{y} = 18x - 70$

Comentários:

Os coeficientes a e b podem ser estimados pelas seguintes relações:

$$\hat{a} = \bar{y} - \hat{b}\bar{x}$$

$$\hat{b} = r \cdot \frac{s_y}{s_x}$$

em que r é o coeficiente de correlação; s_y e s_x são os desvios amostrais de y e x.

Aplicando os valores $R^2 = 0,81$, $s_y = 0,4$ e $s_x = 8$, temos:

$$\hat{b} = \sqrt{0{,}81} \cdot \left(\frac{0{,}4}{8}\right)$$

$$\hat{b} = 0{,}9 \cdot (0{,}05) = 0{,}045$$

Agora, utilizando o valor de $\hat{b} = 0{,}045$, $\bar{y} = 2$ e $\bar{x} = 4$, temos:

$$\hat{a} = \bar{y} - \hat{b}\bar{x}$$

$$\hat{a} = 2 - 0{,}045 \times 4$$

$$\hat{a} = 2 - 0{,}18 = 1{,}82$$

Portanto, a reta ajustada pelo modelo é descrita por:

$$\hat{y} = \hat{b}x + \hat{a}$$

$$\hat{y} = 0{,}045x + 1{,}82$$

Gabarito: A.

2. (CESPE/PC-PB/2022) Para as variáveis *Y* e *X*, em que *Y* denota a variável resposta e *X* representa a variável regressora, a correlação linear de Pearson entre *Y* e *X* é $0,8$, o desvio padrão amostral de *Y* é 2, e o desvio padrão amostral de *X* é 4. Nesse caso, a estimativa de mínimos quadrados ordinários do coeficiente angular da reta de regressão linear simples é igual a

a) 0,40.

b) 1,60.

c) 0,64.

d) 0,80.

e) 0,50.

Comentários:

O coeficiente angular da reta de regressão pode ser calculada por

$$\hat{b} = r \cdot \frac{s_y}{s_x}$$

Como $r = 0{,}8$, $s_y = 2$ e $s_x = 4$, temos:

$$\hat{\beta} = 0{,}8 \times \frac{2}{4} = 0{,}4$$

Gabarito: A.

3. (CESPE/PETROBRAS/2022) Uma determinada repartição pública fez um levantamento do tempo, em minutos, que os cinco funcionários de uma sessão gastam para chegar ao trabalho em

função da distância x, em quilômetros, de suas residências. O resultado da pesquisa realizada com cada um deles é apresentado na tabela a seguir, em que $\overline{x}$ e $\overline{y}$ são, respectivamente, as médias amostrais das variáveis x e y .

i	Tempo y_i	Distância x_i	$x_i - \overline{x}$	$y_i - \overline{y}$	$(x_i - \overline{x}).(y_i - \overline{y})$	$(x_i - \overline{x})^2$
1	10	5	-4	-7	28	16
2	20	5	-4	3	-12	16
3	15	10	1	-2	-2	1
4	10	10	1	-7	-7	1
5	30	15	6	13	78	36
Média	17	9				

Com base nos dados dessa tabela, julgue o próximo item.

Pelo modelo de regressão linear simples, a equação que expressa o relacionamento ajustado entre a variável em função de x e $\hat{y}_i = \frac{85}{70}x_i + \alpha$, em que α é uma constante.

Comentários:

O coeficiente angular da reta de regressão é estimado por

$$\hat{b} = \frac{S_{xy}}{S_{xx}} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})\,(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Temos que somar as duas últimas colunas da tabela:

$$S_{xy} = 28 - 12 - 2 - 7 + 78 = 85$$

$$S_{xx} = 16 + 16 + 1 + 1 + 36 = 70$$

Assim, temos:

$$\hat{b} = \frac{85}{70}$$

Portanto, nosso modelo será:

$$\hat{y}_i = \frac{85}{70}x_i + \alpha$$

Gabarito: Certo.

4. (CESPE/PETROBRAS/2022)

Equação 1: $y_i = a + bXi + e$

Equação 2: $y_i = a + b1X_i + b2X_2 + b3y_i3 + e$

Com base nos modelos de regressão linear simples (equação 1) e de regressão linear múltipla (equação 2), julgue o item a seguir.

O coeficiente b da equação 1 é o resultado da correlação entre os valores amostrais de X e Y, dividida pela variância de X .

Comentários:

A estimativa do coeficiente $\hat{b}$, pelo método dos mínimos quadrados ordinários, é o resultado da covariância entre os valores amostrais de X e Y, dividida pela variância de X:

$$\hat{b} = \frac{Cov(X,Y)}{Var(X)}$$

Gabarito: Errado.

5. (CESPE/TELEBRAS/2022) O quadro a seguir mostra as estimativas de mínimos quadrados ordinários dos coeficientes de um modelo de regressão linear simples na forma $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, em que $i \in \{1, \cdots, 6\}$ e ϵ_i representa o erro aleatório com média zero e variância σ^2.

Coeficiente	Estimativa	Erro Padrão	Razão t
β_0	0,9	0,10	9
β_1	0,2	0,05	4

Considerando essas informações e sabendo que $\sigma^2 = 0,01$, julgue o item seguinte.

$S_{xx} = \sum_{i=1}^{6}(x_i - \bar{x})^2 = 4$ em que $\bar{x} = \sum_{i=1}^{6} x_i/6$.

Comentários:

A variância do estimador de β_1 é definida como:

$$Var(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sigma^2}{S_{xx}}$$

Do enunciado, temos:

$$Var(\hat{\beta}_1) = 0,05^2 = 0,0025$$

e

$$\sigma^2 = 0{,}01.$$

Portanto,

$$0{,}0025 = \frac{0{,}01}{S_{xx}}$$

$$S_{xx} = \frac{0{,}01}{0{,}0025} = 4$$

Gabarito: Certo.

6. (CESPE/TELEBRAS/2022) O quadro a seguir mostra as estimativas de mínimos quadrados ordinários dos coeficientes de um modelo de regressão linear simples na forma $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, em que $i \in \{1, \cdots, 6\}$ e ϵ_i representa o erro aleatório com média zero e variância σ^2.

Coeficiente	Estimativa	Erro Padrão	Razão t
β_0	0,9	0,10	9
β_1	0,2	0,05	4

Considerando essas informações e sabendo que $\sigma^2 = 0,01$, julgue o item seguinte.

A covariância entre a variável resposta (y) e a variável explicativa (x) é igual ou superior a 0,2.

Comentários:

A variância do estimador de β_1 é definida como

$$Var(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sigma^2}{S_{xx}}$$

Do enunciado, temos:

$$Var(\hat{\beta}_1) = 0{,}05^2 = 0{,}0025$$

e

$$\sigma^2 = 0{,}01.$$

Portanto,

$$0{,}0025 = \frac{0{,}01}{S_{xx}}$$

$$S_{xx} = \frac{0{,}01}{0{,}0025} = 4$$

A variância amostral é:

$$Var(X) = \frac{S_{xx}}{n-1} = \frac{4}{5}$$

A covariância entre as variáveis X e Y é obtida por meio da relação:

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{Cov(X,Y)}{Var(X)}$$

Fazendo as substituições, temos:

$$Cov(X,Y) = 0{,}2 \times \frac{4}{5} = 0{,}16 < 0{,}2$$

Gabarito: Errado.

7. (CESPE/TCE-SC/2022) Em artigo publicado em 2004 no Journal of Political Economy, E. Miguel, S. Satyanath e E. Sergenti mostraram o efeito que o crescimento econômico pode ter na ocorrência de conflitos civis, com dados de 41 países africanos, no período de 1981 até 1999. Em certo estágio da pesquisa, para verificar a possibilidade de usar dados sobre precipitação pluviométrica como variável instrumental, foi feita uma regressão entre o crescimento de tais precipitações (variável explicativa) e uma variável resposta que representa um indicador para a ocorrência de conflito: quanto maior for esse indicador, maior a possibilidade de conflitos no ano *t* no país *i*. Os resultados do modelo ajustado pelo método de mínimos quadrados ordinários se encontram na tabela a seguir.

Variável Explicativa	Variável Dependente	
	Conflito civil (mínimo de 25 mortos)	Conflito civil (mínimo de 1000 mortos)
Crescimento na precipitação em *t*	–0,024 (0,043)	–0,062 (0,030)
Crescimento na precipitação em *t–1*	–0,122 (0,052)	–0,069 (0,032)
Efeitos fixos	sim	sim
R^2	0,71	0,70
Observações	743	743

Internet: <https://doi.org/10.1086/421174> (com adaptações).

Os números entre parênteses na tabela apresentada indicam o erro padrão da estimativa dos coeficientes respectivos. Considere os valores críticos t_α da variável t de Student, com significância α para os graus de liberdades adequados aos dados apresentados, como sendo $t_{10\%} = 1,65$, $t_{5\%} = 1,96$ e $t_{1\%} = 2,58$. Considerando as informações precedentes, julgue o próximo item.

Os resultados mostram que um aumento na precipitação pluviométrica no ano anterior resulta no aumento na ocorrência de conflito civil, nas duas regressões.

Comentários:

Como os coeficientes das variáveis regressoras são negativos, verificamos que um aumento nas precipitações está associado a uma diminuição na ocorrência de conflito civil. Isso ocorre tanto para o crescimento da precipitação no tempo t quanto no crescimento defasado (t-1).

Gabarito: Errado.

8. (FGV/EPE/2022) Considere o modelo de regressão linear simples, a seguir.

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \qquad i = 1, 2, \cdots, n$$

Para uma amostra de 20 observações, foram obtidos os seguintes resultados:

$$\sum_{i=1}^{20} x_i = 60, \sum_{i=1}^{20} y_i = 90, \sum_{i=1}^{20} {x_i}^2 = 300, \sum_{i=1}^{20} x_i y_i = 510$$

Os estimadores de mínimos quadrados do modelo são, respectivamente,

a) –1,5 e 0,5.

b) –1,5 e 2.

c) –4,5 e 3.

d) 0,5 e 0,5.

e) 0,5 e 2.

Comentários:

O coeficiente angular da reta de regressão é estimado por

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i Y_i) - n \times \bar{X} \times \bar{Y}}{\sum_{i=1}^{n}(X_i^2) - n \times \bar{X}^2}$$

Substituindo os valores fornecidos no enunciado, temos:

$$\hat{\beta}_1 = \frac{510 - 20 \times \left(\frac{60}{20}\right) \times \left(\frac{90}{20}\right)}{300 - 20 \times \left(\frac{60}{20}\right)^2} = \frac{510 - 270}{300 - 180} = \frac{240}{120} = 2$$

Segundo o método dos mínimos quadrados, a reta de regressão sempre passa pelo ponto formado pela média das variáveis dependente e independente. Ou seja, se

$$\hat{y} = \hat{a}x + \hat{b}$$

é a reta de regressão, então o ponto $(\bar{x}, \bar{y})$ pertence a ela.

Assim, temos que:

$$\hat{\beta}_0 = \left(\frac{90}{20}\right) - 2\left(\frac{60}{20}\right) = 4{,}5 - 6 = -1{,}5$$

Gabarito: B.

9. (CESPE/SEFAZ RR/2021) A tabela a seguir apresenta uma amostra aleatória simples formada por 5 pares de valores (X_i, Y_i), em que $i = 1, 2, \cdots, 5$, X_i é uma variável explicativa e Y_i é uma variável dependente.

i	1	2	3	4	5
X_i	0	1	2	3	4
Y_i	0,5	2,0	2,5	5,0	3,5

Considere o modelo de regressão linear simples na forma $Y_i = bX_i + \epsilon_i$, no qual ϵ representa um erro aleatório normal com média zero e variância σ^2 e b é o coeficiente do modelo.

Com base nos dados da tabela e nas informações apresentadas, é correto afirmar que o valor da estimativa de mínimos quadrados ordinários do coeficiente b é igual a

a) 0,75.

b) 0,9.

c) 1,2.

d) 1,35.

e) 1,45.

Comentários:

Como o modelo proposto passa pela origem, o estimador de mínimos quadrados para o coeficiente angular é dado por:

$$\hat{b} = \frac{\sum xy}{\sum x^2}.$$

Com base nos valores tabelados, temos:

i	1	2	3	4	5
X_i	0	1	2	3	4
Y_i	0,5	2,0	2,5	5,0	3,5
X_iY_i	0	2	5	15	14
${X_i}^2$	0	1	4	9	16

Assim, a estimativa do coeficiente b pelo método dos mínimos quadrados é igual a:

$$\hat{b} = \frac{\sum xy}{\sum x^2}.$$

$$\hat{b} = \frac{0 + 2 + 5 + 15 + 14}{0 + 1 + 4 + 9 + 16}$$

$$\hat{b} = \frac{36}{30}$$

$$\hat{b} = 1{,}2.$$

Gabarito: C.

10. (CESPE/BANESE/2021)

	X	Y
Média	**5**	**10**
Desvio Padrão	**2**	**2**

Com base nas informações apresentadas na tabela precedente e considerando que a covariância entre as variáveis X e Y seja igual a 3, julgue o item que se segue.

O coeficiente de determinação (ou de explicação) da reta de regressão linear da variável X em função da variável Y é igual ou superior a 0,60.

Comentários:

O coeficiente de correlação pode ser expresso por

$$r = \frac{Cov(X, Y)}{\sigma_x \sigma_y}$$

Conforme a tabela apresentada no enunciado, temos que $Cov(X,Y) = 3, \sigma_x = \sigma_y = 2$. Portanto,

$$r = \frac{3}{2 \times 2} = \frac{3}{4}$$

Como sabemos, o coeficiente de determinação é o quadrado desse valor:

$$\left(\frac{3}{4}\right)^2 = \left(\frac{9}{16}\right) = 0{,}5625 < 0{,}60$$

Gabarito: Errado.

11. (CESPE/Pref. Aracaju/2021) Um modelo de regressão linear simples na forma $y = ax + b + \epsilon$, no qual ϵ representa o erro aleatório com média nula e variância constante, foi ajustado para um conjunto de dados no qual as médias aritméticas das variáveis y e x são, respectivamente, $\overline{y} = 10$ e $\overline{x} = 5$. Pelo método dos mínimos quadrados ordinários, se a estimativa do intercepto (coeficiente b) for igual a 20, então a estimativa do coeficiente angular a proporcionada por esse mesmo método deverá ser igual a

a) −2.

b) 2.

c) −1.

d) 0.

e) 1.

Comentários:

Segundo o método dos mínimos quadrados, a reta de regressão sempre passa pelo ponto formado pela média das variáveis dependente e independente. Ou seja, se

$$\hat{y} = \hat{a}x + \hat{b}$$

é a reta de regressão, então o ponto $(\bar{x}, \bar{y})$ pertence a ela.

Segundo o método dos mínimos quadrados, a reta de regressão sempre passa pelo ponto formado pela média das variáveis dependente e independente. Ou seja, o ponto $(\bar{x}, \bar{y})$ pertence à reta de regressão.

Como $\hat{b} = 20, \bar{y} = 10$ e $\bar{x} = 5$, então

$$\hat{y} = \hat{a}x + \hat{b}$$

$$10 = \hat{a} \times 5 + 20$$

$$5\hat{a} = -10$$

$$\hat{a} = -2.$$

Gabarito: A.

12. (CESPE/BANESE/2021) Considere que uma tendência linear na forma $\hat{y} = 4x + 2$ tenha sido obtida com base no método dos mínimos quadrados ordinários. Acerca dessa tendência, sabe-se ainda que o desvio padrão da variável y foi igual a 8; que o desvio padrão da variável x foi igual a 1; e que a média aritmética da variável x foi igual a 2. Com base nessas informações, julgue o item subsequente, relativo a essa tendência linear.

A média aritmética da variável y foi igual a 8.

Comentários:

Segundo o método dos mínimos quadrados, a reta de regressão sempre passa pelo ponto formado pela média das variáveis dependente e independente. Ou seja, o ponto $(\bar{x}, \bar{y})$ pertence à reta de regressão.

Assim, a média de y é encontrada ao resolvermos a equação para $x = \bar{x} = 2$:

$$\bar{y} = 4 \times 2 + 2 = 10$$

Gabarito: Errado.

13. (CESPE/BANESE/2021) Considere que uma tendência linear na forma $\hat{y} = 4x + 2$ tenha sido obtida com base no método dos mínimos quadrados ordinários. Acerca dessa tendência, sabe-se ainda que o desvio padrão da variável y foi igual a 8; que o desvio padrão da variável x foi igual a 1; e que a média aritmética da variável x foi igual a 2. Com base nessas informações, julgue o item subsequente, relativo a essa tendência linear.

A covariância entre as variáveis x e y foi superior a 2.

Comentários:

A estimativa do coeficiente angular da reta de regressão é definida como a razão entre a covariância e a variância da variável explicativa:

$$\beta = \frac{Cov(X, Y)}{Var(X)}$$

Como o modelo de regressão assume a forma $\hat{y} = 4x + 2$, temos que $\beta = 4$.

Agora, como o desvio padrão da variável x vale 1, temos que

$$Var(X) = 1.$$

Logo,

$$Cov(X, Y) = 4 \times 1 = 4 > 2$$

Portanto, a covariância entre as variáveis x e y foi superior a 2

Gabarito: Certo.

14. (CESPE/PF/2021) Um estudo objetivou avaliar a evolução do número mensal Y de milhares de ocorrências de certo tipo de crime em determinado ano. Com base no método dos mínimos quadrados ordinários, esse estudo apresentou um modelo de regressão linear simples da forma

$$\overline{Y} = 5 - 0,1 \times T,$$

em que $\overline{Y}$ representa a reta ajustada em função da variável regressora T, tal que $1 \leq T \leq 12$.

Os erros padrão das estimativas dos coeficientes desse modelo, as razões t e seus respectivos p-valores encontram-se na tabela a seguir.

	Erro Padrão	Razão t	p-valor
Intercepto	0,584	8,547	0,00
Coeficiente Angular	0,064	1,563	0,15

Os desvios padrão amostrais das variáveis y e t foram, respectivamente, 1 e 3,6.

Com base nessas informações, julgue o item a seguir.

Se a média amostral da variável t for igual a 6,5, então a média amostral da variável Y será igual a 4,35 mil ocorrências.

Comentários:

Segundo o método dos mínimos quadrados, a reta de regressão sempre passa pelo ponto formado pela média das variáveis dependente e independente. Ou seja, o ponto $(\bar{x}, \bar{y})$ pertence à reta de regressão.

Assim, se o modelo de regressão é $\hat{Y} = 5 - 0,1 \times T$, podemos afirmar que:

$$\bar{Y} = 5 - 0,1 \times \bar{T}$$

Se a média amostral da variável T for igual a 6,5, teremos:

$$\bar{Y} = 5 - 0,1 \times 6,5 = 5 - 0,65 = 4,35$$

Portanto, quando $\bar{T} = 6,5$ mil, a média amostral da variável Y será igual a 4,35 mil ocorrências.

Gabarito: Certo.

15. (CESPE/MJ-SP/2021) A tabela de análise de variância a seguir se refere a um modelo de regressão linear simples na forma $y = ax + b + \epsilon$, na qual $\epsilon \sim N(0, \sigma^2)$. Os resultados da tabela foram obtidos com base em uma amostra aleatória simples n de pares de observações independentes (x,y).

Fonte de Variação	Graus de Liberdade	Soma de Quadrados

Regressão	1	82
Resíduos	8	8
Total	9	90

Com base nessas informações, julgue o item subsequente.

Se as médias amostrais das variáveis x e y forem iguais a zero, então o estimador de mínimos quadrados ordinários de b será igual a zero.

Comentários:

Segundo o método dos mínimos quadrados, a reta de regressão sempre passa pelo ponto formado pela média das variáveis dependente e independente. Ou seja, o ponto $(\bar{x}, \bar{y})$ pertence à reta de regressão. Considerando o modelo de regressão linear simples $y = a + bx + \epsilon$, teremos:

$$\bar{y} = a + b\bar{x}$$

Caso as médias amostrais das variáveis x e y sejam iguais a zero, vamos ter:

$$\hat{b} = 0 - \hat{a} \times 0$$

$$\hat{b} = 0$$

Gabarito: Certo.

16. (CESPE/BANESE/2021)

	X	Y
Média	**5**	**10**
Desvio Padrão	**2**	**2**

Com base nas informações apresentadas na tabela precedente e considerando que a covariância entre as variáveis X e Y seja igual a 3, julgue o item que se segue.

A reta de regressão linear da variável Y em função da variável X, obtida pelo método de mínimos quadrados ordinários, pode ser escrita como $Y = 0{,}75X + 6{,}25$.

Comentários:

Os coeficientes $\hat{\alpha}$ e $\hat{\beta}$ da reta de regressão $\hat{Y} = \beta X + \alpha$ são obtidos por meio das seguintes relações:

$$\hat{\beta} = \frac{Cov(X,Y)}{Var(X)}$$

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

De acordo com o enunciado, temos que $Cov(X,Y) = 3, Var(X) = 2^2 = 4, \bar{X} = 5$ e $\bar{Y} = 10$.

Aplicando esses valores nas fórmulas apresentadas anteriormente, temos:

$$\hat{\beta} = \frac{3}{4} = 0{,}75$$

$$\hat{\alpha} = 10 - 0{,}75 \times 5 = 10 - 3{,}75 = 6{,}25$$

A reta de regressão da variável Y em função da variável X pode ser escrita como $\hat{Y} = 0{,}75X + 6{,}25$.

Gabarito: Certo.

17. (CESPE/PG DF/2021) O coeficiente de correlação linear de Pearson entre duas variáveis aleatórias discretas X e Y definidas sobre um mesmo espaço amostral é dado por

$$\boldsymbol{CORR(X,Y) = \frac{n(\sum_{i=1}^{n} x_i y_i) - (\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{\sqrt{n(\sum_{i=1}^{n} x_i^2) - (\sum_{i=1}^{n} x_i)^2}\sqrt{n(\sum_{i=1}^{n} y_i^2) - (\sum_{i=1}^{n} y_i)^2}}}$$

Já na reta de melhor ajuste $Y = aX + b$, determinada pelo método dos mínimos quadrados, os coeficientes são dados por

$$\boldsymbol{a = \frac{n(\sum_{i=1}^{n} x_i y_i) - (\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n(\sum_{i=1}^{n} x_i^2) - (\sum_{i=1}^{n} x_i)^2}}$$

$$\boldsymbol{b = \frac{\sum_{i=1}^{n} y_i - a\sum_{i=1}^{n} x_i}{n}}$$

Uma forma de avaliar a precisão do modelo consiste em comparar o estimador não viesado da variância residual, obtidos das diferenças entre os valores observados e os previstos pelo modelo, $\widehat{S}_e = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$, com o estimador não viesado da variância dos valores observados, $S_e = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$.

Tal avaliação também pode ser realizada pela aferição na redução da soma dos quadrados dos resíduos na passagem do modelo simples, em que as observações são aproximadas por sua média, para o modelo de regressão linear, redução esta que é dada por $\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 = a^2\sum_{i=1}^{n}(x_i - \bar{x})^2$.

Com base nessas informações, julgue o item seguinte.

Se, para certo conjunto de dados, o coeficiente angular da reta de melhor ajuste obtida pelo método dos mínimos quadrados for nulo, então o coeficiente de correlação de Pearson entre essas variáveis também será nulo.

Comentários:

Como os numeradores do coeficiente angular da reta e do coeficiente de correlação de Pearson são iguais, quando um for nulo, o outro necessariamente também será nulo. Portanto, a assertiva está correta.

Gabarito: Certo.

18. (CESPE/PG DF/2021) O coeficiente de correlação linear de Pearson entre duas variáveis aleatórias discretas X e Y definidas sobre um mesmo espaço amostral é dado por

$$CORR(X,Y) = \frac{n(\sum_{i=1}^{n} x_i y_i) - (\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{\sqrt{n(\sum_{i=1}^{n} x_i^2) - (\sum_{i=1}^{n} x_i)^2}\sqrt{n(\sum_{i=1}^{n} y_i^2) - (\sum_{i=1}^{n} y_i)^2}}$$

Já na reta de melhor ajuste $Y = aX + b$, determinada pelo método dos mínimos quadrados, os coeficientes são dados por

$$a = \frac{n(\sum_{i=1}^{n} x_i y_i) - (\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n(\sum_{i=1}^{n} x_i^2) - (\sum_{i=1}^{n} x_i)^2}$$

$$b = \frac{\sum_{i=1}^{n} y_i - a\sum_{i=1}^{n} x_i}{n}$$

Uma forma de avaliar a precisão do modelo consiste em comparar o estimador não viesado da variância residual, obtidos das diferenças entre os valores observados e os previstos pelo modelo, $\widehat{S}_e = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$, com o estimador não viesado da variância dos valores observados, $S_e = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2$.

Tal avaliação também pode ser realizada pela aferição na redução da soma dos quadrados dos resíduos na passagem do modelo simples, em que as observações são aproximadas por sua média, para o modelo de regressão linear, redução esta que é dada por $\sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2 = a^2\sum_{i=1}^{n}(x_i - \overline{x})^2$.

Com base nessas informações, julgue o item seguinte.

Quanto mais próximo de -1 estiver o coeficiente de correlação de Pearson entre duas variáveis, menos indicada será a aplicação do método de mínimos quadrados para obter a relação entre as variáveis.

Comentários:

Um coeficiente de correlação de Pearson próximo -1 indica uma forte correlação linear entre as variáveis. Desse modo, o método dos mínimos quadrados conseguirá representar a relação entre as variáveis por meio de uma reta.

Gabarito: Errado.

19. (FCC/ALAP/2020) Em uma empresa de determinado ramo de atividade, utilizando o método de regressão linear, obteve-se a equação de tendência (T) da série temporal abaixo.

Os dados apresentam 10 observações da série temporal Y, que representa o faturamento de uma empresa, em milhões de reais. Supõe-se que essa série é composta apenas de uma tendência T e um ruído branco de média zero e variância constante.

t	1	2	3	4	5	6	7	8	9	10
y_t	6	5	6	8	8	7	8	10	10	11

Observação: t representa o ano e yt o faturamento da empresa no ano t, em milhões de reais.

Dados:

$$\sum_{t=1}^{10} t = 55\,,\ \sum_{t=1}^{10} t^2 = 385\,,\ \sum_{t=1}^{10} y_t = 79\,,\ \sum_{t=1}^{10} t \times y_t = 484$$

A tendência apresenta a forma $T = a + bt$, em que a e b foram obtidos usando o método dos mínimos quadrados. Considerando a equação obtida, tem-se que o acréscimo no faturamento do ano t, com t > 1, para o ano (t + 1) é, em milhões de reais, de

a) 1,2.

b) 1,5.

c) 0,6.

d) 2,4.

e) 1,8.

Comentários:

Precisamos descobrir o valor do coeficiente angular, o enunciado pede a variação entre o ano t e o ano seguinte $(t + 1)$:

$$b = \frac{\sum_{i=1}^{n}(t_i y_i) - n \times \bar{t} \times \bar{y}}{\sum_{i=1}^{n}(t_i^2) - n \times t^2}$$

$$b = \frac{484 - 10 \times 5{,}5 \times 7{,}9}{385 - 10 \times 5{,}5^2}$$

$$b = \frac{484 - 434{,}5}{385 - 302{,}5}$$

$$b = \frac{49{,}5}{82{,}5}$$

$$b = 0{,}6$$

Gabarito: C.

20. (CESPE/TJ-AM/2019) 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:

- $E(e) = 0$, isto é, em média, o erro do modelo deve ser 0;
- $Var(e) = \sigma^2$, a variância deve ser constante, isto é, deve existir homocedasticia;
- $Cov(ei, ej) = 0$, os erros devem ser independentes, ou seja, não há correlação entre os erros.

Nessa questão, o único ponto que precisamos mostrar é que o $Var(e) = 1$. 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.

21. (CESPE/TJ-AM/2019) No modelo de regressão linear simples na forma matricial $Y = X\beta + \epsilon$, Y denota o vetor de respostas, X representa a matriz de delineamento (ou matriz de desenho), β é o vetor de coeficientes do modelo e ϵ é o vetor de erros aleatórios independentes e identicamente distribuídos. Tem-se também que $X'Y = \begin{pmatrix} 2 \\ 10 \end{pmatrix}$ e $(X'X)^{-1} = \begin{pmatrix} 1 & 0,5 \\ 0,5 & 1 \end{pmatrix}$ em que X' é a matriz transposta de X.

Com base nessas informações, julgue o próximo item, considerando que a variância do erro aleatório é $\sigma_\epsilon^2 = 4$

O referido modelo possui uma única variável regressora.

Comentários:

A questão trata de um modelo de regressão linear **simples**, ou seja, um modelo formado por uma única variável regressora e uma variável resposta. A variável regressora também recebe o nome de variável independente, enquanto a variável resposta é a variável dependente. Sendo Y função linear de X, o modelo de regressão linear simples é dado por:

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

Gabarito: Certo.

22. (CESPE/TJ-AM/2019) 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 desvio padrão de x é superior a 1.

Comentários:

Em outra questão dessa mesma prova, determinamos que:

$$R \cong 0,92$$

$$R^2 = 0,85$$

O enunciado nos disse que:

$$\hat{\beta} = 0,8$$

Para o estimador do coeficiente angular temos:

$$\hat{\beta} = \frac{S_{xy}}{S_{xx}}$$

Então,

$$0,8 = \frac{S_{xy}}{S_{xx}}$$

Reorganizando os termos, verificamos que:

$$S_{xy} = 0,8 \times S_{xx}$$

Lembrando que:

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

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

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

O coeficiente de determinação é dado por:

$$R^2 = \frac{S_{xy}^2}{S_{xx} \times S_{yy}}$$

Substituindo, temos:

$$0{,}85 = \frac{(0{,}8 \times S_{xx})^2}{S_{xx} \times S_{yy}}$$

$$0{,}85 = \frac{0{,}64 \times S_{xx}^2}{S_{xx} \times S_{yy}}$$

$$0{,}85 \times S_{yy} = 0{,}64 \times S_{xx}$$

Sabendo que y possui distribuição normal padrão, $S_{yy} = 1$.

$$0{,}85 = 0{,}64 S_{xx}$$

$$S_{xx} = \frac{0{,}85}{0{,}64}$$

$$S_{xx} \cong 1{,}32$$

Gabarito: Certo.

23. (CESPE/TJ-AM/2019) 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 intercepto do referido modelo é igual ou superior a 0,8.

Comentários:

O modelo de regressão linear tem a forma $y = 0{,}8x + b + \epsilon$. Sabemos que a média dos erros é igual a zero. Além disso, como y possui distribuição normal padrão, sabemos que a sua média é 0. Podemos, então, colocar $y = 0$ na regressão.

Assim, temos:

$$0 = 0{,}8\,\bar{x} + b.$$

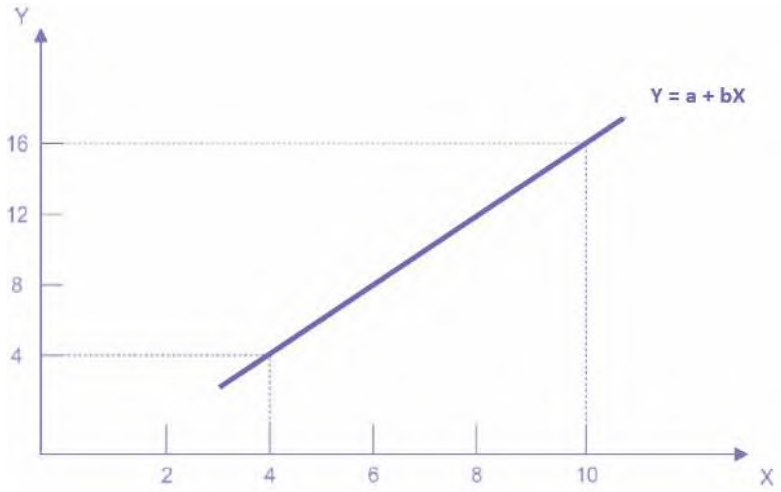
Isolando b:

$$b = -0{,}8\,\bar{x}.$$

Como nada foi dito acerca do valor de $\bar{x}$, nada podemos afirmar sobre o intercepto b. Portanto, a assertiva está incorreta.

Gabarito: Errado.

24. (FCC/SEFAZ-BA/2019) Em uma determinada indústria, foi efetuada uma pesquisa a respeito da possível relação entre o número de horas trabalhadas (X), com $X \geq 2$, e as quantidades produzidas de um produto (Y). Com base em 10 pares de observações (Xi, Yi) e considerando o gráfico de dispersão correspondente, optou-se por utilizar o modelo linear $Y_i = \alpha + \beta X_i + \varepsilon_i$, com i representando a i-ésima observação, ou seja, $i = 1, 2, 3, \ldots 10$. Os parâmetros α e β são desconhecidos e as suas estimativas (a e b, respectivamente) foram obtidas pelo método dos mínimos quadrados. Observação: ε_i é o erro aleatório com as respectivas hipóteses do modelo de regressão linear simples. Considere o gráfico, abaixo, construído utilizando os valores encontrados para as estimativas de α e β.



Dados:

$$\sum_{i=1}^{10} X_i = 120$$

A previsão da quantidade produzida será igual ao dobro da média verificada das 10 observações Y_i quando o número de horas trabalhadas for igual a

a) 20.

b) 24.

c) 22.

d) 18.

e) 12.

Comentários:

Iniciaremos calculando o coeficiente angular da reta:

$$Y = a + bx$$

$$a = Y - bX$$

$$a = 4 - 2 \times 4$$

$$a = 4 - 8$$

$$a = -4$$

Agora, calcularemos a média:

$$\bar{X} = \frac{120}{10} = 12$$

Calculando o coeficiente linear

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

$$-4 = \bar{Y} - 2 \times 12$$

$$\bar{Y} = 20$$

Calculando a previsão para o dobro:

$$Y = a + bx$$

$$40 = -4 + 2x$$

$$x = 22$$

Gabarito: C.

25. (FCC/BANRISUL/2019) Utilizando o método dos mínimos quadrados, obteve-se a equação de tendência $\hat{T}_t = 15 + 2,5t$, sendo $t = 1, 2, 3, \ldots$, com base nos lucros anuais de uma empresa, em milhões de reais, nos últimos 10 anos, em que $t = 1$ representa 2009, $t = 2$ representa 2010 e assim por diante. Por meio dessa equação, obtém-se que a previsão do lucro anual dessa empresa, no valor de 55 milhões de reais, será para o ano

a) 2021.

b) 2025.

c) 2024.

d) 2023.

e) 2022.

Comentários:

O enunciado já nos deu a equação, então, basta substituirmos $\hat{T}_t$ por 55. Assim:

$$\hat{T}_t = 15 + 2,5t$$

$$55 = 15 + 2,5t$$

$$2,5t = 40$$

$$t = \frac{40}{2{,}5}$$

$$t = 16$$

Temos que $t = 0$ representa o ano de 2008. Assim, partiremos do ano de 2008 para obter a previsão de lucro anual:

$$2008 + 16 = 2024$$

Gabarito: C.

26. (VUNESP/Pref. Mogi das Cruzes/2019) Sejam S o valor do salário, em R$ 1.000,00, e t o respectivo tempo de serviço, em anos, de 20 empregados de uma empresa. Optou-se, com o objetivo de previsão do salário de um determinado empregado em função do seu tempo de serviço, por utilizar a relação linear $S_i = \alpha + \beta t_i + \varepsilon_i$, com i representando a i-ésima observação, α e β são parâmetros desconhecidos e εi é o erro aleatório com as respectivas hipóteses da regressão linear simples. Utilizando o método dos mínimos quadrados, com base nas 20 observações correspondentes dos 20 empregados, obtiveram-se as estimativas de α e β (a e b, respectivamente). O valor encontrado para b foi de 1,8 e as médias dos salários dos 20 empregados e dos correspondentes tempos de serviço apresentam os valores de R$ 2.800,00 e 2 anos, respectivamente.

A previsão de salário para um empregado que tenha 5 anos de serviço é de

a) R$ 6.800,00

b) R$ 7.500,00

c) R$ 8.200,00

d) R$ 8.400,00

e) R$ 9.000,00

Comentários:

Tomemos a equação $\hat{S} = a + bt$ para a reta estimada. Sendo b é igual a 1,8.

A média dos salários dos 20 empregados é 2800 e o tempo médio de serviço é 2 anos.

Logo,

$$2{,}8 = a + 1{,}8 \times 2$$

$$2{,}8 = a + 3{,}6$$

$$a = -0{,}8$$

A previsão para 5 anos fica:

$$\hat{S} = (-0{,}8) + 1{,}8 \times 5$$

$$\hat{S} = (-0{,}8) + 9$$

$$\hat{S} = 8{,}2$$

Que corresponde a R$ 8.200,00.

Gabarito: C.

27. (VUNESP/MPE-SP/2019). 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.

28. (CESPE/STM/2018). Considerando que $\hat{Y}$ seja uma variável resposta ajustada por um modelo de regressão em função de uma variável explicativa X, que $x_1, \ldots, x_n$ representem as réplicas de X e que $\hat{\alpha}$ e $\hat{\beta}$ sejam as estimativas dos parâmetros do modelo, julgue o item a seguir.

Em um modelo linear $\hat{Y} = \hat{\alpha} + \hat{\beta}X$, a hipótese de homoscedastiscidade significa que a variância dos erros deve ser constante, e o valor esperado dos erros deve ser zero.

Comentários:

A hipótese de homoscedasticidade diz apenas que a variância dos erros deve ser constante, mas não que o valor esperado dos erros deve ser zero. De fato, o valor esperado dos erros deve ser zero no modelo de regressão linear, porém, isso não representa homoscedasticidade. Portanto, a questão erra ao incluir o valor esperado dos erros nesse conceito.

Gabarito: Errado.

29. (CESPE/ABIN/2018) 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):

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

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.

30. (CESPE/EBSERH/2018) Deseja-se estimar o total de carboidratos existentes em um lote de 500.000 g de macarrão integral. Para esse fim, foi retirada uma amostra aleatória simples constituída por 5 pequenas porções desse lote, conforme a tabela seguinte, que mostra a quantidade x amostrada, em gramas, e a quantidade de carboidratos encontrada, y, em gramas.

Amostra	*X*	*Y*
1	100	60
2	80	40
3	90	40
4	120	50
5	110	60

Com base nas informações e na tabela apresentadas, julgue o item a seguir.

Considerando-se o modelo de regressão linear na forma $y = ax + \varepsilon$, em que ε denota o erro aleatório com média nula e variância V, e a representa o coeficiente angular, a estimativa de mínimos quadrados ordinários do coeficiente a é igual ou superior a 0,5.

Comentários:

Essa questão nos apresenta um modelo de regressão linear que não apresenta coeficiente linear. Portanto, estamos diante de um modelo que obrigatoriamente passa pela origem. Nessa situação, o coeficiente angular é dado por:

$$\hat{\alpha} = \frac{\sum X_i Y_i}{\sum X_i^2}$$

Agora, vamos acrescentar algumas informações à tabela dada:

Amostra	X	Y	$X \times Y$	X^2
1	100	60	6.000	10.000
2	80	40	3.200	6.400
3	90	40	3.600	8.100
4	120	50	6.000	14.400
5	110	60	6.600	12.100
Total	$\bar{X} = 100$	$\bar{Y} = 50$	25.400	51.000

Aplicando os valores da tabela na fórmula anterior, teremos:

$$\hat{\alpha} = \frac{25400}{51000}$$

$$\hat{\alpha} = \frac{254}{510}$$

$$\hat{\alpha} \cong 0{,}498$$

Como 50% de 51.000 é 25.500, nem precisávamos efetuar a divisão para concluirmos que $\hat{\alpha} < 0{,}5$.

Gabarito: Errado.

31. (CESPE/PF/2018) O intervalo de tempo entre a morte de uma vítima até que ela seja encontrada (y em horas) denomina-se intervalo post mortem. Um grupo de pesquisadores mostrou que esse tempo se relaciona com a concentração molar de potássio encontrada na vítima (x, em mmol/dm3). Esses pesquisadores consideraram um modelo de regressão linear

simples na forma $y = ax + b + \varepsilon$, em que a representa o coeficiente angular, b denomina-se intercepto, e ε denota um erro aleatório que segue distribuição normal com média zero e desvio padrão igual a 4.

As estimativas dos coeficientes a e b, obtidas pelo método dos mínimos quadrados ordinários foram, respectivamente, iguais a 2,5 e 10. O tamanho da amostra para a obtenção desses resultados foi n = 101. A média amostral e o desvio padrão amostral da variável x foram, respectivamente, iguais a 9 mmol/dm3 e 1,6 mmol/dm3 e o desvio padrão da variável y foi igual a 5 horas.

A respeito dessa situação hipotética, julgue o item a seguir.

A média amostral da variável resposta y foi superior a 30 horas.

Comentários:

Segundo o enunciado, o modelo de regressão linear é dado por:

$$y = ax + b + \varepsilon$$

As estimativas de a e b são:

$$\hat{a} = 2{,}5$$

$$\hat{b} = 10$$

Assim, a reta de regressão é determinada pela equação:

$$\hat{y} = 2{,}5x + 10$$

Substituindo os valores, temos:

$$\bar{y} = 2{,}5\bar{x} + 10$$

$$\bar{y} = 2{,}5 \times 9 + 10$$

$$\bar{y} = 32{,}5$$

Gabarito: Certo.

32. (CESPE/PF/2018) O intervalo de tempo entre a morte de uma vítima até que ela seja encontrada (y em horas) denomina-se intervalo post mortem. Um grupo de pesquisadores mostrou que esse tempo se relaciona com a concentração molar de potássio encontrada na vítima (x, em mmol/dm3). Esses pesquisadores consideraram um modelo de regressão linear simples na forma $y = ax + b + \varepsilon$, em que a representa o coeficiente angular, b denomina-se intercepto, e ε denota um erro aleatório que segue distribuição normal com média zero e desvio padrão igual a 4.

As estimativas dos coeficientes a e b, obtidas pelo método dos mínimos quadrados ordinários foram, respectivamente, iguais a 2,5 e 10. O tamanho da amostra para a obtenção desses resultados foi n = 101. A média amostral e o desvio padrão amostral da variável x foram,

respectivamente, iguais a 9 mmol/dm3 e 1,6 mmol/dm3 e o desvio padrão da variável y foi igual a 5 horas.

A respeito dessa situação hipotética, julgue o item a seguir.

De acordo com o modelo ajustado, caso a concentração molar de potássio encontrada em uma vítima seja igual a 2 mmol/dm3, o valor predito correspondente do intervalo post mortem será igual a 15 horas.

Comentários:

Segundo o enunciado, o modelo de regressão linear é dado por:

$$y = ax + b + \varepsilon$$

As estimativas de a e b são:

$$\hat{a} = 2{,}5$$

$$\hat{b} = 10$$

Assim, a reta de regressão é determinada pela equação:

$$\hat{y} = 2{,}5x + 10$$

Agora, substituindo x por 2, temos:

$$\hat{y} = 2{,}5 \times 2 + 10$$

$$\hat{y} = 15$$

Gabarito: Certo.

33. (CESPE/STM/2018). Em um modelo de regressão linear simples na forma $yi = a + bx_i + \varepsilon_i$, em que a e b são constantes reais não nulas, y_i representa a resposta da i-ésima observação a um estímulo x_i e ε_i éo erro aleatório correspondente, para $i = 1,\ldots,n$, considere que $\sum_i (x_i - \overline{x})^2 = 10$, em que $\overline{x} = (x_1 + \ldots + x_n)/n$, e que o desvio padrão de cada ε_i seja igual a 10, para $i = 1,\ldots,n$.

A respeito dessa situação hipotética, julgue o item que se segue.

Se $\hat{b}$ representar o estimador de mínimos quadrados ordinários do coeficiente b, então $var[\hat{b}] = 10$.

Comentários:

Pelo método dos mínimos quadrados, a variância do estimador $\hat{b}$ é dada por:

$$var(\hat{b}) = \frac{\sigma^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

Assim, basta substituirmos pelos valores dados no enunciado:

$$var(\hat{b}) = \frac{100}{10}$$

$$var(\hat{b}) = 10$$

Gabarito: Certo.

34. (FCC/Pref. São Luís/2018) 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.

35. (FCC/SEF-SC/2018) A tabela a seguir indica o valor y do salário, em número de salários mínimos (SM) e os respectivos tempos de serviço, em anos, x, de 5 funcionários de uma empresa:

x (anos)	2	3	5	3	2
y (SM)	3	4	7	4	2

Suponha que valha a relação: $y_i = \alpha + \beta x_i + \varepsilon_i$, em que i representa a i-ésima observação, α e β são parâmetros desconhecidos e ε_i é o erro aleatório com as hipóteses para a regressão linear simples. Se as estimativas de α e β forem obtidas pelo método de mínimos quadrados por meio dessas 5 observações, a previsão de salário para um funcionário com 4 anos de serviço será, em SM, igual a

a) 6,1

b) 5,2

c) 6,0

d) 5,5

e) 5,8

Comentários:

Iniciaremos montando uma tabela com os dados para aplicarmos as fórmulas:

X	Y	$(X - \bar{X})$	$(Y - \bar{Y})$	$(X - \bar{X})(Y - \bar{Y})$	$(X - \bar{X})^2$
2	3	-1	-1	1	1
3	4	0	0	0	0
5	7	2	3	6	4

3	4	0	0	0	0
2	2	-1	-2	2	1
$\bar{X} = 3$	$\bar{Y} = 4$	Total		9	6

Calculando b, temos:

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

$$b = \frac{9}{6}$$

$$b = 1{,}5$$

Conhecendo b, podemos descobrir o valor de a:

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

$$a = 4 - 1{,}5 \times 3$$

$$a = -0{,}5$$

Aplicando os valores à reta calculada, temos:

$$\hat{Y} = -0{,}5 + 1{,}5 \times X$$

Substituindo e tomando $x = 4$, para um funcionário com 4 anos de serviço, fica:

$$Y = -0{,}5 + 1{,}5 \times 4$$

$$Y = 5{,}5$$

Gabarito: D.

36. (FGV/ALERO/2018) Se b_0 e b_1 são as estimativas por mínimos quadrados de β_0 e β_1, respectivamente, então seus valores são dados por

a) $b_1 = \sum_{i=1}^{n}(Y_i - \bar{Y})^2 / \sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})\,; b_o = \bar{Y} - b_1\bar{X}$

b) $b_1 = \sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}) / \sum_{i=1}^{n}(Y_i - \bar{Y})^2\,; b_o = \bar{Y} + b_1\bar{X}$

c) $b_1 = \sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}) / \sum_{i=1}^{n}(Y_i - \bar{Y})^2\,; b_o = \bar{Y} - b_1\bar{X}$

d) $b_1 = \sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}) / \sum_{i=1}^{n}(X_i - \bar{X})^2\,; b_o = \bar{Y} - b_1\bar{X}$

e) $b_1 = \sum_{i=1}^{n}(X_i - \bar{X})^2 / \sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})\,; b_o = \bar{Y} + b_1\bar{X}$

Comentários:

Dado o modelo de regressão linear apresentado a seguir:

$$Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + \varepsilon_i.$$

Temos que os estimadores de mínimos quadrados (EMQ) são calculados por meio das seguintes expressões:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i Y_i) - n \times \bar{X} \times \bar{Y}}{\sum_{i=1}^{n}(X_i^2) - n \times \bar{X}^2} = \frac{\sum_{i=1}^{n}[(X_i - \bar{X}) \times (Y_i - \bar{Y})]}{\sum_{i=1}^{n}[(X_i - \bar{X})^2]}$$

$$\hat{\beta}_0 = \frac{\sum_{i=1}^{n} Y_i}{n} - \hat{\beta}_1 \frac{\sum_{i=1}^{n} x_i}{n} = \bar{Y} - \hat{\beta}_1 \bar{X}$$

Gabarito: D.

37. (FGV/ALERO/2018) Se $\widehat{Y} = b_0 + b_1 X$ é a reta ajustada pela regressão e se $e_i = Y_i - \widehat{Y}$ é o resíduo da observação $i, i - 1, \cdots, n$, avalie as afirmativas a seguir.

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

II. $\sum_{i=1}^{n} Y_i = \sum_{i=1}^{n} \widehat{Y}$.

III. O ponto $(\bar{X}, \bar{Y})$ pertence à reta ajustada.

Está correto o que se afirma em

a) I, apenas.

b) I e II, apenas.

c) I e III, apenas.

d) II e III, apenas.

e) I, II e III.

Comentários:

O próprio enunciado nos disse que os resíduos correspondem às diferenças entre os valores observados, Y_i, e os correspondentes valores ajustados, $\hat{Y}_i$:

$$e_i = Y_i - \hat{Y}_i = Y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$$

Essa medida é importante porque nos permite verificar o ajuste do modelo.

A principais propriedades do método dos mínimos quadrados são:

i) a soma dos resíduos é sempre nula:

$$\sum_{i=1}^{n} e_i = \sum_{i=1}^{n} (Y_i - \hat{Y}_i) = 0$$

Para confirmar isso, basta substituirmos os estimadores de mínimos quadrados

$$\sum_{i=1}^{n} e_i = \sum_{i=1}^{n} Y_i - n\hat{\beta}_0 - \hat{\beta}_1 \sum_{i=1}^{n} x_i$$

Da terceira propriedade, temos que:

$$\bar{Y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x}$$

Substituindo na expressão anterior, temos:

$$\sum_{i=1}^{n} e_i = \sum_{i=1}^{n} Y_i - n\left(\bar{Y} - \hat{\beta}_1 \bar{x}\right) - \hat{\beta}_1 \sum_{i=1}^{n} x_i$$

$$\sum_{i=1}^{n} e_i = \sum_{i=1}^{n} Y_i - n\bar{Y} + n\hat{\beta}_1 \bar{x} - \hat{\beta}_1 \sum_{i=1}^{n} x_i$$

$$\sum_{i=1}^{n} e_i = n\bar{Y} - n\bar{Y} + n\hat{\beta}_1 \bar{x} - n\hat{\beta}_1 \bar{x} = 0$$

ii) a soma dos valores observados Y_i é igual à soma dos valores ajustados $\widehat{Y_i}$.

$$\sum_{i=1}^{n} Y_i = \sum_{i=1}^{n} \widehat{Y_i}$$

Decorre exatamente do item anterior, pois:

$$\sum_{i=1}^{n} e_i = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right) = 0$$

$$\sum_{i=1}^{n} Y_i - \sum_{i=1}^{n} \hat{Y}_i = 0$$

$$\sum_{i=1}^{n} Y_i = \sum_{i=1}^{n} \hat{Y}$$

iii) A reta de regressão de mínimos quadrados passa pelo ponto $(\bar{X}, \bar{Y})$.

Realmente,

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i = \beta_0 + \beta_1 (x_i - \bar{x}) + \beta_1 \bar{x} + \varepsilon_i$$

$$= (\beta_0 + \beta_1 \bar{x}) + \beta_1 (x_i - \bar{x}) + \varepsilon_i = \beta_0^* + \beta_1 (x_i - \bar{x}) + \varepsilon_i,$$

com $\beta_0^* = \beta_0 + \beta_1 \bar{x}$. Assim, a reta de regressão ajustada é dada por

$$\hat{Y} = \hat{\beta}_0^* + \hat{\beta}_1 (x_i - \bar{x}) = \hat{\beta}_0 + \hat{\beta}_1 \bar{x} + \hat{\beta}_1 (x_i - \bar{x})$$

$$= \left(\bar{Y} - \hat{\beta}_1 \bar{x}\right) + \hat{\beta}_1 \bar{x} + \hat{\beta}_1 (x_i - \bar{x}) = \bar{Y} + \hat{\beta}_1 (x_i - \bar{x}).$$

Logo, no ponto $x_i = \bar{X}$, temos que:

$$\hat{Y} = \bar{Y} + \hat{\beta}_1(x_i - \bar{x}) = \bar{Y}.$$

Portanto, temos que a reta ajustada passa por $(\bar{X}, \bar{Y})$.

Dessa forma, todos os itens estão corretos.

Gabarito: E.

38. (CESPE/TCE-PE/2017) Um estudo de acompanhamento ambiental considerou, para $j = 1, 2, \ldots, 26$, um modelo de regressão linear simples na forma: $y_j = a + bx_j + e_j$, em que a e b são constantes reais, y_j representa a variável resposta referente ao j-ésimo elemento da amostra, x_j é a variável regressora correspondente, e e_j denota o erro aleatório que segue distribuição normal com média nula e variância V. Aplicando-se, nesse estudo, o método dos mínimos quadrados ordinários, obteve-se a reta ajustada $\hat{y}_j = 1 + 2x_j, para\ j = 1, 2, \ldots, 26$

Considerando que a estimativa da variância V seja igual a 6 e que o coeficiente de explicação do modelo (R quadrado) seja igual a 0,64, julgue o item.

Se $\bar{x} = \frac{\sum_{j=1}^{26} x_j}{26}$ representar a média amostral da variável regressora e se $\bar{y} = \frac{\sum_{j=1}^{26} y_j}{26}$ denotar a média amostral da variável resposta, com $\bar{x} > 0$ e $\bar{y} > 0$, então $\bar{x} < \bar{y}$.

Comentários:

A reta de regressão necessariamente passa pelo ponto $(\bar{x}, \bar{y})$. Aplicando esse ponto na reta de regressão, descobrimos que:

$$\bar{y} = 2\bar{x} + 1$$

Portanto, a média de y é formado pela multiplicação da média de x por 2, e pela soma desse produto com 1. Como todos os valores são positivos, podemos afirmar que $\bar{x} < \bar{y}$.

Gabarito: Certo.

39. (FGV/MPE-BA/2017) Com o objetivo de realizar uma projeção sobre a necessidade de novos servidores para o Ministério Público, foi elaborado um modelo de regressão associando o número de procedimentos em curso e a variável de interesse. A equação do modelo é:

$$NS_i = \alpha + \beta . PC_i + \varepsilon_i$$

onde NS é o número de novos servidores e PC a quantidade de procedimentos em curso.

Através de uma amostra representativa (n=20), em diversas unidades no MP, foram obtidas as seguintes estatísticas:

$$\sum NS^2 = 18000, \sum NS = 200, \sum PC = 800,$$

$$\sum(\boldsymbol{PC}).(\boldsymbol{NS}) = \mathbf{12000}\ \boldsymbol{e}\ \sum \boldsymbol{PC^2} = \mathbf{72000}$$

Com base no modelo e nas estatísticas, é correto afirmar que:

a) ainda que o número de procedimentos não sofra incrementos, novos servidores serão necessários a cada período;

b) se o número de procedimentos sofrer um incremento de 40 unidades, serão necessários mais oito novos servidores;

c) as estimativas de MQO são $\hat{\alpha} = 6$ e $\hat{\beta} = 0{,}1$;

d) a correlação entre o volume de procedimentos e o número de novos servidores é 0,7, comprovando a qualidade do modelo;

e) sendo estimativa de β positiva, o número de funcionários do MP deverá crescer a uma taxa de 10% ao período.

Comentários:

Em uma regressão linear na forma $Y_i = \hat{\alpha} + \hat{\beta} X_i + \varepsilon_i.$, os estimadores mínimos quadrados são dados pelas relações:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i Y_i) - n \times \bar{X} \times \bar{Y}}{\sum_{i=1}^{n}(X_i^2) - n \times \bar{X}^2}$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

em que $\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$ e $\bar{Y} = \frac{\sum_{i=1}^{n} Y_i}{n}$.

Utilizando as fórmulas acima, temos que:

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(PC_i NS_i) - n \times \overline{PC} \times \overline{NS}}{\sum_{i=1}^{n}(PC_i^2) - n \times \overline{PC}^2}$$

$$\hat{\beta} = \frac{12.000 - 20 \times 40 \times 10}{72.000 - 20 \times 40^2}$$

$$\hat{\beta} = \frac{4.000}{40.000} = 0{,}1$$

$$\hat{\alpha} = \overline{NS} - \hat{\beta}_1 \times \overline{PC}$$

$$\hat{\alpha} = 10 - 0{,}1 \times 40$$

$$\hat{\alpha} = 6$$

Dessa forma, o modelo de regressão será:

$$NS = 6 - 0{,}1 \times PC$$

Assim, podemos concluir que a alternativa correta é a Letra C.

Vejamos agora as demais alternativas:

Letra A: **Errada**. A demanda inicial não necessariamente será a mesma a cada período.

Letra B: **Errada**. Tomando $PC = 40$ no modelo de regressão, teremos:

$$NS = 6 + 0{,}1 \times 40 = 10.$$

Letra D: **Errada**. O coeficiente de correlação é dado pela relação:

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

Substituindo os valores fornecidos, temos:

$$r = \sqrt{\frac{0{,}1 \times [12.000 - 40 \times 200]}{72.000 - 20 \times 10^2}}$$

$$r = \sqrt{\frac{0{,}1 \times 4.000}{70.000}}$$

$$r = 0{,}07$$

Letra E: **Errada**. Cresce a uma taxa de 10% para cada procedimento novo.

Gabarito: C.

40. (CESPE/TCE-PA/2016). Uma regressão linear simples é expressa por $Y = a + b \times X + e$, em que o termo e corresponde ao erro aleatório da regressão e os parâmetros a e b são desconhecidos e devem ser estimados a partir de uma amostra disponível. Assumindo que a variável X é não correlacionada com o erro e, julgue o item subsecutivo, nos quais os resíduos das amostras consideradas são IID, com distribuição normal, média zero e variância constante.

Para uma amostra de tamanho $n = 25$, em que a covariância amostral para o par de variáveis X e Y seja $Cov(X, Y) = 20{,}0$, a variância amostral para a variável Y seja $Var(Y) = 4{,}0$ e a variância amostral para a variável X seja $Var(X) = 5{,}0$, a estimativa via estimador de mínimos quadrados ordinários para o coeficiente b é igual a 5,0.

Comentários:

Para calcular o coeficiente b, vamos aplicar a fórmula:

$$b = \frac{cov(X, Y)}{var(X)}$$

Substituindo pelos dados do enunciado, temos:

$$b = \frac{20}{5} = 4$$

Gabarito: Errado.

41. (CESPE/TCE-PA/2016) Uma regressão linear simples é expressa por $Y = a + b \times X + e$, em que o termo e corresponde ao erro aleatório da regressão e os parâmetros a e b são desconhecidos e devem ser estimados a partir de uma amostra disponível. Assumindo que a variável X é não correlacionada com o erro e, julgue o item subsecutivo, nos quais os resíduos das amostras consideradas são IID, com distribuição normal, média zero e variância constante.

Considere que as estimativas via método de mínimos quadrados ordinários para o parâmetro a seja igual a 2,5 e, para o parâmetro b, seja igual a 3,5. Nessa situação, assumindo que $X = 4{,}0$, o valor predito para Y será igual a 16,5, se for utilizada a reta de regressão estimada.

Comentários:

O modelo de regressão linear é expresso por:

$$Y = a + bX + e$$

A reta estimada é dada por:

$$\hat{Y} = \hat{a} + \hat{b}X$$

Substituindo pelos valores dados no enunciado, temos:

$$\hat{Y} = 2{,}5 + 3{,}5 \times 4$$

$$\hat{Y} = 16{,}5$$

Gabarito: Certo.

42. (CESPE/TCE-PA/2016). Uma regressão linear simples é expressa por $Y = a + b \times X + e$, em que o termo e corresponde ao erro aleatório da regressão e os parâmetros a e b são desconhecidos e devem ser estimados a partir de uma amostra disponível. Assumindo que a variável X é não correlacionada com o erro e, julgue o item subsecutivo, nos quais os resíduos das amostras consideradas são IID, com distribuição normal, média zero e variância constante.

A variável Y é denominada variável explicativa, e a variável X é denominada variável dependente.

Comentários:

O examinador inverteu os conceitos para tentar confundir os candidatos. No modelo de regressão linear, Y é a variável cujo comportamento desejamos prever ou explicar, sendo chamada de variável **dependente, explicada** ou **resposta**. Por outro lado, a variável X é utilizada para explicar o comportamento de Y_i, sendo conhecida como **independente**, **regressora**, **explanatória** ou **explicativa**.

Gabarito: Errado.

43. (FGV/DPE-RJ/2014) Considere a equação de regressão $Y_i = \alpha + \beta . X_i + \epsilon_i$, onde Y e X são as variáveis explicada e explicativa, respectivamente, ϵ é o erro aleatório e α e β os parâmetros a estimar. São supostos válidos todos os pressupostos clássicos do Modelo de Regressão Linear Simples (MRLS). Além disso, para determinada amostra de pares (X, Y), foram calculadas as estatísticas $p(X, Y) = 0,8$, $\bar{X} = 6$, $\bar{Y} = 15$, $DP(Y) = 5$ e $DP(X) = 2$. Portanto, a partir do método de Mínimos Quadrados Ordinários os estimadores de α e β são

a) 2 e 3.

b) 3 e 2.

c) -9 e 4.

d) 4 e -9.

e) 6 e 1,5.

Comentários:

Como vimos, o coeficiente de correlação entre as variáveis X e Y também pode ser expresso da seguinte forma:

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

O enunciado nos informou que $\rho(X, Y) = 0,8$, $\sigma_Y = 5$ e $\sigma_X = 2$. Daí, podemos deduzir que:

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

$$0,8 = \frac{Cov(X, Y)}{5 \times 2}$$

$$Cov(X, Y) = 8$$

A partir da covariância já podemos encontrar o coeficiente $\hat{\beta}$ estimado pelo método dos mínimos quadrados:

$$\hat{\beta} = \frac{Cov(X, Y)}{\sigma_X^2}$$

$$\hat{\beta} = \frac{8}{2^2}$$

$$\hat{\beta} = 2.$$

Nesse momento, já poderíamos assinalar a resposta correta (Letra B). Contudo, vamos ver o cálculo do intercepto (coeficiente α). Para tanto, devemos lembrar que a reta de regressão estimada pelo método dos mínimos quadrados sempre passa pelo ponto $(\bar{X}, \bar{Y})$. Desse modo,

$$\bar{Y} = \hat{\alpha} + \hat{\beta}\bar{X}$$

$$15 = \hat{\alpha} + 2 \times 6$$

$$\hat{\alpha} = 3.$$

Gabarito: B.

44. (FGV/SEN/2012) Um modelo probabilístico de primeira ordem, ou seja, de regressão linear pode ser representado da seguinte forma: $y = \beta_0 + \beta_1 x + \varepsilon$.

Com base nessa equação, avalie as afirmativas a seguir:

I. Neste modelo, pode-se sempre assumir que ε, o componente de erro aleatório, seja um ruído branco.

II. Uma vez que o valor esperado do erro, $E(\varepsilon)$, nem sempre será igual a zero, não é correto afirmar que o valor esperado de y, $E(y)$, será igual ao seu componente determinístico, $\beta_0 + \beta_1 x$.

III. Os símbolos gregos β_0 e β_1 são parâmetros populacionais que somente serão conhecidos se tiver acesso às medidas de toda a população de (x, y).

Assinale:

a) se apenas as afirmativas I e II forem verdadeiras.

b) se apenas as afirmativas I e III forem verdadeiras.

c) se apenas as afirmativas II e III forem verdadeiras.

d) se todas as afirmativas forem verdadeiras.

e) se nenhuma afirmativa for verdadeira.

Comentários:

Vamos analisar cada um dos itens:

I. **Correto**. O modelo de regressão linear supõe que o erro segue uma distribuição normal com média zero, o que caracteriza ser chamado de ruído branco. Por isso, tem-se que $E(\varepsilon) = 0$.

II. **Incorreto**. No modelo de regressão linear, assume-se que os resíduos têm distribuição com média zero: $E(\varepsilon_i) = 0$. Sendo assim,

$$E(y|X = x) = \beta_0 + \beta_1 x,$$

ou seja, o valor esperado de y é igual a $\beta_0 + \beta_1 x$.

III. **Correto**. O coeficiente de inclinação β_1 e o intercepto β_0 são de fato parâmetros populacionais e, portanto, desconhecidos. Os estimadores $\hat{\beta}_1$ e $\hat{\beta}_0$ são utilizados no modelo de regressão linear para o melhor ajuste entre valores observados e valores esperados.

Gabarito: B.

45. (FGV/SEFAZ RJ/2011) A tabela abaixo mostra os valores de duas variáveis, X e Y.

X	Y
4	4.5
4	5
3	5
2	5.5

Sabe-se que:

$$\sum X = 13$$

$$\sum Y = 20$$

$$\sum XY = 64$$

$$\sum X^2 = 45$$

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

O valor de "b" na regressão simples $Y = a + bX$ é

a) 11 /5.

b) -3 /8.

c) -4 /11.

d) -4 /17.

e) -11/65.

Comentários:

Primeiro calculamos as médias de X e de Y:

$$\bar{X} = \frac{\sum X}{n} = 134 = 3{,}25$$

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

O valor de b é calculado pela seguinte relação:

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

$$b = \frac{64 - 4 \times 3{,}25 \times 5}{45 - 4 \times 3{,}25^2}$$

$$b = \frac{64 - 65}{45 - 42{,}25}$$

$$b = -\frac{1}{2{,}75}$$

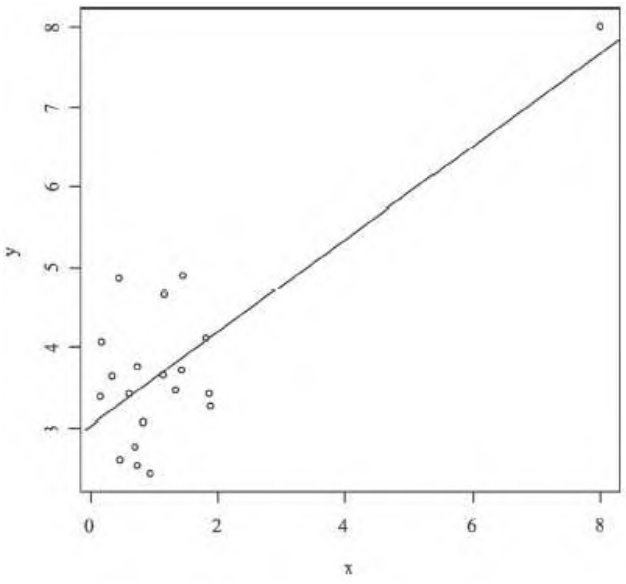
Multiplicando numerador e denominador por 4, temos:

$$b = -\frac{4}{11}$$

Gabarito: C.

46. (FGV/SEN/2008) A figura a seguir representa o diagrama de dispersão de dez pontos (X_i, Y_i) e a reta de regressão ajustada pelo método de mínimos quadrados dada por $Y = 0,42 + 2,45X$.

Quanto ao ponto de coordenadas $X = 8$ *e* $Y = 8$, pode-se afirmar que ele:



a) é o ponto com maior desvio da reta de regressão.

b) é um ponto influente nessa regressão.

c) é um dado legítimo que indica a relação linear entre X e Y.

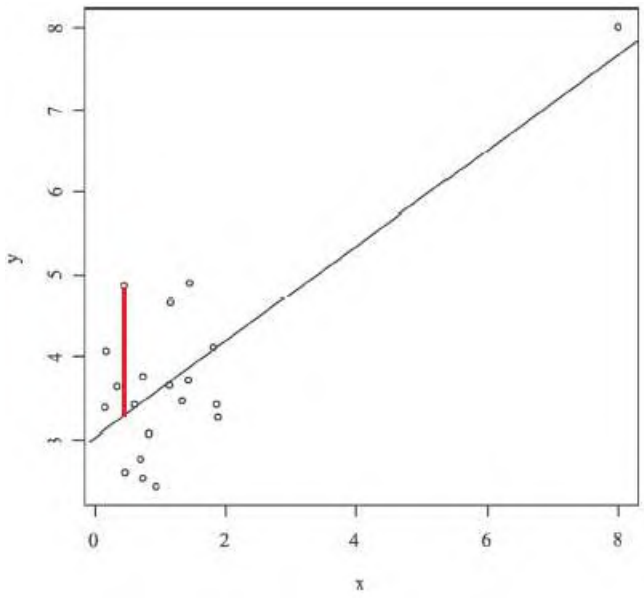
d) indica que o modelo é provavelmente heterocedástico.

e) é uma observação incorreta que deve ser eliminada da análise.

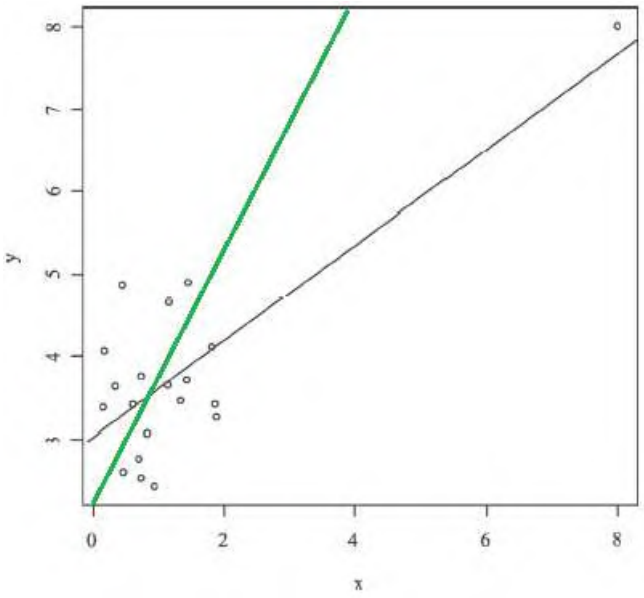
Comentários:

Vamos analisar cada alternativa:

Letra A: **Errada**. Um desvio é representado pela distância vertical de um ponto até a reta de regressão. No caso do ponto em destaque, está representado pela reta vermelha tracejada.



Letra B: **Correta**. Caso o ponto X = 8 e Y = 8 fosse desconsiderado na regressão, a reta de regressão seria algo como a reta verde. Isso indica que o ponto em questão tem forte influência na regressão.



Letra C: **Errada**. Um ponto qualquer de um conjunto de dados não diz nada sobre a linearidade ou não da relação entre duas variáveis. O que vai determinar a natureza dessa relação é o conjunto de dados como um todo. E a relação depende de cada caso.

Letra D: **Errada**. Não há dados suficiente para dizermos que ocorre heterocedasticidade. Para isso, é necessário fazer alguns testes de homoscedasticidade, uma análise da normalidade dos resíduos.

Letra E: **Errada**. Não podemos afirmar que se trata de uma medição incorreta. Existem testes específicos para a detecção de outliers em um conjunto de dados, entretanto, não significa que o dado é incorreto.

Gabarito: B.

QUESTÕES COMENTADAS

Análise de Variância da Regressão

1. (CESPE/TELEBRAS/2022) A tabela ANOVA a seguir se refere ao ajuste de um modelo de regressão linear simples escrito como $y = a + bx + \epsilon$, cujos coeficientes foram estimados pelo método da máxima verossimilhança, com $\epsilon \sim N(0, \sigma^2)$. Os erros em torno da reta esperada são independentes e identicamente distribuídos.

Fonte de Variação	Graus de Liberdade	Soma de Quadrados	Quadrado Médio
Modelo	1	10	10
Erro	99	990	10
Total	**100**	**1.000**	**10**

Com base nessas informações, julgue o item a seguir.

O coeficiente de explicação do modelo é igual a 0,99.

Comentários:

O coeficiente de explicação (ou coeficiente de determinação) resulta da divisão entre a soma dos quadrados do modelo e a soma dos quadrados total:

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

Observando a tabela, temos que:

$$SQM = 10$$

$$SQT = 1.000$$

Aplicando esses valores na fórmula anterior, temos:

$$R^2 = \frac{SQM}{SQT} = \frac{10}{1.000} = 0{,}01$$

Gabarito: Errado.

2. (CESPE/TELEBRAS/2022) O quadro a seguir mostra as estimativas de mínimos quadrados ordinários dos coeficientes de um modelo de regressão linear simples na forma $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, em que $i \in \{1, \cdots, 6\}$ e ϵ_i representa o erro aleatório com média zero e variância σ^2.

Coeficiente	Estimativa	Erro Padrão	Razão t
β_0	0,9	0,10	9
β_1	0,2	0,05	4

Considerando essas informações e sabendo que $\sigma^2 = 0,01$, julgue o item seguinte.

$SQ_{RESÍDUOS} = \sum_{i=1}^{6}(\hat{y}_i - \bar{y})^2 = 0,04$, em que $\hat{y}_i = 0,9 + 0,2x_i$.

Comentários:

A soma dos quadrados dos resíduos é calculada por meio da fórmula:

$$\sigma^2 = \frac{SQR}{n-2}$$

$$0,01 = \frac{SQR}{4} \Rightarrow SQR = 0,04$$

Portanto, a soma dos quadrados dos resíduos vale 0,04, conforme afirma a assertiva. Porém, essa soma é calculada pelo somatório dos quadrados das diferenças entre o valor previsto e o valor observado:

$$SQR = \sum_{i=1}^{6}(\hat{y}_i - y_i)^2$$

A assertiva chamou de soma dos quadrados dos resíduos o que, na verdade, é a soma dos quadrados do modelo:

$$SQM = \sum_{i=1}^{6}(\hat{y}_i - \bar{y})^2$$

Gabarito: Errado.

3. (CESPE/TELEBRAS/2022) O quadro a seguir mostra as estimativas de mínimos quadrados ordinários dos coeficientes de um modelo de regressão linear simples na forma $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, em que $i \in \{1, \cdots, 6\}$ e ϵ_i representa o erro aleatório com média zero e variância σ^2.

Coeficiente	Estimativa	Erro Padrão	Razão t
β_0	0,9	0,10	9
β_1	0,2	0,05	4

Considerando essas informações e sabendo que $\sigma^2 = 0,01$, julgue o item seguinte.

O coeficiente de determinação do modelo (R^2) é igual a 0,8.

Comentários:

O coeficiente de determinação do modelo é definido como

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

A partir da estimativa da variância dos termos de erro (σ^2), podemos calcular a soma dos quadrados dos resíduos:

$$\sigma^2 = \frac{SQR}{n-2}$$

$$0{,}01 = \frac{SQR}{4} \Rightarrow SQR = 0{,}04$$

A variância do estimador de β_1 é definida como

$$Var(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sigma^2}{S_{xx}}$$

Do enunciado, temos:

$$Var(\hat{\beta}_1) = 0{,}05^2 = 0{,}0025$$

e

$$\sigma^2 = 0{,}01.$$

Portanto,

$$0{,}0025 = \frac{0{,}01}{S_{xx}}$$

$$S_{xx} = \frac{0{,}01}{0{,}0025} = 4$$

Além disso,

$$SQM = \hat{\beta}_1^2 S_{xx} = (0{,}2)^2 \times 4 = 0{,}16$$

Portanto,

$$SQT = SQR + SQM = 0{,}04 + 0{,}16 = 0{,}2$$

Substituindo em (1),

$$R^2 = 1 - \frac{0{,}04}{0{,}2} = 0{,}8$$

Gabarito: Certo.

4. (CESPE/TELEBRAS/2022) O quadro a seguir mostra as estimativas de mínimos quadrados ordinários dos coeficientes de um modelo de regressão linear simples na forma $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, em que $i \in \{1, \cdots, 6\}$ e ϵ_i representa o erro aleatório com média zero e variância σ^2.

Coeficiente	Estimativa	Erro Padrão	Razão t
β_0	0,9	0,10	9
β_1	0,2	0,05	4

Considerando essas informações e sabendo que $\sigma^2 = 0,01$, julgue o item seguinte.

$$SQ_{TOTAL} = \sum_{i=1}^{6} (y_i - \bar{y})^2 = 0,2$$

Comentários:

A partir da estimativa da variância dos termos de erro (σ^2), podemos calcular a soma dos quadrados dos resíduos:

$$\sigma^2 = \frac{SQR}{n-2}$$

$$0,01 = \frac{SQR}{4} \Rightarrow SQR = 0,04$$

Do enunciado, temos:

$$Var(\hat{\beta}_1) = 0,05^2 = 0,0025$$

e

$$\sigma^2 = 0,01.$$

Portanto,

$$0,0025 = \frac{0,01}{S_{xx}}$$

$$S_{xx} = \frac{0,01}{0,0025} = 4$$

Além disso,

$$SQM = \hat{\beta}_1^2 S_{xx} = (0,2)^2 \times 4 = 0,16$$

Portanto,

$$SQT = SQR + SQM = 0,04 + 0,16 = 0,2.$$

Gabarito: Certo.

5. (CESPE/TCE-SC/2022) Em artigo publicado em 2004 no Journal of Political Economy, E. Miguel, S. Satyanath e E. Sergenti mostraram o efeito que o crescimento econômico pode ter na ocorrência de conflitos civis, com dados de 41 países africanos, no período de 1981 até 1999. Em certo estágio da pesquisa, para verificar a possibilidade de usar dados sobre precipitação pluviométrica como variável instrumental, foi feita uma regressão entre o crescimento de tais precipitações (variável explicativa) e uma variável resposta que representa um indicador para a ocorrência de conflito: quanto maior for esse indicador, maior a possibilidade de conflitos no ano *t* no país *i*. Os resultados do modelo ajustado pelo método de mínimos quadrados ordinários se encontram na tabela a seguir.

Variável Explicativa	Variável Dependente	
	Conflito civil (mínimo de 25 mortos)	Conflito civil (mínimo de 1000 mortos)
Crescimento na precipitação em *t*	–0,024 (0,043)	–0,062 (0,030)
Crescimento na precipitação em *t–1*	–0,122 (0,052)	–0,069 (0,032)
Efeitos fixos	sim	sim
R^2	0,71	0,70
Observações	743	743

Internet: <https://doi.org/10.1086/421174> (com adaptações).

Os números entre parênteses na tabela apresentada indicam o erro padrão da estimativa dos coeficientes respectivos. Considere os valores críticos t_α da variável *t* de Student, com significância α para os graus de liberdades adequados aos dados apresentados, como sendo $t_{10\%} = 1,65$, $t_{5\%} = 1,96$ e $t_{1\%} = 2,58$. Considerando as informações precedentes, julgue o próximo item.

As variáveis explicativas usadas explicam em torno de 71% das variações na ocorrência de conflito civil com um mínimo de 25 mortos nos países pesquisados, no período analisado.

Comentários:

O coeficiente de determinação R^2 mede a variabilidade da variável dependente explicada pelas variáveis independentes. Na primeira regressão, como $R^2 = 0{,}71 = 71\%$, podemos concluir que as variáveis explicativas são capazes de explicar, aproximadamente, 71% das variações da variável explicada (conflito civil com mínimo de 25 mortos).

Gabarito: Certo.

6. (CESPE/TELEBRAS/2022) A tabela ANOVA a seguir se refere ao ajuste de um modelo de regressão linear simples escrito como $y = a + bx + \epsilon$, cujos coeficientes foram estimados pelo método da máxima verossimilhança, com $\epsilon \sim N(0, \sigma^2)$. Os erros em torno da reta esperada são independentes e identicamente distribuídos.

Fonte de Variação	Graus de Liberdade	Soma de Quadrados	Quadrado Médio
Modelo	1	10	10
Erro	99	990	10
Total	**100**	**1.000**	**10**

Com base nessas informações, julgue o item a seguir.

A variância amostral da variável dependente é inferior a 12.

Comentários:

A variância amostral é a soma dos quadrados total dividida pelos graus de liberdade correspondentes. O resultado é fornecido na própria tabela, na coluna dos quadrados médios:

$$Var(y) = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1} = \frac{1000}{100} = 10$$

Gabarito: Certo.

7. (CESPE/TELEBRAS/2022) A tabela ANOVA a seguir se refere ao ajuste de um modelo de regressão linear simples escrito como $y = a + bx + \epsilon$, cujos coeficientes foram estimados pelo método da máxima verossimilhança, com $\epsilon \sim N(0, \sigma^2)$. Os erros em torno da reta esperada são independentes e identicamente distribuídos.

Fonte de Variação	Graus de Liberdade	Soma de Quadrados	Quadrado Médio
Modelo	1	10	10

Erro	99	990	10
Total	**100**	**1.000**	**10**

Com base nessas informações, julgue o item a seguir.

O R^2 ajustado é maior ou igual a 0,05.

Comentários:

O R^2 ajustado é calculado pela seguinte relação:

$$\overline{R^2} = 1 - \frac{QMR}{QMT}$$

Na tabela, observamos que os quadrados médios são iguais a 10:

$$QMR = QMT = 10.$$

O R^2 ajustado fica:

$$\overline{R^2} = 1 - \frac{10}{10} = 1 - 1 = 0$$

Gabarito: Errado.

8. (CESPE/TELEBRAS/2022) A tabela ANOVA a seguir se refere ao ajuste de um modelo de regressão linear simples escrito como $y = a + bx + \epsilon$, cujos coeficientes foram estimados pelo método da máxima verossimilhança, com $\epsilon \sim N(0, \sigma^2)$. Os erros em torno da reta esperada são independentes e identicamente distribuídos.

Fonte de Variação	Graus de Liberdade	Soma de Quadrados	Quadrado Médio
Modelo	1	10	10
Erro	99	990	10
Total	**100**	**1.000**	**10**

Com base nessas informações, julgue o item a seguir.

$\sigma^2 = 10$.

Comentários:

A estimativa da variância dos resíduos (σ^2) é calculada pela soma dos quadrados dos resíduos dividida pelos graus de liberdade correspondente. Com base na tabela, temos que

$$\sigma^2 = \frac{990}{99} = 10$$

que corresponde ao quadrado médio do erro, já discriminado na própria tabela.

Gabarito: Certo.

9. (CESPE/TELEBRAS/2022) A tabela ANOVA a seguir se refere ao ajuste de um modelo de regressão linear simples escrito como $y = a + bx + \epsilon$, cujos coeficientes foram estimados pelo método da máxima verossimilhança, com $\epsilon \sim N(0, \sigma^2)$. Os erros em torno da reta esperada são independentes e identicamente distribuídos.

Fonte de Variação	Graus de Liberdade	Soma de Quadrados	Quadrado Médio
Modelo	1	10	10
Erro	99	990	10
Total	**100**	**1.000**	**10**

Com base nessas informações, julgue o item a seguir.

Para se testar a hipótese nula $H_0: y = a + \epsilon$ contra a hipótese alternativa $H_1: y = a + bx + \epsilon$, a estatística do teste F proporcionada pela tabela ANOVA é igual ou superior a 2.

Comentários:

Ao verificar a hipótese nula $H_0: y = a + \epsilon$ contra a hipótese alternativa $H_1: y = a + bx + \epsilon$, o que se busca, em verdade, é testar se o coeficiente b é igual a zero, isto é, $H_0: b = 0$ contra $H_1: b \neq 0$.

O teste F tem sua estatística definida pela razão entre o quadrado médio do modelo e o quadrado médio dos resíduos (erros). Na presente questão, ambos os quadrados médios valem 10. Assim, a estatística fica:

$$F = \frac{10}{10} = 1$$

Portanto, a estatística F é inferior a 2.

Gabarito: Errado.

10. (CESPE/TELEBRAS/2022) Considere um modelo de regressão linear simples na forma $Y = aX + b + \epsilon$, em que ϵ representa o erro aleatório com média zero e desvio padrão σ, e a variável

regressora X é binária. A média amostral e o desvio padrão amostral da variável explicativa Y foram, respectivamente, iguais a 10 e 4. Já para a variável regressora X, encontra-se a distribuição de frequências absolutas mostrada no quadro a seguir. Finalmente, sabe-se que a correlação linear entre Y e X é igual a 0,9.

X	Frequência Absoluta
0	55
1	45
Total	**100**

Com base nessas informações, com respeito à reta ajustada pelo método dos mínimos quadrados ordinários, julgue o item subsequente.

O coeficiente de determinação do modelo é igual ou superior a 0,9.

Comentários:

O coeficiente de determinação (R^2) é o quadrado do coeficiente de correlação linear:

$$R^2 = (0{,}9)^2 = 0{,}81 < 0{,}9$$

Gabarito: Errado.

11. (FGV/EPE/2022) A respeito do coeficiente de determinação de uma regressão linear, avalie as afirmativas a seguir.

I. Mede a porcentagem da variância total que é explicada pela regressão.

II. É um número real entre 0 e 1.

III. É igual ao quadrado do coeficiente de correlação amostral.

Está correto o que se afirma em

a) I, apenas.

b) I e II, apenas.

c) I e III, apenas.

d) II e III, apenas.

e) I, II e III.

Comentários:

O coeficiente de determinação R^2 mede a variabilidade da variável dependente explicada pelas variáveis independentes. Portanto, mede a porcentagem da variância total que é explicada pela regressão. Além disso, é igual ao quadrado do coeficiente de correlação amostral e varia entre 0 e 1.

Gabarito: E.

12. (FGV/MPE SC/2022) É possível que o comportamento das bolsas de valores em determinado mês prediga o seu comportamento o ano inteiro. Considere que a variável explicativa X seja a variação percentual do índice da bolsa em janeiro e que a variável de resposta Y seja a variação desse índice para o ano inteiro. O cálculo feito com dados do período de 5 anos teve como resultados:

$$\bar{x} = 1,75\% \; \bar{y} = 9,07\%$$

$$S_x = 5,36\% \; S_y = 15,35\%$$

$$r = 0,59$$

O percentual de variação observado nas alterações anuais do índice que é explicado pela relação linear com a alteração de janeiro é:

a) 2,86%;

b) 5,18%;

c) 34,81%;

d) 35,50%;

e) 59%.

Comentários:

O coeficiente de determinação R^2 mede a variabilidade da variável dependente explicada pelas variáveis independentes. Além disso, é igual ao quadrado do coeficiente de correlação amostral. Assim, temos que:

$$r^2 = 0,59^2 = 0,3481$$

Gabarito: C.

13. (CESPE/MJ-SP/2021) A tabela de análise de variância a seguir se refere a um modelo de regressão linear simples na forma $y = ax + b + \epsilon$, na qual $\epsilon \sim N(0, \sigma^2)$. Os resultados da tabela foram obtidos com base em uma amostra aleatória simples n de pares de observações independentes (x,y).

Fonte de Variação	Graus de Liberdade	Soma de Quadrados

Regressão	1	82
Resíduos	8	8
Total	**9**	**90**

Com base nessas informações, julgue o item subsequente.

O coeficiente de explicação ajustado (R^2 ajustado) é igual a 0,90.

Comentários:

O coeficiente de determinação (R^2) é a razão entre a soma dos quadrados do modelo e a soma dos quadrados total:

$$R^2 = \frac{SQM}{SQT} = \frac{82}{90}$$

O coeficiente ajustado é definido como:

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

em que n é o tamanho amostral.

Com 9 graus de liberdade para o total, temos que:

$$n = 9 + 1 = 10.$$

Substituindo na equação do coeficiente de determinação ajustado, temos:

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

$$\overline{R^2} = 1 - \left(\frac{10-1}{10-2}\right)\left(1 - \frac{82}{90}\right)$$

$$\overline{R^2} = 1 - \left(\frac{9}{8}\right)\left(\frac{90-82}{90}\right)$$

$$\overline{R^2} = 1 - \left(\frac{9}{8}\right)\left(\frac{8}{90}\right)$$

$$\overline{R^2} = 1 - 0{,}1 = 0{,}9$$

Gabarito: Certo.

14. (CESPE/MJ-SP/2021) A tabela de análise de variância a seguir se refere a um modelo de regressão linear simples na forma $y = ax + b + \epsilon$, na qual $\epsilon \sim N(0, \sigma^2)$. Os resultados da tabela foram obtidos com base em uma amostra aleatória simples n de pares de observações independentes (x,y).

Fonte de Variação	Graus de Liberdade	Soma de Quadrados
Regressão	1	82
Resíduos	8	8
Total	9	90

Com base nessas informações, julgue o item subsequente.

O quadrado da razão t do teste de hipóteses $H_0: a = 0$ *versus* $H_1: a \neq 0$ é igual a 16.

Comentários:

Para o teste de hipóteses de um único coeficiente, a estatística F corresponde ao quadrado da estatística t. A estatística F é calculada pela razão entre o quadrado médio do modelo e o quadrado médio dos resíduos:

$$F = \frac{QM_{modelo}}{QM_{resíduos}}$$

Dividindo as somas dos quadrados pelos graus de liberdade correspondentes, temos:

$$QM_{modelo} = \frac{82}{1} = 82$$

$$QM_{resíduos} = \frac{8}{8} = 1$$

$$F = \frac{82}{1} = 82$$

Portanto, o quadrado da razão t do teste é igual a 82.

Gabarito: Errado.

15. (CESPE/ALECE/2021) Um modelo de regressão linear simples tem a forma $y = ax + b + \varepsilon$, em que y denota a variável resposta, x é a variável regressora, a e b são os coeficientes do modelo, e ε representa um erro aleatório com média 0 e variância σ^2. Com base em uma amostra aleatória simples de tamanho $n = 51$, pelo método dos mínimos quadrados ordinários, a estimativa da variância v foi igual 3. A variância amostral da variável y é 42.

Nesse modelo, o valor do coeficiente de determinação (R^2) é igual a

a) 0,07.

b) 0,21.

c) 0,93.

d) 0,42.

e) 0,79.

Comentários:

O coeficiente de determinação (R^2) é a razão entre a soma dos quadrados do modelo e a soma dos quadrados total:

$$R^2 = \frac{SQM}{SQT} = \frac{82}{90}$$

A estimativa da variância dos resíduos é igual a 3, pois o quadrado médio dos erros vale 3. Portanto,

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

$$3 = \frac{SQR}{51-2}$$

$$SQR = 3 \times 49 = 147$$

A variância amostral da variável y é 42. Ao multiplicar esse valor por $n-1$ graus de liberdade, encontramos a soma dos quadrados total:

$$SQT = 42 \times 50 = 2.100$$

Portanto, o coeficiente de determinação é:

$$R^2 = 1 - \frac{147}{2.100} = 0{,}93$$

Gabarito: C.

16. (CESPE/MJ-SP/2021) Acerca de planejamento de pesquisa estatística, julgue o item que se seguem.

Em um modelo estatístico, o erro total corresponde à soma dos desvios das observações em relação ao modelo.

Comentários:

O erro total corresponde à soma dos desvios das observações em relação ao modelo:

$$Erro\ total = \sum_{i=1}^{n} e_i = \sum_{i=1}^{n} y_i - \hat{y}_i$$

O erro total não deve ser confundido com a soma dos quadrados totais, que é a soma dos quadrados dos desvios entre as observações e os valores das predições:

$$SQT = \sum_{i=1}^{N} {e_i}^2 = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2$$

Gabarito: Certo.

17. (FGV/FunSaúde CE/2021)Numa regressão linear, as afirmativas a seguir, acerca do coeficiente de determinação, estão corretas, exceto uma. Assinale-a.

a) Mede a porcentagem da variação total da variável resposta que é explicada pela regressão.

b) É o quadrado do coeficiente de correlação estimado.

c) É um número entre 0 e 1.

d) Determina se as estimativas e predições dos coeficientes são tendenciosas.

e) Em geral, mas nem sempre, quanto maior seu valor, melhor o modelo se ajusta aos dados.

Comentários:

O coeficiente de determinação R^2 mede a variabilidade da variável dependente explicada pelas variáveis independentes. Como principais características, temos:

- mede a porcentagem da variação total da variável resposta que é explicada pela regressão.
- é igual o quadrado do coeficiente de correlação estimado.
- é um número entre 0 e 1.
- em geral, quanto maior seu valor, melhor o modelo se ajusta aos dados.

Gabarito: D.

18. (VUNESP/EBSERH/2020) Numa regressão linear simples em que foi utilizada uma amostra com 52 observações, a soma dos quadrados totais é de 50 e a soma dos quadrados dos resíduos é de 20. O coeficiente de determinação e a estatística F dessa regressão são, respectivamente:

a) 0,6 e 75.

b) 0,6 e 12.

c) 0,8 e 1,5.

d) 0,8 e 12.

e) 0,8 e 75.

Comentários:

O coeficiente de determinação da regressão linear é expresso por:

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

Quanto mais próximo de 1 estiver o coeficiente de determinação, mais forte será a correlação linear. Implica dizer que as diferenças entre os valores observados (Y_i) e a média ($\bar{Y}$) são quase totalmente explicados pela reta de regressão. Por outro lado, quanto mais próximo de 0 estiver o coeficiente de determinação, mais fraca será a correlação linear, isto é, a reta de regressão não é capaz de explicar as diferenças entre os valores observados e a média.

Substituindo os valores, temos:

$$R^2 = 1 - \frac{20}{50}$$

$$R^2 = 1 - 0{,}4$$

$$R^2 = 0{,}6$$

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 52 observações, logo:

$$F = \frac{30 \times (52-2)}{20}$$

$$F = 1{,}5 \times 50$$

$$F = 75$$

Outra forma de fazer é:

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

$$F = \frac{030 \times (52-2)}{1-0{,}6}$$

$$F = \frac{30}{0{,}4}$$

$$F = 75$$

Gabarito: A.

19. (CESPE/TJ-AM/2019) Um modelo de regressão linear foi ajustado para explicar os sintomas de transtornos mentais (T) em função da violência intrafamiliar (V) e do inventário do clima familiar (C). A forma desse modelo é dada por $T = b_0 + b_1V + b_2C + \epsilon$, em que ϵ representa o erro aleatório normal com média zero e desvio padrão σ, e b_0, b_1 e b_2 são os coeficientes do

modelo. A tabela a seguir mostra os resultados da análise de variância (ANOVA) do referido modelo.

Com base na tabela e nas informações apresentadas, julgue o item a seguir.

Fonte de variação	Graus de Liberdade	Soma de Quadrados	Média dos Quadrados	Razão F	p-valor
Regressão	2	608	304	76	<0,0001
Resíduo	98	392	4		
Total	100	1.000			

Conjuntamente, segundo o modelo ajustado, a violência intrafamiliar e o inventário do clima familiar explicam 60,8% da variabilidade total dos sintomas de transtornos mentais.

Comentários:

O coeficiente de determinação da regressão linear é dado por:

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

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

$SQR \rightarrow Soma\ dos\ quadrados\ dos\ resíduos$

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

Substituindo pelos valores da tabela, temos:

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

$$R^2 = \frac{608}{1000} = 0{,}608 = 60{,}8\%$$

Gabarito: Certo.

20. (CESPE/COGE-CE/2019) Considerando-se que, em uma regressão múltipla de dados estatísticos, a soma dos quadrados da regressão seja igual a 60.000 e a soma dos quadrados dos erros seja igual a 15.000, é correto afirmar que o coeficiente de determinação — R^2 — é igual a

a) 0,75.

b) 0,25.

c) 0,50.

d) 0,20.

e) 0,80.

Comentários:

O coeficiente de determinação da regressão linear é dado por:

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

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

$SQR \rightarrow Soma\ dos\ quadrados\ dos\ resíduos$

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

Do enunciado temos:

$SQR = 15.000$

$SQM = 60.000$

Logo,

$SQT = 75.000$

Aplicando fica:

$$R^2 = \frac{60.000}{75.000} = 0{,}8$$

Gabarito: E.

21. (CESPE/TJ-AM/2019) 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.

A correlação linear entre as variáveis x e y é superior a 0,9.

Comentários:

Sabemos que o coeficiente de correlação linear é igual a R e que o coeficiente de determinação é igual a R^2. Então, temos:

$$R = \sqrt{R^2}$$

$$R = \sqrt{0{,}85}$$

$$R \cong 0{,}92$$

Gabarito: Certo.

22. (CESPE/PF/2018). Um pesquisador estudou a relação entre a taxa de criminalidade (Y) e a taxa de desocupação da população economicamente ativa (X) em determinada região do país. Esse 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	225
Erro	899	175
Total	900	400

A respeito dessa situação hipotética, julgue o item, sabendo que b > 0 e que o desvio padrão amostral da variável X é igual a 2.

A correlação linear de Pearson entre a variável resposta Y e a variável regressora X é igual a 0,75.

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:

$$R^2 = \frac{225}{400}$$

$$R^2 = 0{,}5625$$

$$R = \sqrt{0{,}5625}$$

$$R = 0{,}75$$

Gabarito: Certo.

23. (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.

24. (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:

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

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

Pela tabela, verificamos que $SQT = 1000$ e $SQM = 900$. 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.

25. (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.

26. (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.

27. (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.

28. (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 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 $SQM = 900$. 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.

29. (CESPE/TCE-PE/2017). Um estudo de acompanhamento ambiental considerou, para $j = 1, 2, \ldots, 26$, um modelo de regressão linear simples na forma: $y_j = a + bx_j + e_j$, em que a e b são constantes reais, y_j representa a variável resposta referente ao j-ésimo elemento da amostra, x_j é a variável regressora correspondente, e e_j denota o erro aleatório que segue distribuição normal com média nula e variância V. Aplicando-se, nesse estudo, o método dos mínimos quadrados ordinários, obteve-se a reta ajustada $\hat{y}_j = 1 + 2x_j, para\ j = 1, 2, \ldots, 26$

Considerando que a estimativa da variância V seja igual a 6 e que o coeficiente de explicação do modelo (R quadrado) seja igual a 0,64, julgue o item.

A correlação linear entre as variáveis x e y é igual a 0,5, pois a reta invertida proporcionada pelo método de mínimos quadrados ordinários é expressa por $\hat{x}_j = 0{,}5y_i - 0{,}5$, para $j = 1{,}2, \ldots, 26$

Comentários:

Sabemos que o coeficiente de correlação linear é igual a R, e que o coeficiente de determinação é igual a R^2. Então, temos:

$$R = \sqrt{R^2}$$

Com os dados do enunciado temos:

$$R = \sqrt{0{,}64}$$

$$R = 0{,}8$$

Gabarito: Errado.

30. (FGV/IBGE/2017) Após formular e estimar um modelo de regressão simples, o estatístico responsável pela análise trabalha nos resultados, defrontando-se com a tabela a seguir:

Fonte	S. Quadrados	G.L.	Q. Médio	F-Snedecor	p-value
Equação	450	1	450	12,00	0,21%
Resíduos	300	8	37,50		
Total	750	9	83,33		

A partir desses números, é correto concluir que:

a) com uma amostra de tamanho 10, o modelo é capaz de explicar 60% da variação total;

b) a variância estimada do erro aleatório é inferior a 6;

c) apesar de um poder de explicação de 60%, o modelo não passa no teste de significância da estatística F;

d) a variância da variável explicativa do modelo é igual a 75;

e) a estimativa do coeficiente angular da equação é igual a 2.

Comentários:

Letra A: **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:

$$R^2 = \frac{450}{750} = 0{,}6 = 60\%.$$

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

$$\sigma^2 = QM_{Res} = 37{,}50.$$

Letra C: **Errado**. O p-valor próximo de zero significa que o modelo proposto é adequado, pois fornece evidências contra a hipótese nula. Se fixarmos um nível de significância, α, podemos dizer que uma hipótese nula é rejeitada a este nível quando o p-valor é menor do que esse α.

Letra D: **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 450.

Letra E: **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 2.

Gabarito: A.

31. (CESPE/TCE-SC/2016). Um auditor foi convocado para verificar se o valor de Y, doado para a campanha de determinado candidato, estava relacionado ao valor de X, referente a contratos firmados após a sua eleição.

Tabela de análise de variância de dados					
Fonte de variação	**Graus de Liberdade**	**Soma de Quadrados**	**Quadrados Médios**	**F**	**Pr > F**
Modelo	1	4,623		9,76	0,0261
Erro	5	2,371			
Total	6	7,000			

Com base na situação hipotética e na tabela apresentadas, julgue o item que se segue, considerando-se que $\sum(x_i - \bar{x})^2 = 17,5$ e $E(y^2) = 7,25$

O coeficiente angular é maior que 1.

Comentários:

Sabemos que:

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

Em que

$$b \rightarrow \text{é } a\ estimativa\ do\ coeficiente\ angular\ da\ reta\ de\ regressão.$$

Substituindo, temos:

$$4{,}623 = b^2 \times 17{,}5$$

$$b^2 \cong 0{,}26$$

Logo, o coeficiente angular é menor que 1.

Gabarito: Errado.

32. (CESPE/TCE-PA/2016). Uma regressão linear simples é expressa por $Y = a + b \times X + e$, em que o termo e corresponde ao erro aleatório da regressão e os parâmetros a e b são desconhecidos e devem ser estimados a partir de uma amostra disponível. Assumindo que a variável X é não correlacionada com o erro e, julgue o item subsecutivo, nos quais os resíduos das amostras consideradas são IID, com distribuição normal, média zero e variância constante.

Se, em uma amostra de tamanho n = 25, o coeficiente de correlação entre as variáveis X e Y for igual a 0,8, o coeficiente de determinação da regressão estimada via mínimos quadrados ordinários, com base nessa amostra, terá valor $R^2 = 0{,}64$.

Comentários:

Sabemos que o coeficiente de correlação linear é igual a R, e que o coeficiente de determinação é igual a R^2. Então:

$$R^2 = 0{,}8^2$$

$$R^2 = 0{,}64$$

Gabarito: Certo.

33. (CESPE/TCE-PA/2016)

Fonte de variação	Graus de Liberdade	Soma de Quadrados	Quadrados Médios	F	Pr > F
Modelo			900		
Erro	98				
Total			90		

Considerando um modelo de regressão linear simples, para averiguar se existe alguma relação entre o salário pago — Y — para uma pessoa em cargo comissionado e o tempo de trabalho — X — dessa pessoa na campanha de determinado padrinho político eleito, foi escolhida uma

amostra de indivíduos em cargos comissionados cujos resultados estão apresentados nessa tabela.

Com base nessa situação hipotética e nos dados apresentados na tabela, julgue o item que se segue, relativo à análise de regressão e amostragem.

A variância de Y é maior que 100.

Comentários:

Analisando a tabela dada, temos:

- 98 é o grau de liberdade do resíduo, corresponde a (n - 2);
- 900 é o Quadrado Médio do Modelo (QMM);
- 90 é o Quadrado Médio Total (QMT);

Logo, o QMT corresponde à variância da amostra, no caso, é 90. Ou seja, inferior a 100.

Gabarito: Errado.

34. (FGV/Pref. Recife/2014) Numa regressão linear simples, obteve-se um coeficiente de correlação igual a 0,78. O coeficiente de determinação é aproximadamente igual a

a) 0,36.

b) 0,48.

c) 0,50.

d) 0,61.

e) 0,69.

Comentários:

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

$$R^2 = 0{,}78^2 = 0{,}6084$$

Gabarito: D.

35. (FGV/SEAD-AP/2010) Se no ajuste de uma reta de regressão linear simples de uma variável Y em uma variável X o coeficiente de determinação observado foi igual a 0,64, então o módulo do coeficiente de correlação amostral entre X e Y é igual a:

a) 0,24

b) 0,36

c) 0,50

d) 0,64

e) 0,80

Comentários:

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

$$R^2 = 0{,}64$$

Logo,

$$R = \pm\sqrt{0{,}64}$$

$$R = \pm 0{,}8$$

O módulo de R vale 0,8.

Gabarito: E.

Lista de Questões

Correlação Linear

1. (CESPE/TJ-PA/2020) Em um gráfico de dispersão, por meio de transformações convenientes, a origem foi colocada no centro da nuvem de dispersão e as variáveis foram reduzidas a uma mesma escala. Se, nesse gráfico, for observado que a grande maioria dos pontos está situada no segundo e no quarto quadrantes, e que aqueles que não estão nessa posição situam-se próximos da origem, então a correlação linear entre as variáveis

a) Será necessariamente fortemente positiva.

b) Poderá ser fracamente positiva.

c) Será necessariamente nula.

d) Poderá ser fracamente negativa.

e) Será necessariamente fortemente negativa.

2. (CESPE/TJ-PA/2020) Texto 7A3-I. O coeficiente de correlação linear de Pearson entre duas variáveis aleatórias discretas X e Y definidas sobre um mesmo espaço amostral é dado por

$$CORR(X,Y) = \frac{n(\sum_{i=1}^{n} x_i y_i) - (\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{\sqrt{n(\sum_{i=1}^{n} x_i^2) - (\sum_{i=1}^{n} x_i)^2}\sqrt{n(\sum_{i=1}^{n} y_i^2) - (\sum_{i=1}^{n} y_i)^2}}.$$

Já na reta de melhor ajuste $Y = aX + b$, determinada pelo método dos mínimos quadrados, os coeficientes são dados por

$$a = \frac{n(\sum_{i=1}^{n} x_i y_i) - (\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n(\sum_{i=1}^{n} x_i^2) - (\sum_{i=1}^{n} x_i)^2}$$

$$b = \frac{\sum_{i=1}^{n} y_i - a\sum_{i=1}^{n} x_i}{n}$$

Uma forma de avaliar a precisão do modelo consiste em comparar o estimador não viesado da variância residual, obtido das diferenças entre os valores observados e os previstos pelo modelo, $\widehat{S}_e = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2$, com o estimador não viesado da variância dos valores observados, $\widehat{S}_e = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2$.

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, em que a última coluna foi obtida usando a reta ajustada pelo método dos mínimos quadrados.

Réu	P	R	$P \times R$	P^2	R^2	$(R - \overline{R})^2$	$(R - \widehat{R})^2$

1	14	0,25	3,5	196	0,0625	3,0625	0,0547560
2	12	0,5	6	144	0,25	2,25	0,0001440
3	10,9	1	10,9	118,81	1	1	0,0463110
4	6	1,5	9	36	2,25	0,25	0,2500000
5	5	1,75	8,75	25	3,0625	0,0625	0,2480040
6	3	2	6	9	4	0	0,5535360
7	3	2,5	7,5	9	6,25	0,25	0,0595360
8	2,3	3	6,9	5,29	9	1	0,0067898
9	1,8	3,5	6,3	3,24	12,25	2,25	0,2101306
10	2	4	8	4	16	4	1,0160640
Totais	**60**	**20**	**72,85**	**550,34**	**54,125**	**14,125**	**2,4452714**

Dados:

$\mathbf{1903,4^{1/2} = 43,63}$

$\mathbf{141,25^{1/2} = 11,88}$

A partir das informações do texto 7A3-I, o coeficiente de correlação linear entre as variáveis R e P é

a) – 0,33.

b) – 0,51.

c) – 0,67.

d) – 0,82.

e) – 0,91.

3. (CESPE/TJ-PA/2020) Texto 7A3-I. O coeficiente de correlação linear de Pearson entre duas variáveis aleatórias discretas X e Y definidas sobre um mesmo espaço amostral é dado por

$$CORR(X,Y) = \frac{n(\sum_{i=1}^{n} x_i y_i) - (\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{\sqrt{n(\sum_{i=1}^{n} x_i^2) - (\sum_{i=1}^{n} x_i)^2}\sqrt{n(\sum_{i=1}^{n} y_i^2) - (\sum_{i=1}^{n} y_i)^2}}.$$

Já na reta de melhor ajuste $Y = aX + b$, determinada pelo método dos mínimos quadrados, os coeficientes são dados por

$$a = \frac{n(\sum_{i=1}^{n} x_i y_i) - (\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n(\sum_{i=1}^{n} x_i^2) - (\sum_{i=1}^{n} x_i)^2}$$

$$b = \frac{\sum_{i=1}^{n} y_i - a\sum_{i=1}^{n} x_i}{n}$$

Uma forma de avaliar a precisão do modelo consiste em comparar o estimador não viesado da variância residual, obtido das diferenças entre os valores observados e os previstos pelo modelo, $\widehat{S}_e = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$, com o estimador não viesado da variância dos valores observados, $\widehat{S}_e = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$.

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, em que a última coluna foi obtida usando a reta ajustada pelo método dos mínimos quadrados.

Réu	P	R	$P \times R$	P^2	R^2	$(R - \bar{R})^2$	$(R - \widehat{R})^2$
1	14	0,25	3,5	196	0,0625	3,0625	0,0547560
2	12	0,5	6	144	0,25	2,25	0,0001440
3	10,9	1	10,9	118,81	1	1	0,0463110
4	6	1,5	9	36	2,25	0,25	0,2500000
5	5	1,75	8,75	25	3,0625	0,0625	0,2480040
6	3	2	6	9	4	0	0,5535360
7	3	2,5	7,5	9	6,25	0,25	0,0595360
8	2,3	3	6,9	5,29	9	1	0,0067898
9	1,8	3,5	6,3	3,24	12,25	2,25	0,2101306
10	2	4	8	4	16	4	1,0160640
Totais	**60**	**20**	**72,85**	**550,34**	**54,125**	**14,125**	**2,4452714**

Dados:

$1903,4^{1/2} = 43,63$

$141,25^{1/2} = 11,88$

Considerando-se o texto 7A3-I, a relação entre o coeficiente de correlação linear entre as variáveis X e Y e o coeficiente angular, da reta de melhor ajuste aos dados determinada pelo método dos mínimos quadrados pode ser expressa por

a) $a = CORR(X, Y)$.

b) $b = CORR(X, Y)$.

c) $a \times \sqrt{n(\sum_{i=1}^{n} x_i^2) - (\sum_{i=1}^{n} x_i)^2} = CORR(X, Y) \times \sqrt{n(\sum_{i=1}^{n} y_i^2) - (\sum_{i=1}^{n} y_i)^2}$.

d) $b \times \sqrt{n(\sum_{i=1}^{n} x_i^2) - (\sum_{i=1}^{n} x_i)^2} = CORR(X, Y) \times \sqrt{n(\sum_{i=1}^{n} y_i^2) - (\sum_{i=1}^{n} y_i)^2}$.

e) $a = \frac{1}{CORR(X,Y)}$.

4. (CESPE/TJ-PA/2020) Texto 7A3-I. O coeficiente de correlação linear de Pearson entre duas variáveis aleatórias discretas X e Y definidas sobre um mesmo espaço amostral é dado por

$$CORR(X, Y) = \frac{n(\sum_{i=1}^{n} x_i y_i) - (\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{\sqrt{n(\sum_{i=1}^{n} x_i^2) - (\sum_{i=1}^{n} x_i)^2}\sqrt{n(\sum_{i=1}^{n} y_i^2) - (\sum_{i=1}^{n} y_i)^2}}.$$

Já na reta de melhor ajuste $Y = aX + b$, determinada pelo método dos mínimos quadrados, os coeficientes são dados por

$$a = \frac{n(\sum_{i=1}^{n} x_i y_i) - (\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n(\sum_{i=1}^{n} x_i^2) - (\sum_{i=1}^{n} x_i)^2}$$

$$b = \frac{\sum_{i=1}^{n} y_i - a\sum_{i=1}^{n} x_i}{n}$$

Uma forma de avaliar a precisão do modelo consiste em comparar o estimador não viesado da variância residual, obtido das diferenças entre os valores observados e os previstos pelo modelo, $\widehat{S}_e = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$, com o estimador não viesado da variância dos valores observados, $\widehat{S}_e = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$.

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, em que a última coluna foi obtida usando a reta ajustada pelo método dos mínimos quadrados.

Réu	P	R	$P \times R$	P^2	R^2	$(R - \bar{R})^2$	$(R - \hat{R})^2$
1	14	0,25	3,5	196	0,0625	3,0625	0,0547560
2	12	0,5	6	144	0,25	2,25	0,0001440

3	10,9	1	10,9	118,81	1	1	0,0463110
4	6	1,5	9	36	2,25	0,25	0,2500000
5	5	1,75	8,75	25	3,0625	0,0625	0,2480040
6	3	2	6	9	4	0	0,5535360
7	3	2,5	7,5	9	6,25	0,25	0,0595360
8	2,3	3	6,9	5,29	9	1	0,0067898
9	1,8	3,5	6,3	3,24	12,25	2,25	0,2101306
10	2	4	8	4	16	4	1,0160640
Totais	**60**	**20**	**72,85**	**550,34**	**54,125**	**14,125**	**2,4452714**

Dados:

$\mathbf{1903,4^{1/2} = 43,63}$

$\mathbf{141,25^{1/2} = 11,88}$

Com base no texto 7A3-I, a renda familiar per capita esperada X, em número de salários-mínimos, obtida aplicando-se a reta de melhor ajuste aos dados determinada pelo método dos mínimos quadrados para um réu ao qual tenha sido cominada uma pena de 4 anos de reclusão é

a) $2{,}3 < X < 2{,}6$.

b) $2{,}1 < X < 2{,}3$.

c) $1{,}9 < X < 2{,}1$.

d) $1{,}2 < X < 1{,}9$.

e) $1{,}0 < X < 1{,}2$.

5. (CESPE/TJ-PA/2020) Texto 7A3-I. O coeficiente de correlação linear de Pearson entre duas variáveis aleatórias discretas X e Y definidas sobre um mesmo espaço amostral é dado por

$$CORR(X,Y) = \frac{n(\sum_{i=1}^{n} x_i y_i) - (\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{\sqrt{n(\sum_{i=1}^{n} x_i^2) - (\sum_{i=1}^{n} x_i)^2}\sqrt{n(\sum_{i=1}^{n} y_i^2) - (\sum_{i=1}^{n} y_i)^2}}.$$

Já na reta de melhor ajuste $Y = aX + b$, determinada pelo método dos mínimos quadrados, os coeficientes são dados por

$$a = \frac{n(\sum_{i=1}^{n} x_i y_i) - (\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n(\sum_{i=1}^{n} x_i^2) - (\sum_{i=1}^{n} x_i)^2}$$

$$b = \frac{\sum_{i=1}^{n} y_i - a\sum_{i=1}^{n} x_i}{n}$$

Uma forma de avaliar a precisão do modelo consiste em comparar o estimador não viesado da variância residual, obtido das diferenças entre os valores observados e os previstos pelo modelo, $\widehat{S}_e = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$, com o estimador não viesado da variância dos valores observados, $\widehat{S}_e = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2$.

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, em que a última coluna foi obtida usando a reta ajustada pelo método dos mínimos quadrados.

Réu	P	R	$P \times R$	P^2	R^2	$(R - \overline{R})^2$	$(R - \widehat{R})^2$
1	14	0,25	3,5	196	0,0625	3,0625	0,0547560
2	12	0,5	6	144	0,25	2,25	0,0001440
3	10,9	1	10,9	118,81	1	1	0,0463110
4	6	1,5	9	36	2,25	0,25	0,2500000
5	5	1,75	8,75	25	3,0625	0,0625	0,2480040
6	3	2	6	9	4	0	0,5535360
7	3	2,5	7,5	9	6,25	0,25	0,0595360
8	2,3	3	6,9	5,29	9	1	0,0067898
9	1,8	3,5	6,3	3,24	12,25	2,25	0,2101306
10	2	4	8	4	16	4	1,0160640
Totais	**60**	**20**	**72,85**	**550,34**	**54,125**	**14,125**	**2,4452714**

Dados:

$1903,4^{1/2} = 43,63$

$141,25^{1/2} = 11,88$

Levando-se em consideração o texto 7A3-I, a discrepância na renda familiar per capita X, em número de salários-mínimos, obtida entre o valor observado e aquele em que se aplica a reta de melhor ajuste aos dados determinada pelo método dos mínimos quadrados para o nono réu é

a) $0{,}47 < X < 0{,}50$.

b) $0{,}44 < X < 0{,}47$.

c) $0{,}42 < X < 0{,}44$.

d) $0{,}39 < X < 0{,}42$.

e) $0{,}38 < X < 0{,}39$.

6. (VUNESP/EBSERH/2020) Dados para responder à questão.

A variável x tem média 4 e desvio padrão 2, enquanto a variável y tem média 3 e desvio padrão 1. A covariância entre x e y é –1.

O coeficiente de correlação entre x e y é

a) 0,5.

b) –0,5.

c) 1.

d) –1.

e) –0,25.

GABARITO

Correlação Linear

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

LISTA DE QUESTÕES

Regressão Linear Simples

1. (CESPE/SEFAZ-SE/2022) Para a obtenção de projeções de resultados financeiros de empresas de determinado ramo de negócios, será ajustado um modelo de regressão linear simples na forma $y = ax + b + \epsilon$, no qual x representa o grau de endividamento; y denota um índice contábil; o termo ϵ é o erro aleatório, que segue uma distribuição com média nula e variância σ^2; e a e b são os coeficientes do modelo, com $b \neq 0$. A correlação linear entre as variáveis x e y é positiva e algumas medidas descritivas referentes às variáveis x e y se encontram na tabela a seguir.

	y	x
Média Amostral	2	4
Desvio Padrão Amostral	0,4	8

Com base nessa situação hipotética e considerando que o coeficiente de determinação proporcionado pelo modelo em tela seja $R^2 = 0,81$, assinale a opção em que é apresentada a reta ajustada pelo critério de mínimos quadrados ordinários.

a) $\hat{y} = 0,045x + 1,82$

b) $\hat{y} = 0,5x$

c) $\hat{y} = 0,4x + 0,4 + \epsilon$

d) $\hat{y} = 18x - 70 + \epsilon$

e) $\hat{y} = 18x - 70$

2. (CESPE/PC-PB/2022) Para as variáveis Y e X, em que Y denota a variável resposta e X representa a variável regressora, a correlação linear de Pearson entre Y e X é $0,8$, o desvio padrão amostral de Y é 2, e o desvio padrão amostral de X é 4. Nesse caso, a estimativa de mínimos quadrados ordinários do coeficiente angular da reta de regressão linear simples é igual a

a) 0,40.

b) 1,60.

c) 0,64.

d) 0,80.

e) 0,50.

3. (CESPE/PETROBRAS/2022) Uma determinada repartição pública fez um levantamento do tempo, em minutos, que os cinco funcionários de uma sessão gastam para chegar ao trabalho em função da distância x, em quilômetros, de suas residências. O resultado da pesquisa realizada com cada um deles é apresentado na tabela a seguir, em que $\overline{x}$ e $\overline{y}$ são, respectivamente, as médias amostrais das variáveis x e y .

i	Tempo y_i	Distância x_i	$x_i - \overline{x}$	$y_i - \overline{y}$	$(x_i - \overline{x}).(y_i - \overline{y})$	$(x_i - \overline{x})^2$
1	10	5	-4	-7	28	16
2	20	5	-4	3	-12	16
3	15	10	1	-2	-2	1
4	10	10	1	-7	-7	1
5	30	15	6	13	78	36
Média	17	9				

Com base nos dados dessa tabela, julgue o próximo item.

Pelo modelo de regressão linear simples, a equação que expressa o relacionamento ajustado entre a variável em função de x e $\hat{y}_i = \frac{85}{70} x_i + \alpha$, em que α é uma constante.

4. (CESPE/PETROBRAS/2022)

Equação 1: $y_i = a + bXi + e$

Equação 2: $y_i = a + b1X_i + b2X_2 + b3y_i3 + e$

Com base nos modelos de regressão linear simples (equação 1) e de regressão linear múltipla (equação 2), julgue o item a seguir.

O coeficiente b da equação 1 é o resultado da correlação entre os valores amostrais de X e Y, dividida pela variância de X .

5. (CESPE/TELEBRAS/2022) O quadro a seguir mostra as estimativas de mínimos quadrados ordinários dos coeficientes de um modelo de regressão linear simples na forma $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, em que $i \in \{1, \cdots, 6\}$ e ϵ_i representa o erro aleatório com média zero e variância σ^2.

Coeficiente	Estimativa	Erro Padrão	Razão t
β_0	0,9	0,10	9
β_1	0,2	0,05	4

Considerando essas informações e sabendo que $\sigma^2 = 0,01$, julgue o item seguinte.

$S_{xx} = \sum_{i=1}^{6}(x_i - \bar{x})^2 = 4$ em que $\bar{x} = \sum_{i=1}^{6} x_i/6$.

6. (CESPE/TELEBRAS/2022) O quadro a seguir mostra as estimativas de mínimos quadrados ordinários dos coeficientes de um modelo de regressão linear simples na forma $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, em que $i \in \{1, \cdots, 6\}$ e ϵ_i representa o erro aleatório com média zero e variância σ^2.

Coeficiente	Estimativa	Erro Padrão	Razão t
β_0	0,9	0,10	9
β_1	0,2	0,05	4

Considerando essas informações e sabendo que $\sigma^2 = 0,01$, julgue o item seguinte.

A covariância entre a variável resposta (y) e a variável explicativa (x) é igual ou superior a 0,2.

7. (CESPE/TCE-SC/2022) Em artigo publicado em 2004 no Journal of Political Economy, E. Miguel, S. Satyanath e E. Sergenti mostraram o efeito que o crescimento econômico pode ter na ocorrência de conflitos civis, com dados de 41 países africanos, no período de 1981 até 1999. Em certo estágio da pesquisa, para verificar a possibilidade de usar dados sobre precipitação pluviométrica como variável instrumental, foi feita uma regressão entre o crescimento de tais precipitações (variável explicativa) e uma variável resposta que representa um indicador para a ocorrência de conflito: quanto maior for esse indicador, maior a possibilidade de conflitos no ano t no país i. Os resultados do modelo ajustado pelo método de mínimos quadrados ordinários se encontram na tabela a seguir.

Variável Explicativa	Variável Dependente

	Conflito civil (mínimo de 25 mortos)	Conflito civil (mínimo de 1000 mortos)
Crescimento na precipitação em *t*	–0,024 (0,043)	–0,062 (0,030)
Crescimento na precipitação em *t–1*	–0,122 (0,052)	–0,069 (0,032)
Efeitos fixos	sim	sim
R^2	0,71	0,70
Observações	743	743

Internet: <https://doi.org/10.1086/421174> (com adaptações).

Os números entre parênteses na tabela apresentada indicam o erro padrão da estimativa dos coeficientes respectivos. Considere os valores críticos t_α da variável t de Student, com significância α para os graus de liberdades adequados aos dados apresentados, como sendo $t_{10\%} = 1,65$, $t_{5\%} = 1,96$ e $t_{1\%} = 2,58$. Considerando as informações precedentes, julgue o próximo item.

Os resultados mostram que um aumento na precipitação pluviométrica no ano anterior resulta no aumento na ocorrência de conflito civil, nas duas regressões.

8. (FGV/EPE/2022) Considere o modelo de regressão linear simples, a seguir.

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \qquad i = 1, 2, \cdots, n$$

Para uma amostra de 20 observações, foram obtidos os seguintes resultados:

$$\sum_{i=1}^{20} x_i = 60, \sum_{i=1}^{20} y_i = 90, \sum_{i=1}^{20} {x_i}^2 = 300, \sum_{i=1}^{20} x_i y_i = 510$$

Os estimadores de mínimos quadrados do modelo são, respectivamente,

a) –1,5 e 0,5.

b) –1,5 e 2.

c) –4,5 e 3.

d) 0,5 e 0,5.

e) 0,5 e 2.

9. (CESPE/SEFAZ RR/2021) A tabela a seguir apresenta uma amostra aleatória simples formada por 5 pares de valores (X_i, Y_i), em que $i = 1, 2, \cdots, 5$, X_i é uma variável explicativa e Y_i é uma variável dependente.

i	1	2	3	4	5
X_i	0	1	2	3	4
Y_i	0,5	2,0	2,5	5,0	3,5

Considere o modelo de regressão linear simples na forma $Y_i = bX_i + \epsilon_i$, no qual ϵ representa um erro aleatório normal com média zero e variância σ^2 e b é o coeficiente do modelo.

Com base nos dados da tabela e nas informações apresentadas, é correto afirmar que o valor da estimativa de mínimos quadrados ordinários do coeficiente b é igual a

a) 0,75.

b) 0,9.

c) 1,2.

d) 1,35.

e) 1,45.

10. (CESPE/BANESE/2021)

	X	Y
Média	**5**	**10**
Desvio Padrão	**2**	**2**

Com base nas informações apresentadas na tabela precedente e considerando que a covariância entre as variáveis X e Y seja igual a 3, julgue o item que se segue.

O coeficiente de determinação (ou de explicação) da reta de regressão linear da variável X em função da variável Y é igual ou superior a 0,60.

11. (CESPE/Pref. Aracaju/2021) Um modelo de regressão linear simples na forma $y = ax + b + \epsilon$, no qual ϵ representa o erro aleatório com média nula e variância constante, foi ajustado para um conjunto de dados no qual as médias aritméticas das variáveis y e x são, respectivamente, $\bar{y} = 10$ e $\bar{x} = 5$. Pelo método dos mínimos quadrados ordinários, se a estimativa do intercepto

(coeficiente b) for igual a 20, então a estimativa do coeficiente angular a proporcionada por esse mesmo método deverá ser igual a

a) −2.

b) 2.

c) −1.

d) 0.

e) 1.

12. (CESPE/BANESE/2021) Considere que uma tendência linear na forma $\hat{y} = 4x + 2$ tenha sido obtida com base no método dos mínimos quadrados ordinários. Acerca dessa tendência, sabe-se ainda que o desvio padrão da variável y foi igual a 8; que o desvio padrão da variável x foi igual a 1; e que a média aritmética da variável x foi igual a 2. Com base nessas informações, julgue o item subsequente, relativo a essa tendência linear.

A média aritmética da variável y foi igual a 8.

13. (CESPE/BANESE/2021) Considere que uma tendência linear na forma $\hat{y} = 4x + 2$ tenha sido obtida com base no método dos mínimos quadrados ordinários. Acerca dessa tendência, sabe-se ainda que o desvio padrão da variável y foi igual a 8; que o desvio padrão da variável x foi igual a 1; e que a média aritmética da variável x foi igual a 2. Com base nessas informações, julgue o item subsequente, relativo a essa tendência linear.

A covariância entre as variáveis x e y foi superior a 2.

14. (CESPE/PF/2021) Um estudo objetivou avaliar a evolução do número mensal Y de milhares de ocorrências de certo tipo de crime em determinado ano. Com base no método dos mínimos quadrados ordinários, esse estudo apresentou um modelo de regressão linear simples da forma

$$\overline{Y} = 5 - 0,1 \times T,$$

em que $\overline{Y}$ representa a reta ajustada em função da variável regressora T, tal que $1 \leq T \leq 12$.

Os erros padrão das estimativas dos coeficientes desse modelo, as razões t e seus respectivos p-valores encontram-se na tabela a seguir.

	Erro Padrão	Razão t	p-valor
Intercepto	0,584	8,547	0,00
Coeficiente Angular	0,064	1,563	0,15

Os desvios padrão amostrais das variáveis y e t foram, respectivamente, 1 e 3,6.

Com base nessas informações, julgue o item a seguir.

Se a média amostral da variável t for igual a 6,5, então a média amostral da variável Y será igual a 4,35 mil ocorrências.

15. (CESPE/MJ-SP/2021) A tabela de análise de variância a seguir se refere a um modelo de regressão linear simples na forma $y = ax + b + \epsilon$, na qual $\epsilon \sim N(0, \sigma^2)$. Os resultados da tabela foram obtidos com base em uma amostra aleatória simples n de pares de observações independentes (x,y).

Fonte de Variação	Graus de Liberdade	Soma de Quadrados
Regressão	1	82
Resíduos	8	8
Total	9	90

Com base nessas informações, julgue o item subsequente.

Se as médias amostrais das variáveis x e y forem iguais a zero, então o estimador de mínimos quadrados ordinários de b será igual a zero.

16. (CESPE/BANESE/2021)

	X	Y
Média	**5**	**10**
Desvio Padrão	**2**	**2**

Com base nas informações apresentadas na tabela precedente e considerando que a covariância entre as variáveis X e Y seja igual a 3, julgue o item que se segue.

A reta de regressão linear da variável Y em função da variável X, obtida pelo método de mínimos quadrados ordinários, pode ser escrita como $Y = 0{,}75X + 6{,}25$.

17. (CESPE/PG DF/2021) O coeficiente de correlação linear de Pearson entre duas variáveis aleatórias discretas X e Y definidas sobre um mesmo espaço amostral é dado por

$$CORR(X,Y) = \frac{n(\sum_{i=1}^{n} x_i y_i) - (\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{\sqrt{n(\sum_{i=1}^{n} x_i^2) - (\sum_{i=1}^{n} x_i)^2}\sqrt{n(\sum_{i=1}^{n} y_i^2) - (\sum_{i=1}^{n} y_i)^2}}$$

Já na reta de melhor ajuste $Y = aX + b$, determinada pelo método dos mínimos quadrados, os coeficientes são dados por

$$a = \frac{n(\sum_{i=1}^{n} x_i y_i) - (\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n(\sum_{i=1}^{n} x_i^2) - (\sum_{i=1}^{n} x_i)^2}$$

$$b = \frac{\sum_{i=1}^{n} y_i - a\sum_{i=1}^{n} x_i}{n}$$

Uma forma de avaliar a precisão do modelo consiste em comparar o estimador não viesado da variância residual, obtidos das diferenças entre os valores observados e os previstos pelo modelo, $\widehat{S}_e = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2$, com o estimador não viesado da variância dos valores observados, $S_e = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2$.

Tal avaliação também pode ser realizada pela aferição na redução da soma dos quadrados dos resíduos na passagem do modelo simples, em que as observações são aproximadas por sua média, para o modelo de regressão linear, redução esta que é dada por $\sum_{i=1}^{n}(\widehat{y}_i - \overline{y})^2 = a^2\sum_{i=1}^{n}(x_i - \overline{x})^2$.

Com base nessas informações, julgue o item seguinte.

Se, para certo conjunto de dados, o coeficiente angular da reta de melhor ajuste obtida pelo método dos mínimos quadrados for nulo, então o coeficiente de correlação de Pearson entre essas variáveis também será nulo.

18. (CESPE/PG DF/2021) O coeficiente de correlação linear de Pearson entre duas variáveis aleatórias discretas X e Y definidas sobre um mesmo espaço amostral é dado por

$$CORR(X,Y) = \frac{n(\sum_{i=1}^{n} x_i y_i) - (\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{\sqrt{n(\sum_{i=1}^{n} x_i^2) - (\sum_{i=1}^{n} x_i)^2}\sqrt{n(\sum_{i=1}^{n} y_i^2) - (\sum_{i=1}^{n} y_i)^2}}$$

Já na reta de melhor ajuste $Y = aX + b$, determinada pelo método dos mínimos quadrados, os coeficientes são dados por

$$a = \frac{n(\sum_{i=1}^{n} x_i y_i) - (\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n(\sum_{i=1}^{n} x_i^2) - (\sum_{i=1}^{n} x_i)^2}$$

$$b = \frac{\sum_{i=1}^{n} y_i - a\sum_{i=1}^{n} x_i}{n}$$

Uma forma de avaliar a precisão do modelo consiste em comparar o estimador não viesado da variância residual, obtidos das diferenças entre os valores observados e os previstos pelo

modelo, $\widehat{S}_e = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$, com o estimador não viesado da variância dos valores observados, $S_e = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$.

Tal avaliação também pode ser realizada pela aferição na redução da soma dos quadrados dos resíduos na passagem do modelo simples, em que as observações são aproximadas por sua média, para o modelo de regressão linear, redução esta que é dada por $\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 = a^2\sum_{i=1}^{n}(x_i - \bar{x})^2$.

Com base nessas informações, julgue o item seguinte.

Quanto mais próximo de -1 estiver o coeficiente de correlação de Pearson entre duas variáveis, menos indicada será a aplicação do método de mínimos quadrados para obter a relação entre as variáveis.

19. (FCC/ALAP/2020) Em uma empresa de determinado ramo de atividade, utilizando o método de regressão linear, obteve-se a equação de tendência (T) da série temporal abaixo.

Os dados apresentam 10 observações da série temporal Y, que representa o faturamento de uma empresa, em milhões de reais. Supõe-se que essa série é composta apenas de uma tendência T e um ruído branco de média zero e variância constante.

t	1	2	3	4	5	6	7	8	9	10
y_t	6	5	6	8	8	7	8	10	10	11

Observação: t representa o ano e yt o faturamento da empresa no ano t, em milhões de reais.

Dados:

$$\sum_{t=1}^{10} t = 55\,,\ \sum_{t=1}^{10} t^2 = 385\,,\ \sum_{t=1}^{10} y_t = 79\,,\ \sum_{t=1}^{10} t \times y_t = 484$$

A tendência apresenta a forma $T = a + bt$, em que a e b foram obtidos usando o método dos mínimos quadrados. Considerando a equação obtida, tem-se que o acréscimo no faturamento do ano t, com t > 1, para o ano (t + 1) é, em milhões de reais, de

a) 1,2.

b) 1,5.

c) 0,6.

d) 2,4.

e) 1,8.

20. (CESPE/TJ-AM/2019) 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.

21. (CESPE/TJ-AM/2019) No modelo de regressão linear simples na forma matricial $Y = X\beta + \epsilon$, Y denota o vetor de respostas, X representa a matriz de delineamento (ou matriz de desenho), β é o vetor de coeficientes do modelo e ϵ é o vetor de erros aleatórios independentes e identicamente distribuídos. Tem-se também que $X'Y = \begin{pmatrix} 2 \\ 10 \end{pmatrix}$ e $(X'X)^{-1} = \begin{pmatrix} 1 & 0,5 \\ 0,5 & 1 \end{pmatrix}$ em que X' é a matriz transposta de X.

Com base nessas informações, julgue o próximo item, considerando que a variância do erro aleatório é $\sigma_\epsilon^2 = 4$

O referido modelo possui uma única variável regressora.

22. (CESPE/TJ-AM/2019) 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 desvio padrão de x é superior a 1.

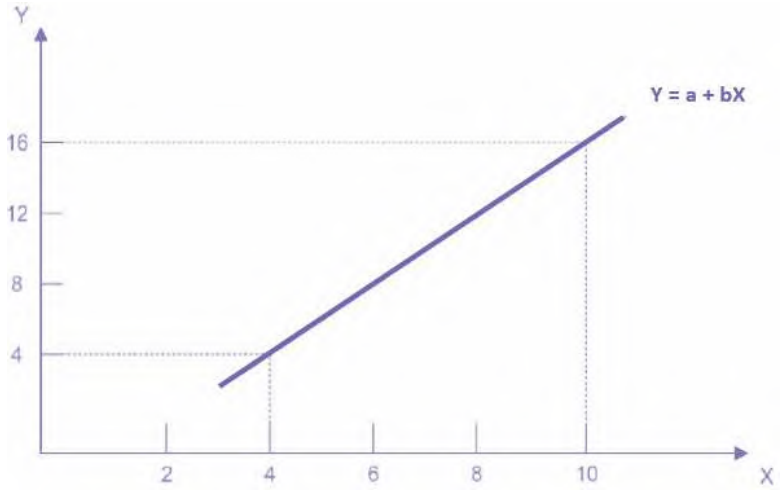
23. (CESPE/TJ-AM/2019) Um estudo considerou um modelo de regressão linear simples na forma $y = 0,8x + b + \epsilon$, em que y é a variável dependente, x representa a variável explicativa do modelo, o coeficiente b denomina-se intercepto e ϵ é um erro aleatório que possui média nula e desvio padrão σ. 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 intercepto do referido modelo é igual ou superior a 0,8.

24. (FCC/SEFAZ-BA/2019) Em uma determinada indústria, foi efetuada uma pesquisa a respeito da possível relação entre o número de horas trabalhadas (X), com $X \geq 2$, e as quantidades produzidas de um produto (Y). Com base em 10 pares de observações (Xi, Yi) e considerando o gráfico de dispersão correspondente, optou-se por utilizar o modelo linear $Y_i = \alpha + \beta X_i + \varepsilon_i$, com i representando a i-ésima observação, ou seja, $i = 1, 2, 3, \ldots 10$. Os parâmetros α e β são desconhecidos e as suas estimativas (a e b, respectivamente) foram obtidas pelo método dos

mínimos quadrados. Observação: ε_i é o erro aleatório com as respectivas hipóteses do modelo de regressão linear simples. Considere o gráfico, abaixo, construído utilizando os valores encontrados para as estimativas de α e β.



Dados:

$$\sum_{i=1}^{10} X_i = 120$$

A previsão da quantidade produzida será igual ao dobro da média verificada das 10 observações Y_i quando o número de horas trabalhadas for igual a

a) 20.

b) 24.

c) 22.

d) 18.

e) 12.

25. (FCC/BANRISUL/2019) Utilizando o método dos mínimos quadrados, obteve-se a equação de tendência $\widehat{T}_t = 15 + 2,5t$, sendo $t = 1, 2, 3, \ldots$, com base nos lucros anuais de uma empresa, em milhões de reais, nos últimos 10 anos, em que $t = 1$ representa 2009, $t = 2$ representa 2010 e assim por diante. Por meio dessa equação, obtém-se que a previsão do lucro anual dessa empresa, no valor de 55 milhões de reais, será para o ano

a) 2021.

b) 2025.

c) 2024.

d) 2023.

e) 2022.

26. (VUNESP/Pref. Mogi das Cruzes/2019) Sejam S o valor do salário, em R$ 1.000,00, e t o respectivo tempo de serviço, em anos, de 20 empregados de uma empresa. Optou-se, com o objetivo de previsão do salário de um determinado empregado em função do seu tempo de serviço, por utilizar a relação linear $S_i = \alpha + \beta t_i + \varepsilon_i$, com i representando a i-ésima observação, α e β são parâmetros desconhecidos e εi é o erro aleatório com as respectivas hipóteses da regressão linear simples. Utilizando o método dos mínimos quadrados, com base nas 20 observações correspondentes dos 20 empregados, obtiveram-se as estimativas de α e β (a e b, respectivamente). O valor encontrado para b foi de 1,8 e as médias dos salários dos 20 empregados e dos correspondentes tempos de serviço apresentam os valores de R$ 2.800,00 e 2 anos, respectivamente.

A previsão de salário para um empregado que tenha 5 anos de serviço é de

a) R$ 6.800,00

b) R$ 7.500,00

c) R$ 8.200,00

d) R$ 8.400,00

e) R$ 9.000,00

27. (VUNESP/MPE-SP/2019). 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.

28. (CESPE/STM/2018). Considerando que $\widehat{Y}$ seja uma variável resposta ajustada por um modelo de regressão em função de uma variável explicativa X, que $x_1, \ldots, x_n$ representem as réplicas de X e que $\widehat{\alpha}$ e $\widehat{\beta}$ sejam as estimativas dos parâmetros do modelo, julgue o item a seguir.

Em um modelo linear $\hat{Y} = \hat{\alpha} + \hat{\beta}X$, a hipótese de homoscedastiscidade significa que a variância dos erros deve ser constante, e o valor esperado dos erros deve ser zero.

29. (CESPE/ABIN/2018) Ao avaliar o efeito das variações de uma grandeza X sobre outra grandeza Y por meio de uma regressão linear da forma $\widehat{Y} = \widehat{\alpha} + \widehat{\beta}X$, 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):

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

A partir desses resultados, julgue o item a seguir.

30. (CESPE/EBSERH/2018) Deseja-se estimar o total de carboidratos existentes em um lote de 500.000 g de macarrão integral. Para esse fim, foi retirada uma amostra aleatória simples constituída por 5 pequenas porções desse lote, conforme a tabela seguinte, que mostra a quantidade x amostrada, em gramas, e a quantidade de carboidratos encontrada, y, em gramas.

Amostra	*X*	*Y*
1	100	60
2	80	40
3	90	40
4	120	50
5	110	60

Com base nas informações e na tabela apresentadas, julgue o item a seguir.

Considerando-se o modelo de regressão linear na forma $y = ax + \varepsilon$, em que ε denota o erro aleatório com média nula e variância V, e a representa o coeficiente angular, a estimativa de mínimos quadrados ordinários do coeficiente a é igual ou superior a 0,5.

31. (CESPE/PF/2018) O intervalo de tempo entre a morte de uma vítima até que ela seja encontrada (y em horas) denomina-se intervalo post mortem. Um grupo de pesquisadores mostrou que esse tempo se relaciona com a concentração molar de potássio encontrada na vítima (x, em mmol/dm3). Esses pesquisadores consideraram um modelo de regressão linear simples na forma $y = ax + b + \varepsilon$, em que a representa o coeficiente angular, b denomina-se intercepto, e ε denota um erro aleatório que segue distribuição normal com média zero e desvio padrão igual a 4.

As estimativas dos coeficientes a e b, obtidas pelo método dos mínimos quadrados ordinários foram, respectivamente, iguais a 2,5 e 10. O tamanho da amostra para a obtenção desses resultados foi n = 101. A média amostral e o desvio padrão amostral da variável x foram, respectivamente, iguais a 9 mmol/dm3 e 1,6 mmol/dm3 e o desvio padrão da variável y foi igual a 5 horas.

A respeito dessa situação hipotética, julgue o item a seguir.

A média amostral da variável resposta y foi superior a 30 horas.

32. (CESPE/PF/2018) O intervalo de tempo entre a morte de uma vítima até que ela seja encontrada (y em horas) denomina-se intervalo post mortem. Um grupo de pesquisadores mostrou que esse tempo se relaciona com a concentração molar de potássio encontrada na vítima (x, em mmol/dm3). Esses pesquisadores consideraram um modelo de regressão linear simples na forma $y = ax + b + \varepsilon$, em que a representa o coeficiente angular, b denomina-se intercepto, e ε denota um erro aleatório que segue distribuição normal com média zero e desvio padrão igual a 4.

As estimativas dos coeficientes a e b, obtidas pelo método dos mínimos quadrados ordinários foram, respectivamente, iguais a 2,5 e 10. O tamanho da amostra para a obtenção desses resultados foi n = 101. A média amostral e o desvio padrão amostral da variável x foram, respectivamente, iguais a 9 mmol/dm3 e 1,6 mmol/dm3 e o desvio padrão da variável y foi igual a 5 horas.

A respeito dessa situação hipotética, julgue o item a seguir.

De acordo com o modelo ajustado, caso a concentração molar de potássio encontrada em uma vítima seja igual a 2 mmol/dm3, o valor predito correspondente do intervalo post mortem será igual a 15 horas.

33. (CESPE/STM/2018). Em um modelo de regressão linear simples na forma $yi = a + bx_i + \varepsilon_i$, em que a e b são constantes reais não nulas, y_i representa a resposta da i-ésima observação a um estímulo x_i e ε_i éo erro aleatório correspondente, para $i = 1, \ldots, n$, considere que $\sum_i (x_i - \overline{x})^2 = 10$, em que $\overline{x} = (x_1 + \ldots + x_n)/n$, e que o desvio padrão de cada ε_i seja igual a 10, para $i = 1, \ldots, n$.

A respeito dessa situação hipotética, julgue o item que se segue.

Se $\hat{b}$ representar o estimador de mínimos quadrados ordinários do coeficiente b, então $var[\hat{b}] = 10$.

34. (FCC/Pref. São Luís/2018) 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

35. (FCC/SEF-SC/2018) A tabela a seguir indica o valor y do salário, em número de salários mínimos (SM) e os respectivos tempos de serviço, em anos, x, de 5 funcionários de uma empresa:

x (anos)	2	3	5	3	2
y (SM)	3	4	7	4	2

Suponha que valha a relação: $y_i = \alpha + \beta x_i + \varepsilon_i$, em que i representa a i-ésima observação, α e β são parâmetros desconhecidos e ε_i é o erro aleatório com as hipóteses para a regressão linear simples. Se as estimativas de α e β forem obtidas pelo método de mínimos quadrados por meio dessas 5 observações, a previsão de salário para um funcionário com 4 anos de serviço será, em SM, igual a

a) 6,1

b) 5,2

c) 6,0

d) 5,5

e) 5,8

36. (FGV/ALERO/2018) Se b_0 e b_1 são as estimativas por mínimos quadrados de β_0 e β_1, respectivamente, então seus valores são dados por

a) $b_1 = \sum_{i=1}^{n}(Y_i - \bar{Y})^2 / \sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})\,; b_o = \bar{Y} - b_1\bar{X}$

b) $b_1 = \sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}) / \sum_{i=1}^{n}(Y_i - \bar{Y})^2\,; b_o = \bar{Y} + b_1\bar{X}$

c) $b_1 = \sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}) / \sum_{i=1}^{n}(Y_i - \bar{Y})^2\,; b_o = \bar{Y} - b_1\bar{X}$

d) $b_1 = \sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}) / \sum_{i=1}^{n}(X_i - \bar{X})^2\,; b_o = \bar{Y} - b_1\bar{X}$

e) $b_1 = \sum_{i=1}^{n}(X_i - \bar{X})^2 / \sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})\,; b_o = \bar{Y} + b_1\bar{X}$

37. (FGV/ALERO/2018) Se $\hat{Y} = b_0 + b_1 X$ é a reta ajustada pela regressão e se $e_i = Y_i - \hat{Y}$ é o resíduo da observação $i, i - 1, \cdots, n$, avalie as afirmativas a seguir.

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

II. $\sum_{i=1}^{n} Y_i = \sum_{i=1}^{n} \hat{Y}$.

III. O ponto $(\bar{X}, \bar{Y})$ pertence à reta ajustada.

Está correto o que se afirma em

a) I, apenas.

b) I e II, apenas.

c) I e III, apenas.

d) II e III, apenas.

e) I, II e III.

38. (CESPE/TCE-PE/2017) Um estudo de acompanhamento ambiental considerou, para $j = 1, 2, \ldots, 26$, um modelo de regressão linear simples na forma: $y_j = a + bx_j + e_j$, em que a e b são constantes reais, y_j representa a variável resposta referente ao j-ésimo elemento da amostra, x_j é a variável regressora correspondente, e e_j denota o erro aleatório que segue distribuição normal com média nula e variância V. Aplicando-se, nesse estudo, o método dos mínimos quadrados ordinários, obteve-se a reta ajustada $\hat{y}_j = 1 + 2x_j, para\ j = 1, 2, \ldots, 26$

Considerando que a estimativa da variância V seja igual a 6 e que o coeficiente de explicação do modelo (R quadrado) seja igual a 0,64, julgue o item.

Se $\bar{x} = \frac{\sum_{j=1}^{26} x_j}{26}$ representar a média amostral da variável regressora e se $\bar{y} = \frac{\sum_{j=1}^{26} y_j}{26}$ denotar a média amostral da variável resposta, com $\bar{x} > 0$ e $\bar{y} > 0$, então $\bar{x} < \bar{y}$.

39. (FGV/MPE-BA/2017) Com o objetivo de realizar uma projeção sobre a necessidade de novos servidores para o Ministério Público, foi elaborado um modelo de regressão associando o número de procedimentos em curso e a variável de interesse. A equação do modelo é:

$$NS_i = \alpha + \beta . PC_i + \varepsilon_i$$

onde NS é o número de novos servidores e PC a quantidade de procedimentos em curso.

Através de uma amostra representativa (n=20), em diversas unidades no MP, foram obtidas as seguintes estatísticas:

$$\sum NS^2 = 18000, \sum NS = 200, \sum PC = 800,$$

$$\sum (PC).(NS) = 12000\ e \sum PC^2 = 72000$$

Com base no modelo e nas estatísticas, é correto afirmar que:

a) ainda que o número de procedimentos não sofra incrementos, novos servidores serão necessários a cada período;

b) se o número de procedimentos sofrer um incremento de 40 unidades, serão necessários mais oito novos servidores;

c) as estimativas de MQO são $\hat{\alpha} = 6$ e $\hat{\beta} = 0{,}1$;

d) a correlação entre o volume de procedimentos e o número de novos servidores é 0,7, comprovando a qualidade do modelo;

e) sendo estimativa de β positiva, o número de funcionários do MP deverá crescer a uma taxa de 10% ao período.

40. (CESPE/TCE-PA/2016). Uma regressão linear simples é expressa por $Y = a + b \times X + e$, em que o termo e corresponde ao erro aleatório da regressão e os parâmetros a e b são desconhecidos e devem ser estimados a partir de uma amostra disponível. Assumindo que a variável X é não correlacionada com o erro e, julgue o item subsecutivo, nos quais os resíduos das amostras consideradas são IID, com distribuição normal, média zero e variância constante.

Para uma amostra de tamanho $n = 25$, em que a covariância amostral para o par de variáveis X e Y seja $Cov(X, Y) = 20{,}0$, a variância amostral para a variável Y seja $Var(Y) = 4{,}0$ e a variância amostral para a variável X seja $Var(X) = 5{,}0$, a estimativa via estimador de mínimos quadrados ordinários para o coeficiente b é igual a 5,0.

41. (CESPE/TCE-PA/2016) Uma regressão linear simples é expressa por $Y = a + b \times X + e$, em que o termo e corresponde ao erro aleatório da regressão e os parâmetros a e b são desconhecidos e devem ser estimados a partir de uma amostra disponível. Assumindo que a variável X é não correlacionada com o erro e, julgue o item subsecutivo, nos quais os resíduos das amostras consideradas são IID, com distribuição normal, média zero e variância constante.

Considere que as estimativas via método de mínimos quadrados ordinários para o parâmetro a seja igual a 2,5 e, para o parâmetro b, seja igual a 3,5. Nessa situação, assumindo que $X = 4{,}0$, o valor predito para Y será igual a 16,5, se for utilizada a reta de regressão estimada.

42. (CESPE/TCE-PA/2016). Uma regressão linear simples é expressa por $Y = a + b \times X + e$, em que o termo e corresponde ao erro aleatório da regressão e os parâmetros a e b são desconhecidos e devem ser estimados a partir de uma amostra disponível. Assumindo que a variável X é não correlacionada com o erro e, julgue o item subsecutivo, nos quais os resíduos das amostras consideradas são IID, com distribuição normal, média zero e variância constante.

A variável Y é denominada variável explicativa, e a variável X é denominada variável dependente.

43. (FGV/DPE-RJ/2014) Considere a equação de regressão $Y_i = \alpha + \beta . X_i + \epsilon_i$, onde Y e X são as variáveis explicada e explicativa, respectivamente, ϵ é o erro aleatório e α e β os parâmetros a estimar. São supostos válidos todos os pressupostos clássicos do Modelo de Regressão Linear Simples (MRLS). Além disso, para determinada amostra de pares (X, Y), foram calculadas as estatísticas $p(X, Y) = 0{,}8$, $\overline{X} = 6$, $\overline{Y} = 15$, $DP(Y) = 5$ e $DP(X) = 2$. Portanto, a partir do método de Mínimos Quadrados Ordinários os estimadores de α e β são

a) 2 e 3.

b) 3 e 2.

c) -9 e 4.

d) 4 e -9.

e) 6 e 1,5.

44. (FGV/SEN/2012) Um modelo probabilístico de primeira ordem, ou seja, de regressão linear pode ser representado da seguinte forma: $y = \beta_0 + \beta_1 x + \varepsilon$.

Com base nessa equação, avalie as afirmativas a seguir:

I. Neste modelo, pode-se sempre assumir que ε, o componente de erro aleatório, seja um ruído branco.

II. Uma vez que o valor esperado do erro, $E(\varepsilon)$, nem sempre será igual a zero, não é correto afirmar que o valor esperado de y, $E(y)$, será igual ao seu componente determinístico, $\beta_0 + \beta_1 x$.

III. Os símbolos gregos β_0 e β_1 são parâmetros populacionais que somente serão conhecidos se tiver acesso às medidas de toda a população de (x, y).

Assinale:

a) se apenas as afirmativas I e II forem verdadeiras.

b) se apenas as afirmativas I e III forem verdadeiras.

c) se apenas as afirmativas II e III forem verdadeiras.

d) se todas as afirmativas forem verdadeiras.

e) se nenhuma afirmativa for verdadeira.

45. (FGV/SEFAZ RJ/2011) A tabela abaixo mostra os valores de duas variáveis, X e Y.

X	Y
4	4.5
4	5
3	5
2	5.5

Sabe-se que:

$$\sum X = 13$$

$$\sum Y = 20$$

$$\sum XY = 64$$

$$\sum X^2 = 45$$

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

O valor de "b" na regressão simples $Y = a + bX$ é

a) 11 /5.

b) -3 /8.

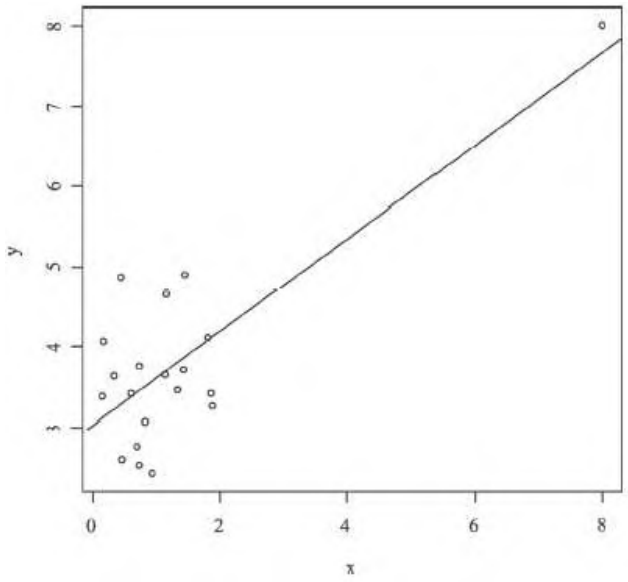
c) -4 /11.

d) -4 /17.

e) -11/65.

46. (FGV/SEN/2008) A figura a seguir representa o diagrama de dispersão de dez pontos (X_i, Y_i) e a reta de regressão ajustada pelo método de mínimos quadrados dada por $Y = 0,42 + 2,45X$.

Quanto ao ponto de coordenadas $X = 8$ e $Y = 8$, pode-se afirmar que ele:



a) é o ponto com maior desvio da reta de regressão.

b) é um ponto influente nessa regressão.

c) é um dado legítimo que indica a relação linear entre X e Y.

d) indica que o modelo é provavelmente heterocedástico.

e) é uma observação incorreta que deve ser eliminada da análise.

GABARITO

Regressão Linear Simples

1. LETRA A
2. LETRA A
3. CERTO
4. ERRADO
5. CERTO
6. ERRADO
7. ERRADO
8. LETRA B
9. LETRA C
10. ERRADO
11. LETRA A
12. ERRADO
13. CERTO
14. CERTO
15. CERTO
16. CERTO
17. CERTO
18. ERRADO
19. LETRA C
20. CERTO
21. CERTO
22. CERTO
23. ERRADO
24. LETRA C
25. LETRA C
26. LETRA C
27. LETRA C
28. ERRADO
29. CERTO
30. ERRADO
31. CERTO
32. CERTO
33. CERTO
34. LETRA D
35. LETRA D
36. LETRA D
37. LETRA E
38. CERTO
39. LETRA C
40. ERRADO
41. CERTO
42. ERRADO
43. LETRA B
44. LETRA B
45. LETRA C
46. LETRA B

Lista de Questões

Análise de Variância da Regressão

1. (CESPE/TELEBRAS/2022) A tabela ANOVA a seguir se refere ao ajuste de um modelo de regressão linear simples escrito como $y = a + bx + \epsilon$, cujos coeficientes foram estimados pelo método da máxima verossimilhança, com $\epsilon \sim N(0, \sigma^2)$. Os erros em torno da reta esperada são independentes e identicamente distribuídos.

Fonte de Variação	Graus de Liberdade	Soma de Quadrados	Quadrado Médio
Modelo	1	10	10
Erro	99	990	10
Total	**100**	**1.000**	**10**

Com base nessas informações, julgue o item a seguir.

O coeficiente de explicação do modelo é igual a 0,99.

2. (CESPE/TELEBRAS/2022) O quadro a seguir mostra as estimativas de mínimos quadrados ordinários dos coeficientes de um modelo de regressão linear simples na forma $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, em que $i \in \{1, \cdots, 6\}$ e ϵ_i representa o erro aleatório com média zero e variância σ^2.

Coeficiente	Estimativa	Erro Padrão	Razão t
β_0	0,9	0,10	9
β_1	0,2	0,05	4

Considerando essas informações e sabendo que $\sigma^2 = 0,01$, julgue o item seguinte.

$SQ_{RESÍDUOS} = \sum_{i=1}^{6}(\hat{y}_i - \bar{y})^2 = 0,04$, em que $\hat{y}_i = 0,9 + 0,2x_i$.

3. (CESPE/TELEBRAS/2022) O quadro a seguir mostra as estimativas de mínimos quadrados ordinários dos coeficientes de um modelo de regressão linear simples na forma $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, em que $i \in \{1, \cdots, 6\}$ e ϵ_i representa o erro aleatório com média zero e variância σ^2.

Coeficiente	Estimativa	Erro Padrão	Razão t
β_0	0,9	0,10	9
β_1	0,2	0,05	4

Considerando essas informações e sabendo que $\sigma^2 = 0,01$, julgue o item seguinte.

O coeficiente de determinação do modelo (R^2) é igual a 0,8.

4. (CESPE/TELEBRAS/2022) O quadro a seguir mostra as estimativas de mínimos quadrados ordinários dos coeficientes de um modelo de regressão linear simples na forma $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, em que $i \in \{1, \cdots, 6\}$ e ϵ_i representa o erro aleatório com média zero e variância σ^2.

Coeficiente	Estimativa	Erro Padrão	Razão t
β_0	0,9	0,10	9
β_1	0,2	0,05	4

Considerando essas informações e sabendo que $\sigma^2 = 0,01$, julgue o item seguinte.

$$SQ_{TOTAL} = \sum_{i=1}^{6} (y_i - \bar{y})^2 = 0,2$$

5. (CESPE/TCE-SC/2022) Em artigo publicado em 2004 no Journal of Political Economy, E. Miguel, S. Satyanath e E. Sergenti mostraram o efeito que o crescimento econômico pode ter na ocorrência de conflitos civis, com dados de 41 países africanos, no período de 1981 até 1999. Em certo estágio da pesquisa, para verificar a possibilidade de usar dados sobre precipitação pluviométrica como variável instrumental, foi feita uma regressão entre o crescimento de tais precipitações (variável explicativa) e uma variável resposta que representa um indicador para a ocorrência de conflito: quanto maior for esse indicador, maior a possibilidade de conflitos no ano t no país i. Os resultados do modelo ajustado pelo método de mínimos quadrados ordinários se encontram na tabela a seguir.

Variável Explicativa	Variável Dependente	
	Conflito civil (mínimo de 25 mortos)	Conflito civil (mínimo de 1000 mortos)

Crescimento na precipitação em *t*	–0,024 (0,043)	–0,062 (0,030)
Crescimento na precipitação em *t–1*	–0,122 (0,052)	–0,069 (0,032)
Efeitos fixos	sim	sim
R^2	0,71	0,70
Observações	743	743

Internet: <https://doi.org/10.1086/421174> (com adaptações).

Os números entre parênteses na tabela apresentada indicam o erro padrão da estimativa dos coeficientes respectivos. Considere os valores críticos t_α da variável t de Student, com significância α para os graus de liberdades adequados aos dados apresentados, como sendo $t_{10\%} = 1,65$, $t_{5\%} = 1,96$ e $t_{1\%} = 2,58$. Considerando as informações precedentes, julgue o próximo item.

As variáveis explicativas usadas explicam em torno de 71% das variações na ocorrência de conflito civil com um mínimo de 25 mortos nos países pesquisados, no período analisado.

6. (CESPE/TELEBRAS/2022) A tabela ANOVA a seguir se refere ao ajuste de um modelo de regressão linear simples escrito como $y = a + bx + \epsilon$, cujos coeficientes foram estimados pelo método da máxima verossimilhança, com $\epsilon \sim N(0, \sigma^2)$. Os erros em torno da reta esperada são independentes e identicamente distribuídos.

Fonte de Variação	Graus de Liberdade	Soma de Quadrados	Quadrado Médio
Modelo	1	10	10
Erro	99	990	10
Total	**100**	**1.000**	**10**

Com base nessas informações, julgue o item a seguir.

A variância amostral da variável dependente é inferior a 12.

7. (CESPE/TELEBRAS/2022) A tabela ANOVA a seguir se refere ao ajuste de um modelo de regressão linear simples escrito como $y = a + bx + \epsilon$, cujos coeficientes foram estimados pelo

método da máxima verossimilhança, com $\epsilon \sim N(0, \sigma^2)$. Os erros em torno da reta esperada são independentes e identicamente distribuídos.

Fonte de Variação	Graus de Liberdade	Soma de Quadrados	Quadrado Médio
Modelo	1	10	10
Erro	99	990	10
Total	**100**	**1.000**	**10**

Com base nessas informações, julgue o item a seguir.

O R^2 ajustado é maior ou igual a 0,05.

8. (CESPE/TELEBRAS/2022) A tabela ANOVA a seguir se refere ao ajuste de um modelo de regressão linear simples escrito como $y = a + bx + \epsilon$, cujos coeficientes foram estimados pelo método da máxima verossimilhança, com $\epsilon \sim N(0, \sigma^2)$. Os erros em torno da reta esperada são independentes e identicamente distribuídos.

Fonte de Variação	Graus de Liberdade	Soma de Quadrados	Quadrado Médio
Modelo	1	10	10
Erro	99	990	10
Total	**100**	**1.000**	**10**

Com base nessas informações, julgue o item a seguir.

$\sigma^2 = 10$.

9. (CESPE/TELEBRAS/2022) A tabela ANOVA a seguir se refere ao ajuste de um modelo de regressão linear simples escrito como $y = a + bx + \epsilon$, cujos coeficientes foram estimados pelo método da máxima verossimilhança, com $\epsilon \sim N(0, \sigma^2)$. Os erros em torno da reta esperada são independentes e identicamente distribuídos.

Fonte de Variação	Graus de Liberdade	Soma de Quadrados	Quadrado Médio
Modelo	1	10	10

Erro	99	990	10
Total	**100**	**1.000**	**10**

Com base nessas informações, julgue o item a seguir.

Para se testar a hipótese nula $H_0: y = a + \epsilon$ contra a hipótese alternativa $H_1: y = a + bx + \epsilon$, a estatística do teste F proporcionada pela tabela ANOVA é igual ou superior a 2.

10. (CESPE/TELEBRAS/2022) Considere um modelo de regressão linear simples na forma $Y = aX + b + \epsilon$, em que ϵ representa o erro aleatório com média zero e desvio padrão σ, e a variável regressora X é binária. A média amostral e o desvio padrão amostral da variável explicativa Y foram, respectivamente, iguais a 10 e 4. Já para a variável regressora X, encontra-se a distribuição de frequências absolutas mostrada no quadro a seguir. Finalmente, sabe-se que a correlação linear entre Y e X é igual a 0,9.

X	Frequência Absoluta
0	55
1	45
Total	**100**

Com base nessas informações, com respeito à reta ajustada pelo método dos mínimos quadrados ordinários, julgue o item subsequente.

O coeficiente de determinação do modelo é igual ou superior a 0,9.

11. (FGV/EPE/2022) A respeito do coeficiente de determinação de uma regressão linear, avalie as afirmativas a seguir.

I. Mede a porcentagem da variância total que é explicada pela regressão.

II. É um número real entre 0 e 1.

III. É igual ao quadrado do coeficiente de correlação amostral.

Está correto o que se afirma em

a) I, apenas.

b) I e II, apenas.

c) I e III, apenas.

d) II e III, apenas.

e) I, II e III.

12. (FGV/MPE SC/2022) É possível que o comportamento das bolsas de valores em determinado mês prediga o seu comportamento o ano inteiro. Considere que a variável explicativa X seja a variação percentual do índice da bolsa em janeiro e que a variável de resposta Y seja a variação desse índice para o ano inteiro. O cálculo feito com dados do período de 5 anos teve como resultados:

$$\bar{x} = 1,75\% \; \bar{y} = 9,07\%$$

$$S_x = 5,36\% \; S_y = 15,35\%$$

$$r = 0,59$$

O percentual de variação observado nas alterações anuais do índice que é explicado pela relação linear com a alteração de janeiro é:

a) 2,86%;

b) 5,18%;

c) 34,81%;

d) 35,50%;

e) 59%.

13. (CESPE/MJ-SP/2021) A tabela de análise de variância a seguir se refere a um modelo de regressão linear simples na forma $y = ax + b + \epsilon$, na qual $\epsilon \sim N(0, \sigma^2)$. Os resultados da tabela foram obtidos com base em uma amostra aleatória simples n de pares de observações independentes (x,y).

Fonte de Variação	Graus de Liberdade	Soma de Quadrados
Regressão	1	82
Resíduos	8	8
Total	**9**	**90**

Com base nessas informações, julgue o item subsequente.

O coeficiente de explicação ajustado (R^2 ajustado) é igual a 0,90.

14. (CESPE/MJ-SP/2021) A tabela de análise de variância a seguir se refere a um modelo de regressão linear simples na forma $y = ax + b + \epsilon$, na qual $\epsilon \sim N(0, \sigma^2)$. Os resultados da tabela foram obtidos com base em uma amostra aleatória simples n de pares de observações independentes (x,y).

Fonte de Variação	Graus de Liberdade	Soma de Quadrados
Regressão	1	82
Resíduos	8	8
Total	9	90

Com base nessas informações, julgue o item subsequente.

O quadrado da razão t do teste de hipóteses $H_0: a = 0$ *versus* $H_1: a \neq 0$ é igual a 16.

15. (CESPE/ALECE/2021) Um modelo de regressão linear simples tem a forma $y = ax + b + \varepsilon$, em que y denota a variável resposta, x é a variável regressora, a e b são os coeficientes do modelo, e ε representa um erro aleatório com média 0 e variância σ^2. Com base em uma amostra aleatória simples de tamanho $n = 51$, pelo método dos mínimos quadrados ordinários, a estimativa da variância v foi igual 3. A variância amostral da variável y é 42.

Nesse modelo, o valor do coeficiente de determinação (R^2) é igual a

a) 0,07.

b) 0,21.

c) 0,93.

d) 0,42.

e) 0,79.

16. (CESPE/MJ-SP/2021) Acerca de planejamento de pesquisa estatística, julgue o item que se seguem.

Em um modelo estatístico, o erro total corresponde à soma dos desvios das observações em relação ao modelo.

17. (FGV/FunSaúde CE/2021)Numa regressão linear, as afirmativas a seguir, acerca do coeficiente de determinação, estão corretas, exceto uma. Assinale-a.

a) Mede a porcentagem da variação total da variável resposta que é explicada pela regressão.

b) É o quadrado do coeficiente de correlação estimado.

c) É um número entre 0 e 1.

d) Determina se as estimativas e predições dos coeficientes são tendenciosas.

e) Em geral, mas nem sempre, quanto maior seu valor, melhor o modelo se ajusta aos dados.

18. (VUNESP/EBSERH/2020) Numa regressão linear simples em que foi utilizada uma amostra com 52 observações, a soma dos quadrados totais é de 50 e a soma dos quadrados dos resíduos é de 20. O coeficiente de determinação e a estatística F dessa regressão são, respectivamente:

a) 0,6 e 75.

b) 0,6 e 12.

c) 0,8 e 1,5.

d) 0,8 e 12.

e) 0,8 e 75.

19. (CESPE/TJ-AM/2019) Um modelo de regressão linear foi ajustado para explicar os sintomas de transtornos mentais (T) em função da violência intrafamiliar (V) e do inventário do clima familiar (C). A forma desse modelo é dada por $T = b_0 + b_1V + b_2C + \epsilon$, em que ϵ representa o erro aleatório normal com média zero e desvio padrão σ, e b_0, b_1 e b_2 são os coeficientes do modelo. A tabela a seguir mostra os resultados da análise de variância (ANOVA) do referido modelo.

Com base na tabela e nas informações apresentadas, julgue o item a seguir.

Fonte de variação	Graus de Liberdade	Soma de Quadrados	Média dos Quadrados	Razão F	p-valor
Regressão	2	608	304	76	<0,0001
Resíduo	98	392	4		
Total	100	1.000			

Conjuntamente, segundo o modelo ajustado, a violência intrafamiliar e o inventário do clima familiar explicam 60,8% da variabilidade total dos sintomas de transtornos mentais.

20. (CESPE/COGE-CE/2019) Considerando-se que, em uma regressão múltipla de dados estatísticos, a soma dos quadrados da regressão seja igual a 60.000 e a soma dos quadrados dos erros seja igual a 15.000, é correto afirmar que o coeficiente de determinação — R^2 — é igual a

a) 0,75.

b) 0,25.

c) 0,50.

d) 0,20.

e) 0,80.

21. (CESPE/TJ-AM/2019) 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.

A correlação linear entre as variáveis x e y é superior a 0,9.

22. (CESPE/PF/2018). Um pesquisador estudou a relação entre a taxa de criminalidade (Y) e a taxa de desocupação da população economicamente ativa (X) em determinada região do país. Esse 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	225
Erro	899	175
Total	900	400

A respeito dessa situação hipotética, julgue o item, sabendo que b > 0 e que o desvio padrão amostral da variável X é igual a 2.

A correlação linear de Pearson entre a variável resposta Y e a variável regressora X é igual a 0,75.

23. (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.

24. (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.

25. (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 $H_A: \beta_0 \neq 0$.

26. (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.

27. (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.

28. (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.

29. (CESPE/TCE-PE/2017). Um estudo de acompanhamento ambiental considerou, para $j = 1, 2, \ldots, 26$, um modelo de regressão linear simples na forma: $y_j = a + bx_j + e_j$, em que a e b são constantes reais, y_j representa a variável resposta referente ao j-ésimo elemento da amostra, x_j é a variável regressora correspondente, e e_j denota o erro aleatório que segue distribuição normal com média nula e variância V. Aplicando-se, nesse estudo, o método dos mínimos quadrados ordinários, obteve-se a reta ajustada $\hat{y}_j = 1 + 2x_j, para\ j = 1, 2, \ldots, 26$

Considerando que a estimativa da variância V seja igual a 6 e que o coeficiente de explicação do modelo (R quadrado) seja igual a 0,64, julgue o item.

A correlação linear entre as variáveis x e y é igual a 0,5, pois a reta invertida proporcionada pelo método de mínimos quadrados ordinários é expressa por $\hat{x}_j = 0{,}5y_i - 0{,}5$, para $j = 1{,}2, \ldots, 26$

30. (FGV/IBGE/2017) Após formular e estimar um modelo de regressão simples, o estatístico responsável pela análise trabalha nos resultados, defrontando-se com a tabela a seguir:

Fonte	S. Quadrados	G.L.	Q. Médio	F-Snedecor	p-value
Equação	450	1	450	12,00	0,21%
Resíduos	300	8	37,50		
Total	750	9	83,33		

A partir desses números, é correto concluir que:

a) com uma amostra de tamanho 10, o modelo é capaz de explicar 60% da variação total;

b) a variância estimada do erro aleatório é inferior a 6;

c) apesar de um poder de explicação de 60%, o modelo não passa no teste de significância da estatística F;

d) a variância da variável explicativa do modelo é igual a 75;

e) a estimativa do coeficiente angular da equação é igual a 2.

31. (CESPE/TCE-SC/2016). Um auditor foi convocado para verificar se o valor de Y, doado para a campanha de determinado candidato, estava relacionado ao valor de X, referente a contratos firmados após a sua eleição.

Tabela de análise de variância de dados					
Fonte de variação	**Graus de Liberdade**	**Soma de Quadrados**	**Quadrados Médios**	**F**	**Pr > F**
Modelo	1	4,623		9,76	0,0261
Erro	5	2,371			
Total	6	7,000			

Com base na situação hipotética e na tabela apresentadas, julgue o item que se segue, considerando-se que $\sum(x_i - \bar{x})^2 = 17,5$ e $E(y^2) = 7,25$

O coeficiente angular é maior que 1.

32. (CESPE/TCE-PA/2016). Uma regressão linear simples é expressa por $Y = a + b \times X + e$, em que o termo e corresponde ao erro aleatório da regressão e os parâmetros a e b são desconhecidos e devem ser estimados a partir de uma amostra disponível. Assumindo que a variável X é não correlacionada com o erro e, julgue o item subsecutivo, nos quais os resíduos das amostras consideradas são IID, com distribuição normal, média zero e variância constante.

Se, em uma amostra de tamanho n = 25, o coeficiente de correlação entre as variáveis X e Y for igual a 0,8, o coeficiente de determinação da regressão estimada via mínimos quadrados ordinários, com base nessa amostra, terá valor $R^2 = 0,64$.

33. (CESPE/TCE-PA/2016)

Fonte de variação	**Graus de Liberdade**	**Soma de Quadrados**	**Quadrados Médios**	**F**	**Pr > F**
Modelo			900		
Erro	98				
Total			90		

Considerando um modelo de regressão linear simples, para averiguar se existe alguma relação entre o salário pago — Y — para uma pessoa em cargo comissionado e o tempo de trabalho — X — dessa pessoa na campanha de determinado padrinho político eleito, foi escolhida uma amostra de indivíduos em cargos comissionados cujos resultados estão apresentados nessa tabela.

Com base nessa situação hipotética e nos dados apresentados na tabela, julgue o item que se segue, relativo à análise de regressão e amostragem.

A variância de Y é maior que 100.

34. (FGV/Pref. Recife/2014) Numa regressão linear simples, obteve-se um coeficiente de correlação igual a 0,78. O coeficiente de determinação é aproximadamente igual a

a) 0,36.

b) 0,48.

c) 0,50.

d) 0,61.

e) 0,69.

35. (FGV/SEAD-AP/2010) Se no ajuste de uma reta de regressão linear simples de uma variável Y em uma variável X o coeficiente de determinação observado foi igual a 0,64, então o módulo do coeficiente de correlação amostral entre X e Y é igual a:

a) 0,24

b) 0,36

c) 0,50

d) 0,64

e) 0,80

GABARITO

Análise de Variância da Regressão

1. ERRADO
2. ERRADO
3. CERTO
4. CERTO
5. CERTO
6. CERTO
7. ERRADO
8. CERTO
9. ERRADO
10. ERRADO
11. LETRA E
12. LETRA C
13. CERTO
14. ERRADO
15. LETRA C
16. CERTO
17. LETRA D
18. LETRA A
19. CERTO
20. LETRA E
21. CERTO
22. CERTO
23. ERRADO
24. ERRADO
25. ERRADO
26. ERRADO
27. CERTO
28. CERTO
29. ERRADO
30. LETRA A
31. ERRADO
32. CERTO
33. ERRADO
34. LETRA D
35. LETRA E