Estratégia Concursos

Aula 00

ANA (Especialista em Regulação de Recursos Hídricos e Saneamento Básico - Especialidade 1) Econometria (Pós-Edital)

Autor:
Equipe Exatas Estratégia Concursos

20 de Janeiro de 2024

Índice

1) Apresentação do Curso 3
2) Aviso 4
3) Correlação Linear 5
4) Regressão Linear Simples 29
5) Análise de Variância da Regressão 43
6) Análise de Resíduos 63
7) Questões Comentadas - Correlação Linear - Cebraspe 75
8) Questões Comentadas - Regressão Linear Simples - Cebraspe 89
9) Questões Comentadas - Análise de Variância da Regressão - Cebraspe 117
10) Questões Comentadas - Análise de resíduos - Cebraspe 150
11) Lista de Questões - Correlação Linear - Cebraspe 155
12) Lista de Questões - Regressão Linear Simples - Cebraspe 164
13) Lista de Questões - Análise de Variância da Regressão - Cebraspe 178
14) Lista de Questões - Análise de Resíduos - Cebraspe 195

APRESENTAÇÃO DO CURSO

Olá, pessoal! Tudo bem?

É com grande satisfação damos início ao nosso curso de **Estatística**!

Os professores **Luana Brandão** e **Djefferson Maranhão** ficarão responsáveis pela elaboração do **Livro Digital**.

Antes de continuarmos, vamos apresentar os professores do material escrito:

Luana Brandão: Oi, pessoal! O meu nome é Luana Brandão e sou professora de Estatística do Estratégia Concursos. Sou Graduada, Mestre e Doutora em Engenharia de Produção, pela Universidade Federal Fluminense. Passei nos concursos de Auditor Fiscal (2009/2010) e Analista Tributário (2009) da Receita Federal e de Auditor Fiscal do Estado do Rio de Janeiro (2010). Sou Auditora Fiscal do Estado do RJ desde 2010. Vamos juntos nesse caminho até a aprovação? **@professoraluanabrandao**

Djefferson Maranhão: Olá, caros alunos! Meu nome é Djefferson Maranhão, sou professor de Estatística do Estratégia Concursos. Graduado e Mestre em Ciência da Computação pela Universidade Federal do Maranhão (UFMA). Desde 2015, sou Auditor da Controladoria Geral do Estado do Maranhão (2015 - 5º lugar). Também exerci os cargos de Analista de Sistemas na UFMA (2010 - 1º lugar) e no TJ-MA (2011 - 1º lugar). Sinto-me honrado em fazer parte de sua jornada rumo à aprovação.

O material escrito em **PDF** está sendo construído para ser sua fonte **autossuficiente** de estudos. Isso significa que o livro digital será **completo** e **voltado para o seu edital**, justamente para que você não perca o seu precioso tempo procurando o conteúdo que será cobrado na sua prova. Ademais, sempre que necessário, você poderá fazer perguntas sobre as aulas no **fórum de dúvidas**. **Bons estudos!**

AVISO IMPORTANTE!

Olá, Alunos (as)!

Passando para informá-los a respeito da **disposição das questões** dentro do nosso material didático. Informamos que a escolha das bancas, dentro dos nossos Livros Digitais, é feita de maneira estratégica e pedagógica pelos nossos professores a fim de proporcionar a melhor didática e o melhor direcionamento daquilo que mais se aproxima do formato de cobrança da banca do seu concurso.

Assim, o formato de questões divididas por tópico facilitará o seu processo de estudo, deixando mais alinhado às disposições constantes no edital.

No mais, continuaremos à disposição de todos no Fórum de dúvidas!

Atenciosamente,

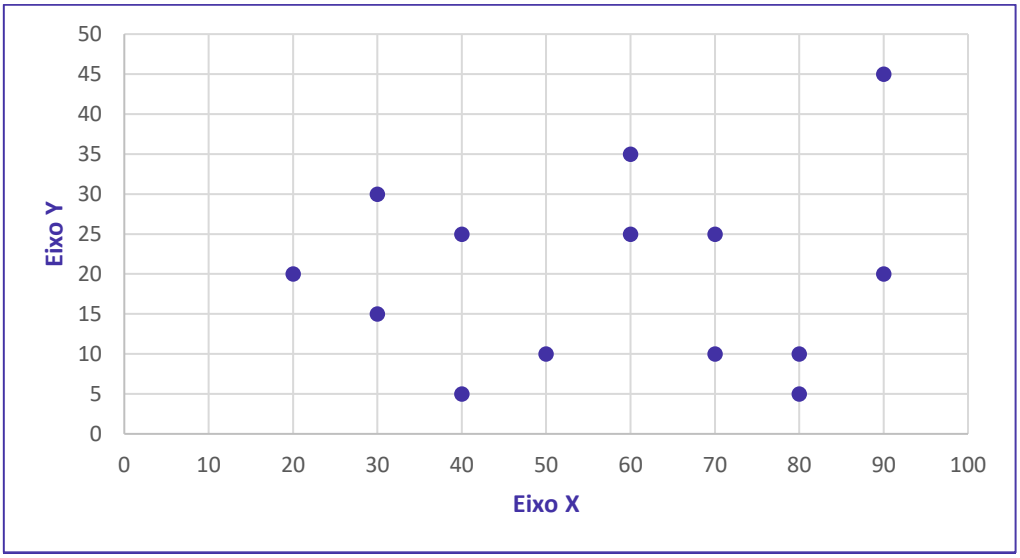
Equipe Exatas

Estratégia Concursos

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).

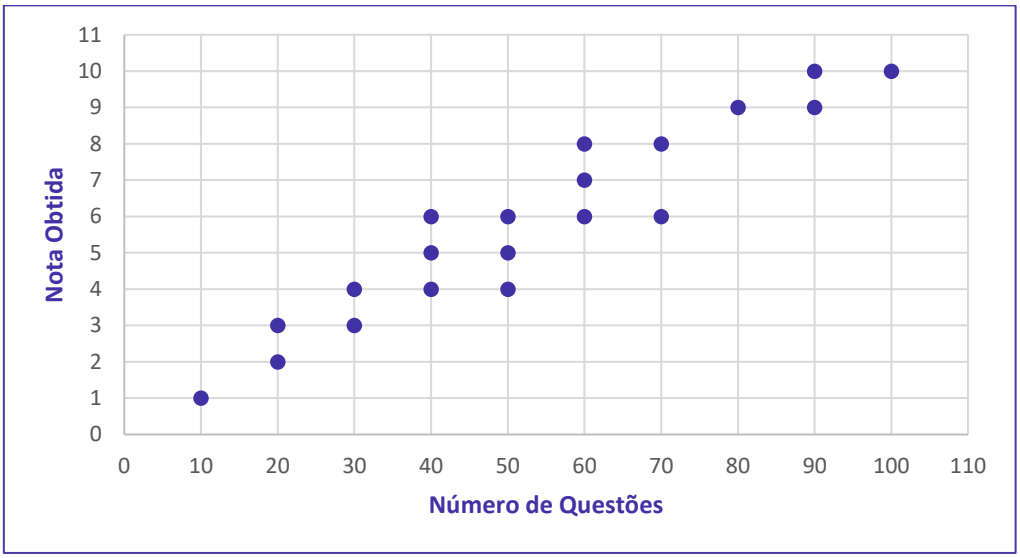


A título exemplificativo, as informações da tabela a seguir referem-se ao número de questões resolvidas por um determinado aluno e a nota obtida por ele em uma avaliação. Observe que quanto maior o número de questões resolvidas, maior é a nota obtida na avaliação.

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

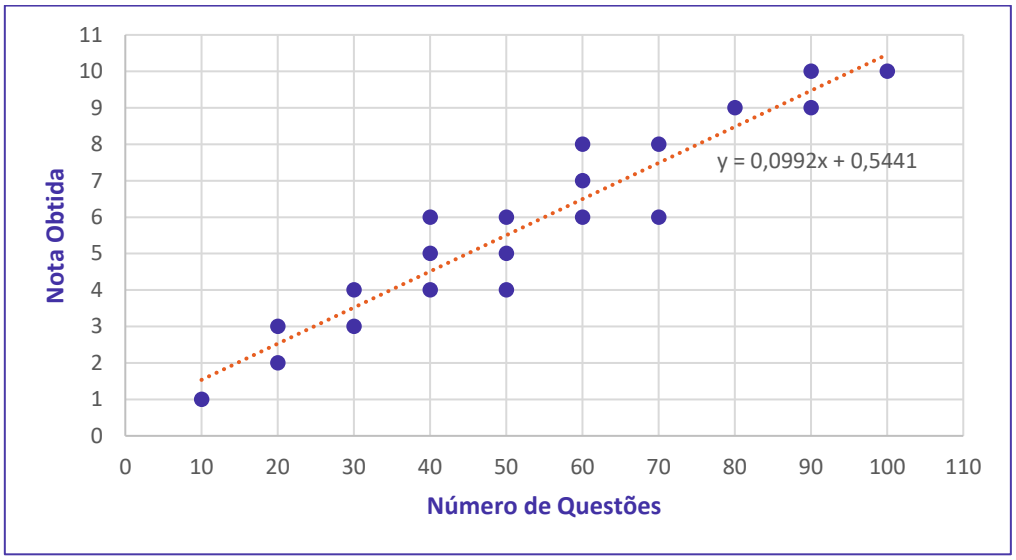
10	30	4
11	10	1
12	60	6
13	50	5
14	70	6
15	90	9
16	100	10
17	20	3
18	40	5
19	60	7
20	50	4

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:

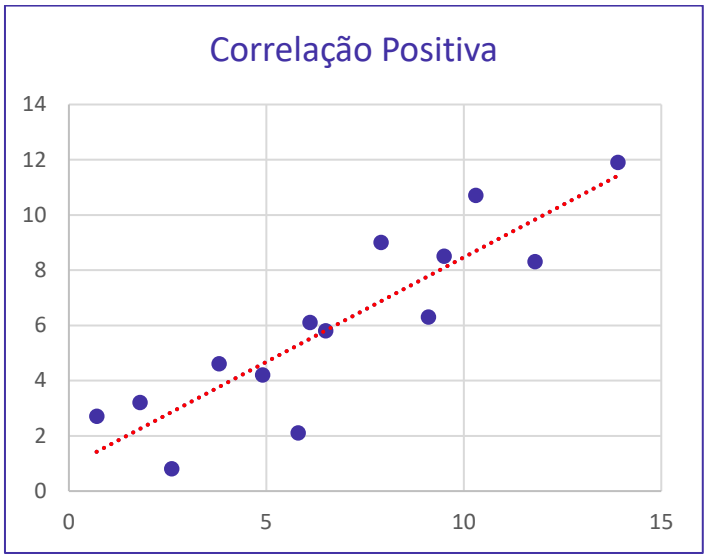
a) direta ou positiva – quando temos dois fenômenos que variam no mesmo sentido. Se aumentarmos ou diminuirmos um deles, o outro também aumentará ou diminuirá;

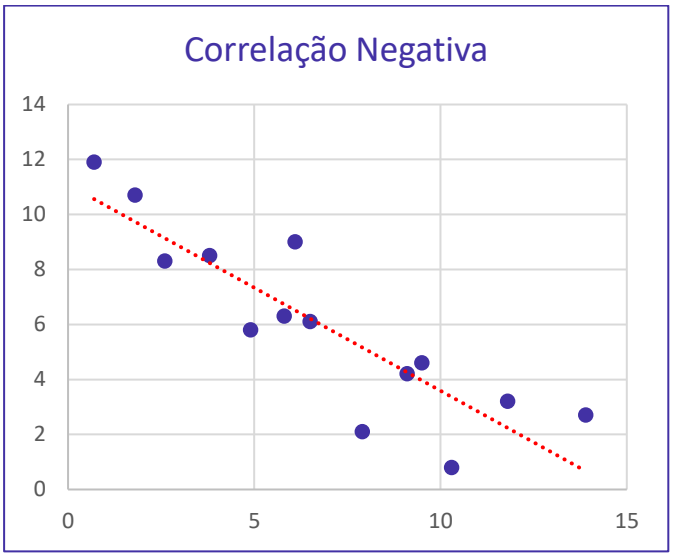
b) inversa ou negativa – quando temos dois fenômenos que variam em sentido contrário. Se aumentarmos ou diminuirmos um deles, acontecerá o contrário com o outro, no caso, diminuirá ou aumentará;

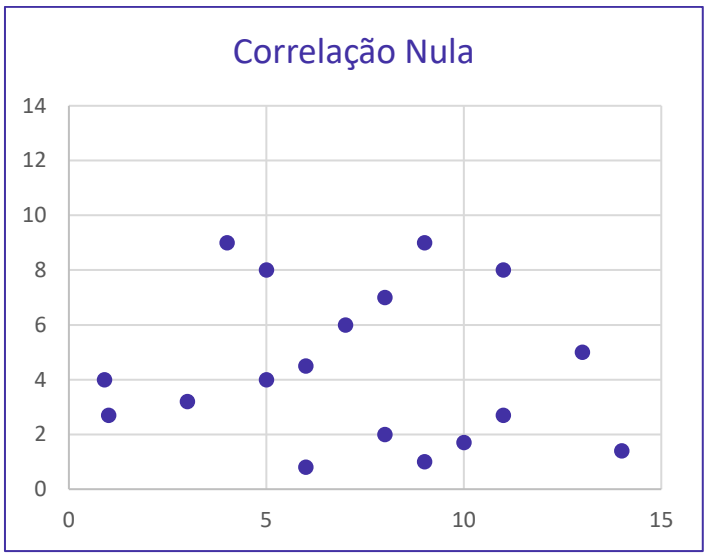
c) inexistente ou nula – quando não existe correlação ou dependência entre os dois fenômenos. Nessa situação, o valor do coeficiente de correlação linear será zero ($r = 0$) ou um valor aproximadamente igual a zero ($r \cong 0$); e

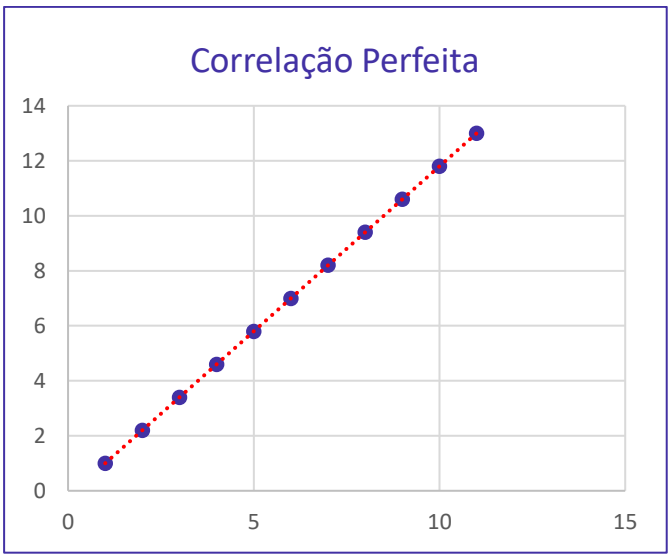
d) perfeita – quando os fenômenos se ajustam perfeitamente a uma reta.

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:

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

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

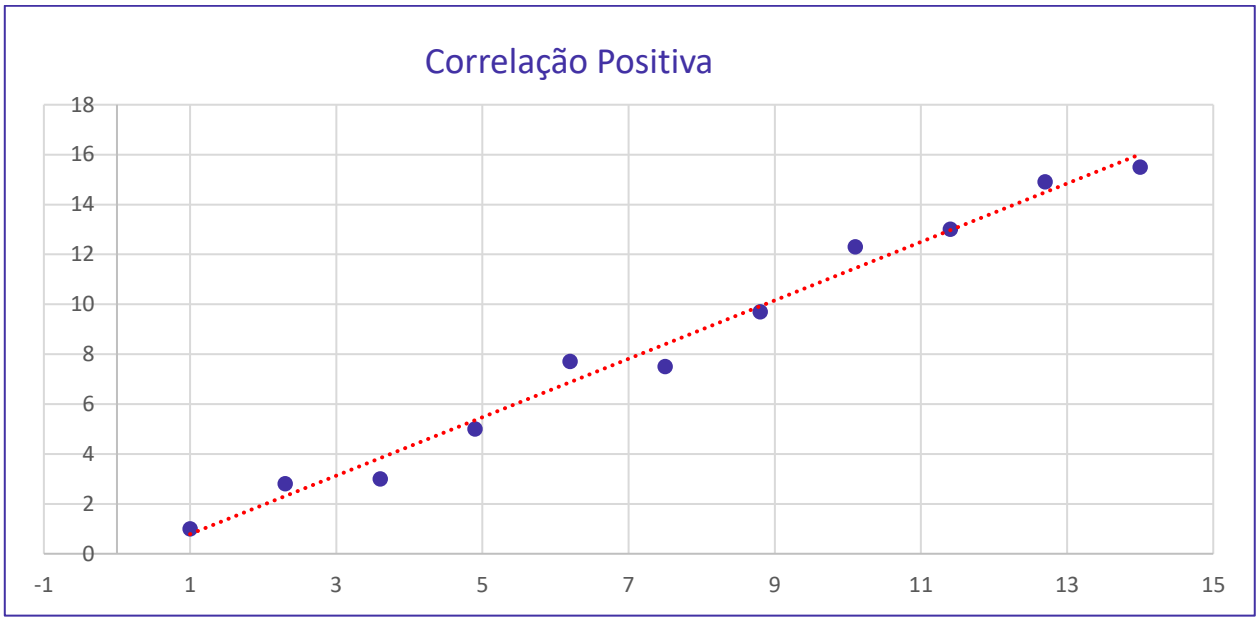
As últimas duas fórmulas são formas alternativas que podem ser empregadas no cálculo do denominador do coeficiente de correlação.

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

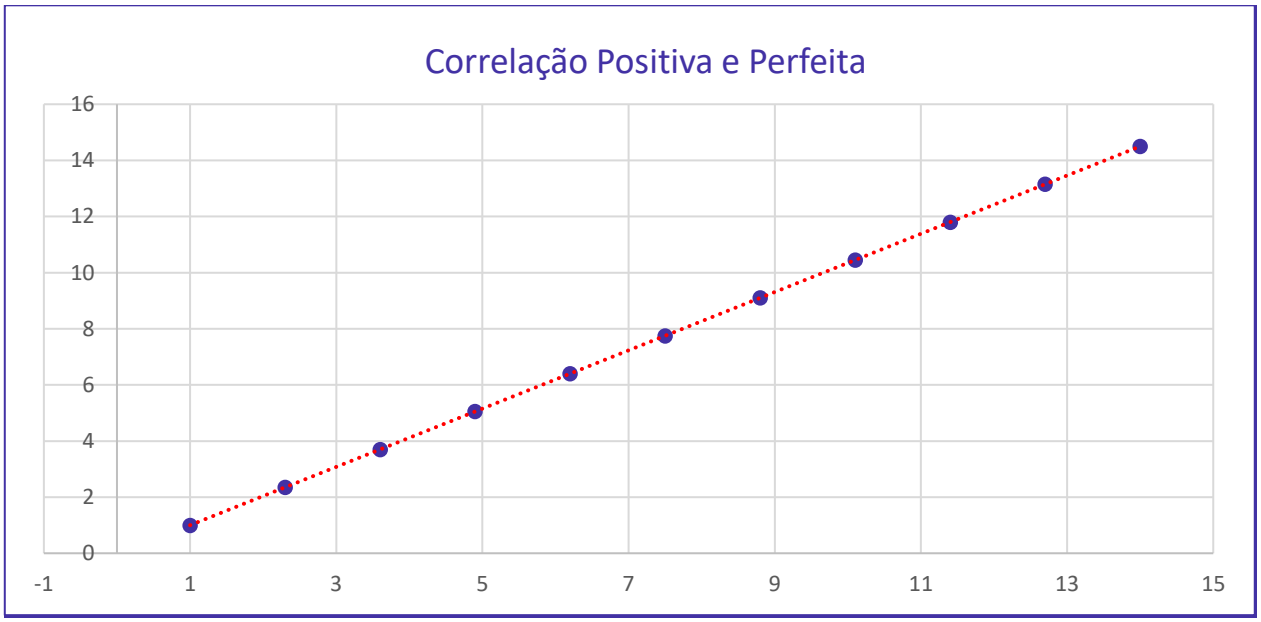
$$\boldsymbol{-1 \leq r \leq 1}$$

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.

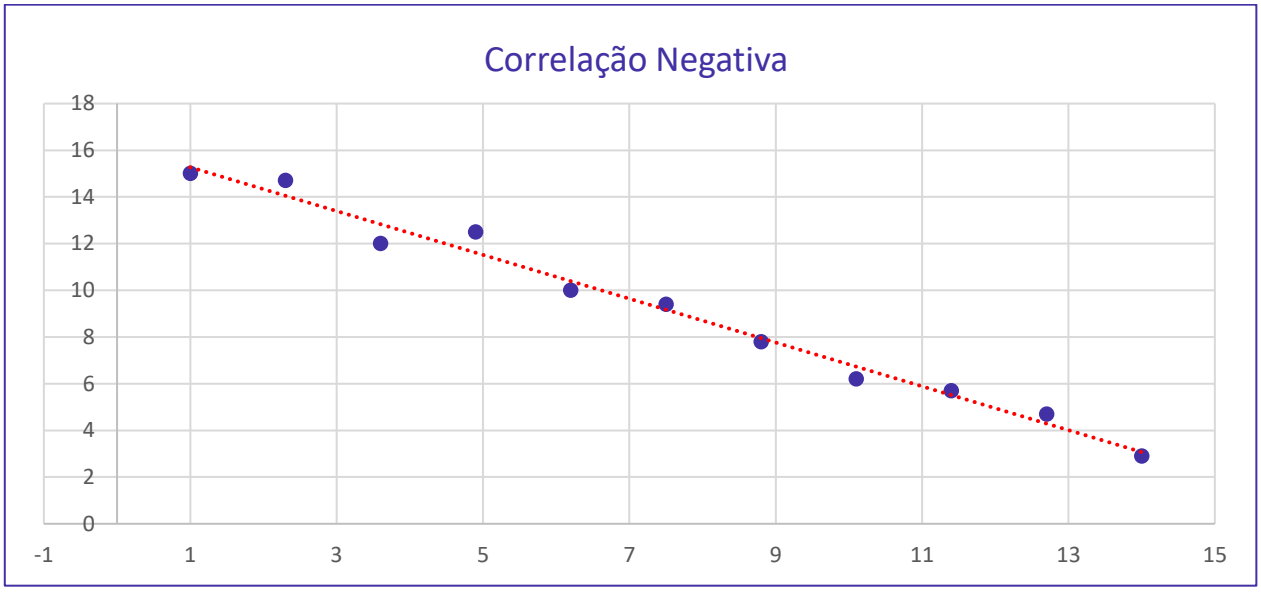


Se a correlação for positiva e todos os pontos estiverem sobre uma mesma reta, o valor de r será exatamente 1. Nesse caso, dizemos que a correlação é **positiva e perfeita**.

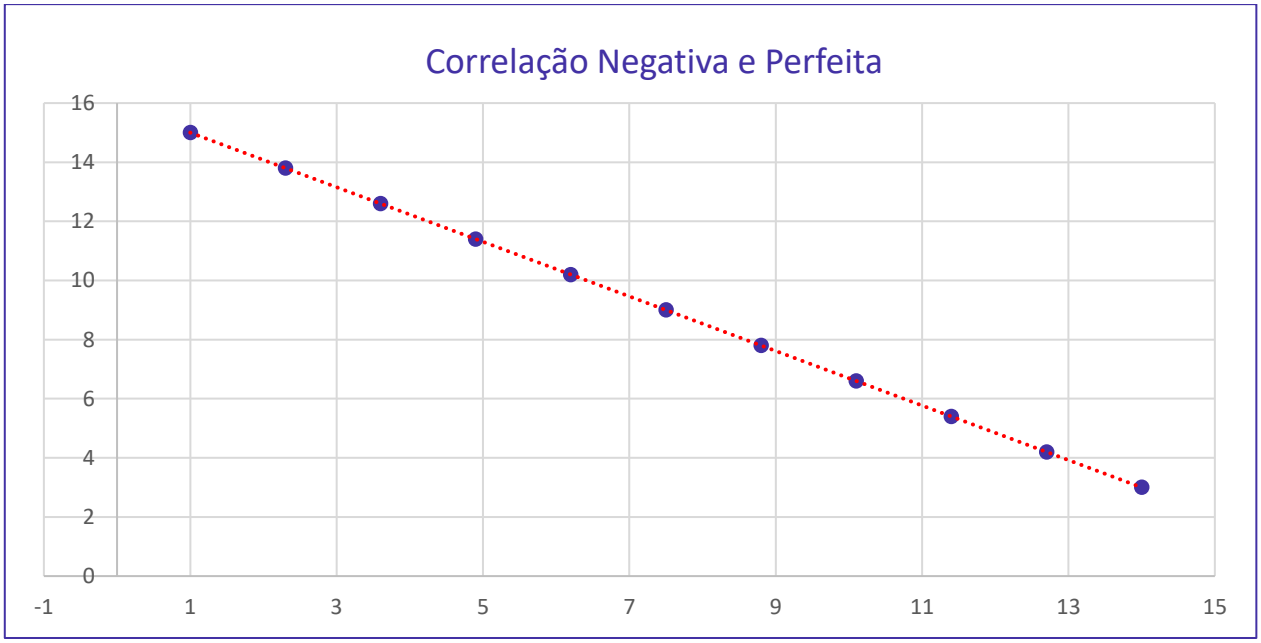


O valor de r é negativo quando a variável Y tende a diminuir ou aumentar quando X aumentar ou diminuir, respectivamente. Nessa situação, dizemos que as variáveis estão negativamente correlacionadas. No

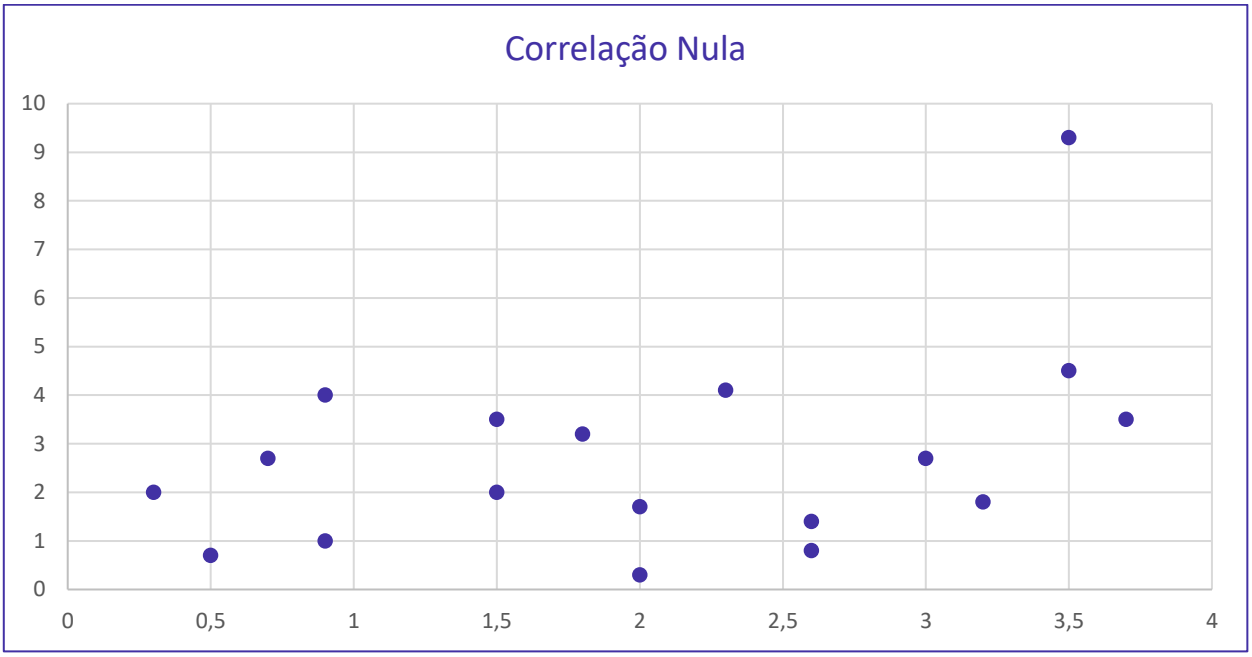
exemplo a seguir, os dados estão praticamente em cima de uma reta, indicando a existência de uma correlação negativa forte, isto é, r muito próximo de -1. No caso, o coeficiente de correlação foi de -0,9918.



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.

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

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

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

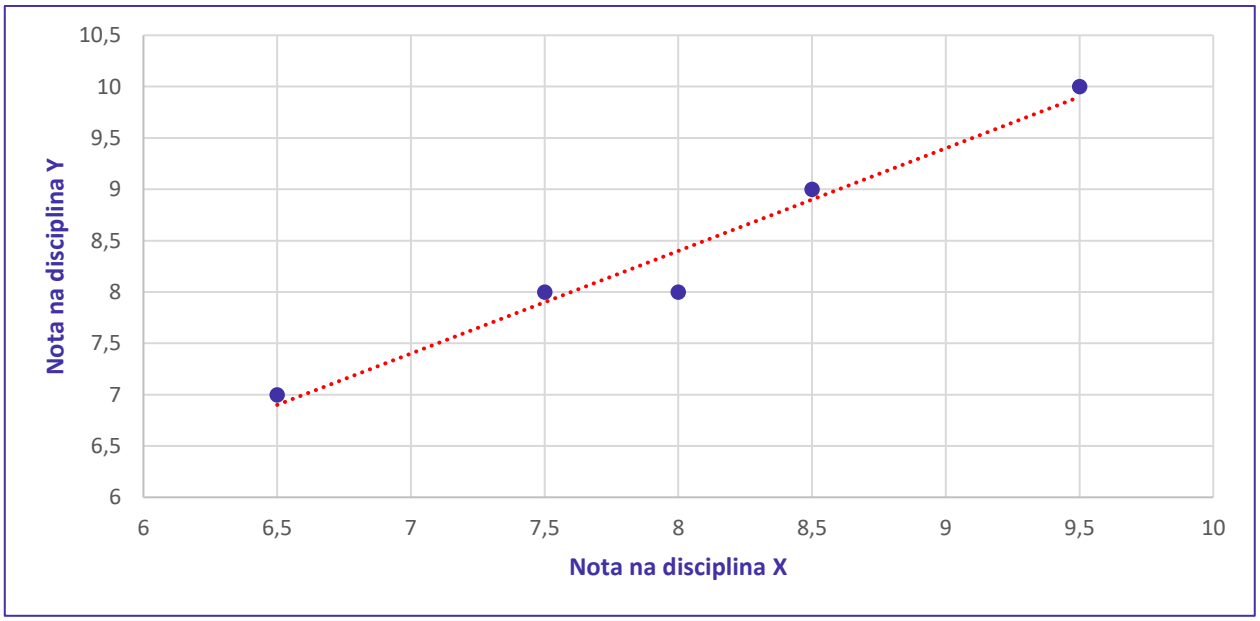
Aplicando esses valores na fórmula do coeficiente de correlação linear, temos:

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

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

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

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.

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

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

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

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

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

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

Encontrando as médias das variáveis transformadas:

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

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

Organizando a tabela auxiliar, teríamos:

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

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

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

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

REGRESSÃO LINEAR SIMPLES

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

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

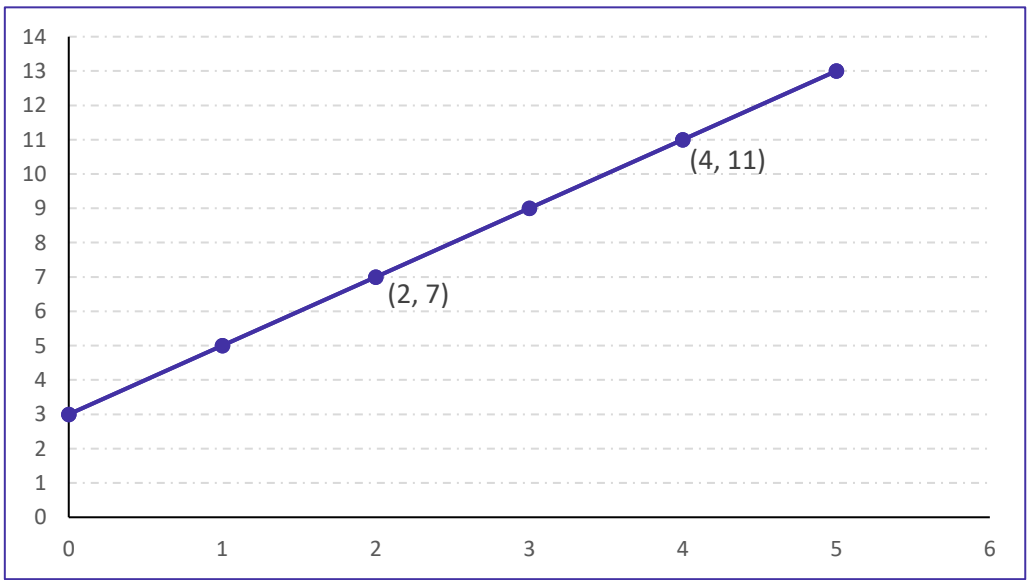
$$y = m \cdot x + b$$

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

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

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

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



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

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

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

Dessa forma, a equação da reta fica:

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

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

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

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

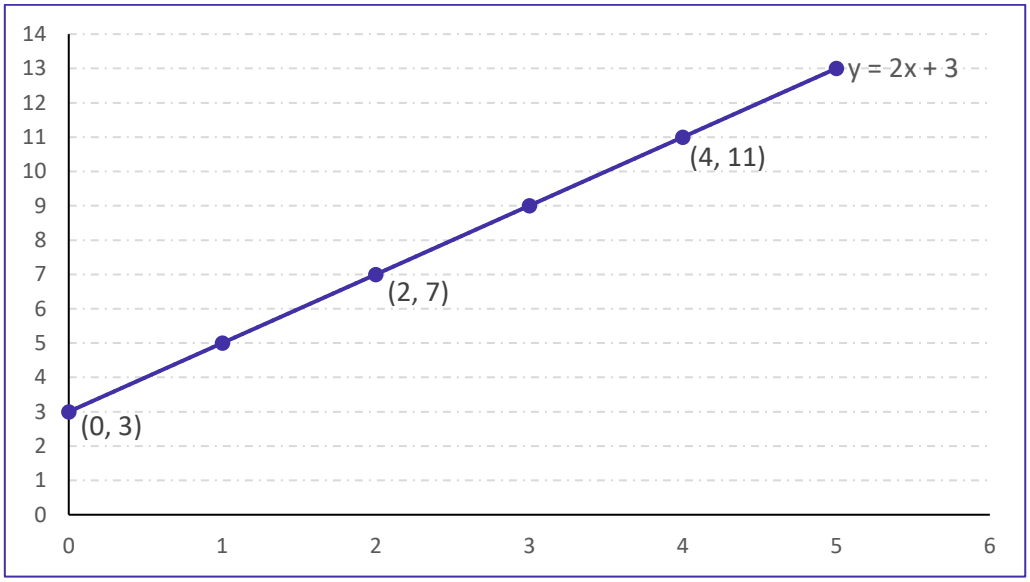
$$7 = 4 + b$$

$$b = 3$$

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

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

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



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

Observe:

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

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

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

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

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

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

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

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

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

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

O erro aleatório ϵ segue a distribuição normal padrão.

Comentários:

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

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

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

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

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

Gabarito: Certo.

Método dos Mínimos Quadrados

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

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

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

A expressão usada para determinar a reta de regressão é:

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

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

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

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

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

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

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

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

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

Para o numerador da fórmula, temos:

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

Para o denominador da fórmula, temos:

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

Logo,

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

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

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

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

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

Calculando o valor de b:

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

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

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

O valor de a é calculado por:

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

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

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

$$a = 10{,}499$$

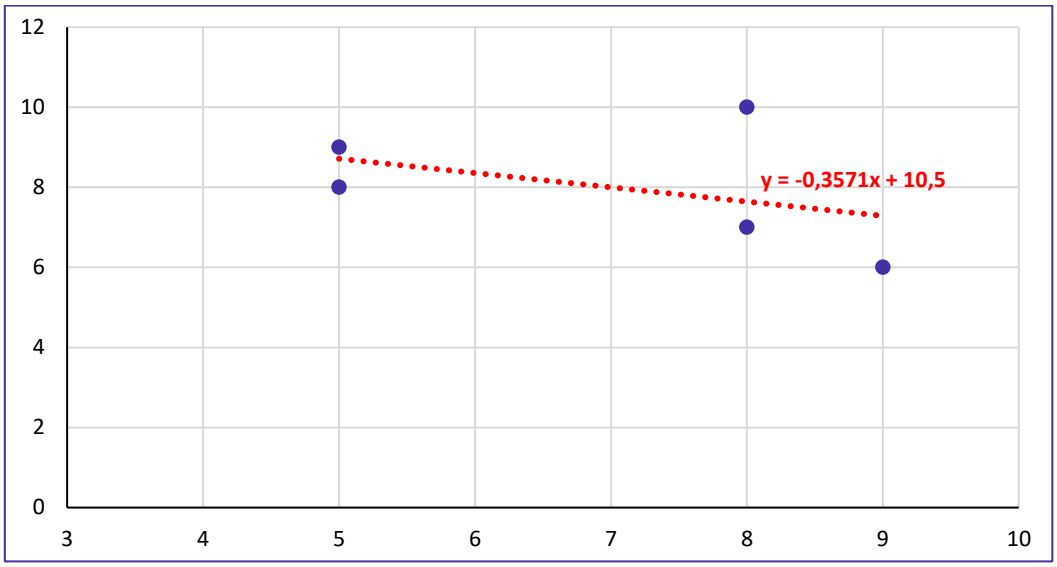
Assim, a reta de regressão estimada é:

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

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

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

Montando o gráfico com os valores estimados, temos:



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

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

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

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

HORA DE PRATICAR!

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

a) 4.

b) 4,5.

c) 5.

d) 5,5.

e) 6.

Comentários:

Vamos montar uma tabela com os dados fornecidos:

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

Pelo método dos mínimos quadrados, temos:

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

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

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

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

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

$$a = 2{,}25$$

Nossa reta de regressão será:

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

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

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

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

$$\hat{Y} = 5$$

Gabarito: C.

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

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

A partir desses resultados, julgue o item a seguir.

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

Comentários:

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

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

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

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

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

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

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

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

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

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

Assim, o modelo de regressão é dado por:

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

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

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

$$\hat{Y} = 12$$

Gabarito: Certo.

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

a) 11,6

b) 15,0

c) 13,2

d) 12,4

e) 14,4

Comentários:

A reta calculada é expressa por:

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

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

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

Agora, vamos calcular a média de t:

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

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

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

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

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

$$b = 0{,}8$$

A reta fica assim:

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

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

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

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

Gabarito: D.

Reta Passando pela Origem

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

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

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

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

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

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

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

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

A reta de regressão ajustada é:

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

Os desvios ou resíduos são dados por:

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

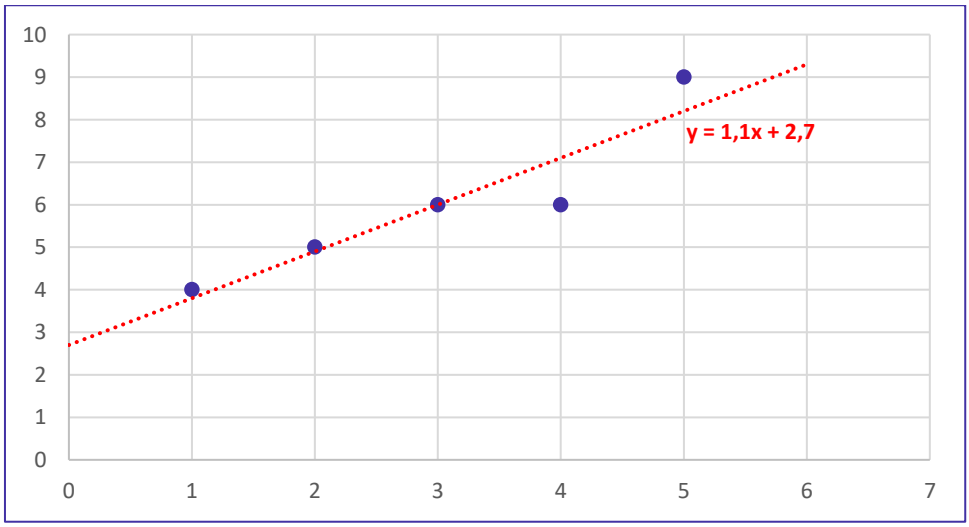
Nesse caso, não há garantia de que o somatório dos resíduos seja igual a zero.

EXEMPLIFICANDO

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

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

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



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

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

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

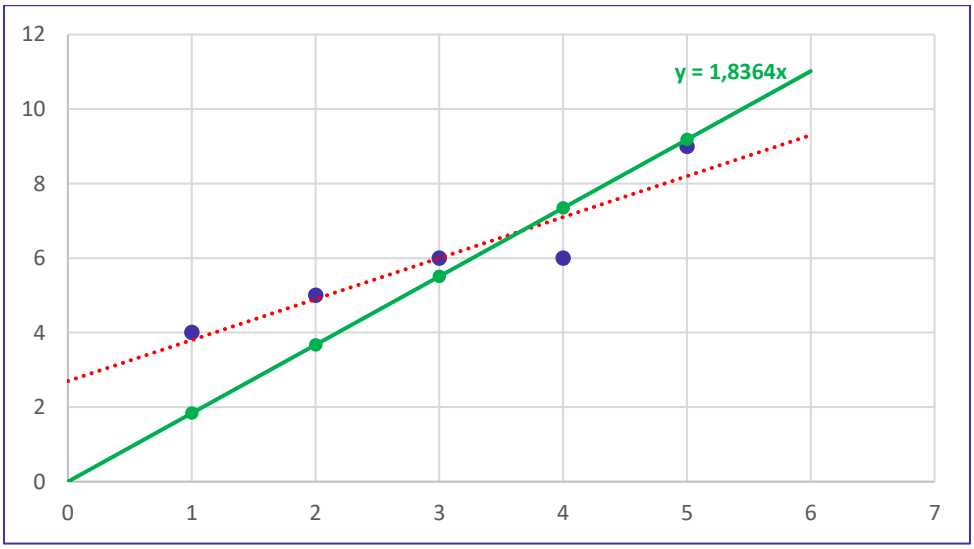
$$b = \frac{\sum X_i \times Y_i}{\sum {X_i}^2} = \frac{101}{55} \cong 1{,}836$$

Portanto, a reta de regressão ajustada é:

$$\widehat{Y_i} = bX_i$$

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

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



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

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

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

ANÁLISE DE VARIÂNCIA DA REGRESSÃO

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

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

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

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

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

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

Graus de Liberdade

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

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

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

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

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

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

Daí, concluímos que:

$$n - 1 = 1 + GL_{Resíduos}$$

$$GL_{Resíduos} = n - 2$$

ATENÇÃO
DECORE!

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

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

O número de graus de liberdade dos resíduos é:

$$GL_{Resíduos} = n - 2$$

O número de graus de liberdade total é:

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

Somas de Quadrados

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

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

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

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

Agora, elevando os dois lados ao quadrado:

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

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

Somando tudo, temos:

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

É possível demonstrar que :

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

Logo,

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

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

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

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

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

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

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

Podemos demonstrar que:

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

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

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

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

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

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

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

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

$$SQT = SQM + SQR$$

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

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

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

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

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

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

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

Coeficiente de Determinação

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

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

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

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

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

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

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

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

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

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

Coeficiente de Determinação Ajustado

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

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

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

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

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

Quadrados Médios

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

a) quadrado médio do modelo (QMM):

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

b) quadrado médio dos resíduos (QMR):

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

c) quadrado médio total (QMT):

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

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

Estatística F (Razão F)

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

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

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

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

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

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

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

Tabela de Análise de Variância da Regressão

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

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

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

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

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

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

Comentários:

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

$$n - 1 = 11$$

$$n = 12\ municípios.$$

Gabarito: Errado.

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

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

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

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

Comentários:

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

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

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

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

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

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

Gabarito: Errado.

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

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

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

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

Comentários:

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

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

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

Gabarito: Errado.

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

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

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

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

Comentários:

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

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

A variância amostral é calculada por:

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

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

$$n - 1 = 11$$

Logo, a variância amostral é:

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

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

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

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

Gabarito: Errado.

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

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

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

A estimativa de σ^2 foi igual a 10.

Comentários:

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

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

Gabarito: Certo.

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

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

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

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

Comentários:

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

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

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

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

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

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

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

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

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

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

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

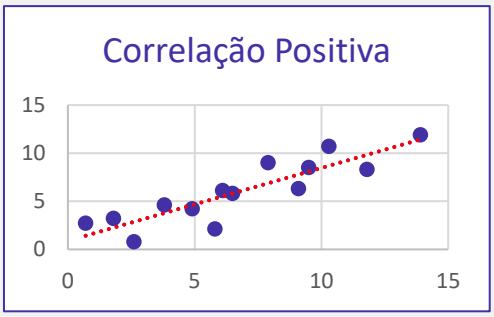
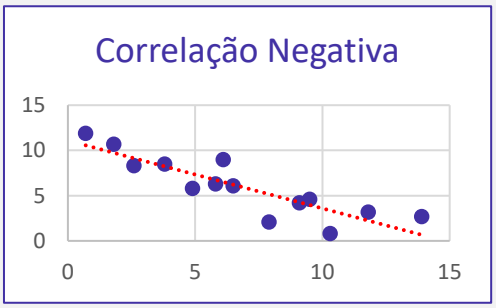
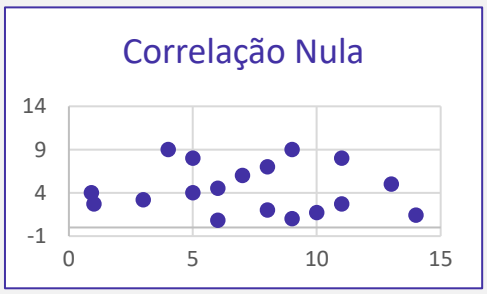
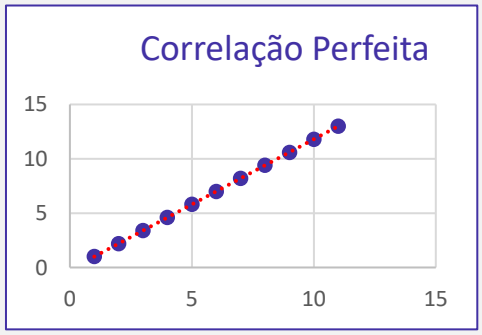
$$\overline{R^2} = 0{,}89.$$

Gabarito: Certo.

RESUMO DA AULA

CORRELAÇÃO LINEAR

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

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

Coeficiente de Correlação de Pearson

COEFICIENTE DE CORRELAÇÃO LINEAR DE PEARSON

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

FÓRMULA

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

FÓRMULAS ALTERNATIVAS

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

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

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

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

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

Propriedades do Coeficiente de Correlação

1º Propriedade

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

2º Propriedade

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

REGRESSÃO LINEAR SIMPLES

REGRESSÃO LINEAR SIMPLES

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

Trata-se da equação que representa uma reta:

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

- Propriedades

Sobre a Regressão Linear Simples, podemos afirmar que:

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

Método dos Mínimos Quadrados

MÉTODO DOS MÍNIMOS QUADRADOS

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

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

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

Expressão usada para determinar a reta de regressão é:

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

Reta Passando pela Origem

MODELO DE REGRESSÃO QUE PASSA OBRIGATORIAMENTE PELA ORIGEM É:

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

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

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

ANÁLISE DE VARIÂNCIA DA REGRESSÃO

ANÁLISE DE VARIÂNCIA DA REGRESSÃO

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

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

Testar a hipótese:

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

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

Graus de Liberdade

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

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

O número de graus de liberdade dos resíduos é:

$$GL_{Resíduos} = n - 2$$

O número de graus de liberdade total é:

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

Somas de Quadrados

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

$$SQT = SQM + SQR$$

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

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

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

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

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

A soma dos quadrados dos resíduos é calculada pela fórmula:

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

Coeficiente de Determinação

O coeficiente de determinação é calculado pela fórmula:

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

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

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

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

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

Coeficiente de Determinação Ajustado

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

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

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

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

Quadrados Médios

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

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

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

Estatística F (Razão F)

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

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

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

Tabela de Análise de Variância da Regressão

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

ANÁLISE DE RESÍDUOS

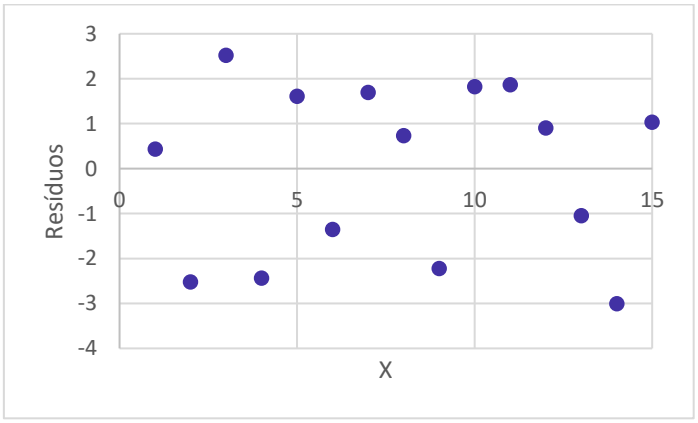
Os resíduos do modelo de regressão são definidos por $e_i = y_i - \hat{y}_i$, em que y_i é uma observação real e $\hat{y}_i$ é o valor ajustado correspondentes, proveniente do modelo de regressão. **Os resíduos desempenham um papel importante no julgamento da adequação do modelo**.

Ao ajustar um modelo de regressão, devemos levar em consideração uma série de suposições. Por exemplo, a estimação dos parâmetros do modelo requer a suposição de que os erros sejam variáveis aleatórias não correlacionadas, com média zero e variância constante. Além disso, os testes de hipótese e estimação do intervalo requerem que os erros sejam normalmente distribuídos.

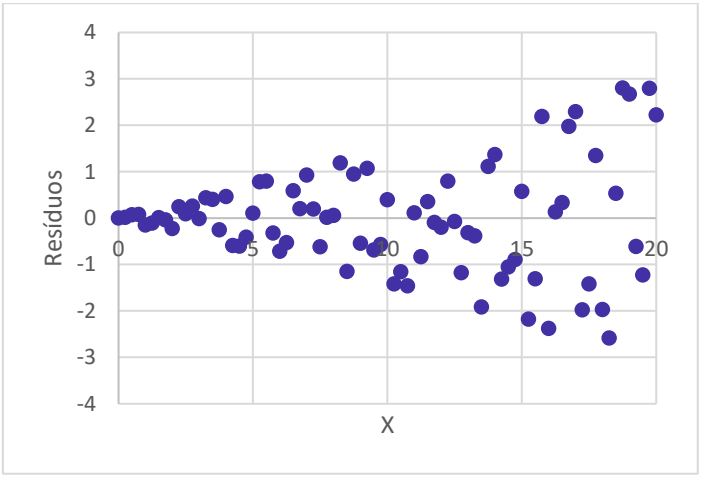
Por isso, devemos sempre duvidar da validade dessas suposições e conduzir análises para examinar a **adequação do modelo**. **A análise dos resíduos normalmente é utilizada na avaliação da suposição de que os erros sejam distribuídos de forma aproximadamente normal, com variância constante, assim como na determinação da utilidade dos termos adicionais no modelo**.

Como uma verificação aproximada da normalidade, podemos construir um **histograma de frequência dos resíduos** ou um **gráfico de probabilidade normal dos resíduos**. Como os tamanhos das amostras na regressão normalmente são muito pequenos para um histograma ser significativo, o **método de plotar a probabilidade normal é preferido**.

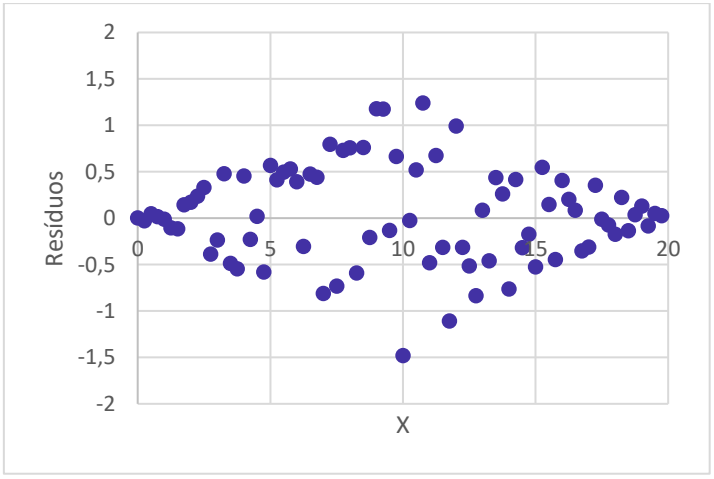
O gráfico de resíduos mais comum envolve os resíduos e_i representados em uma sequência temporal contra os valores de $\hat{y}_i$ e contra a variável independente x. O **gráfico nulo**, que representa a **situação ideal**, ocorre quando **todas as suposições são atendidas**, mostrando que os **resíduos se comportam de forma aleatória**, com **dispersão em torno de zero** e **nenhuma tendência forte para ser maior ou menor que zero**.



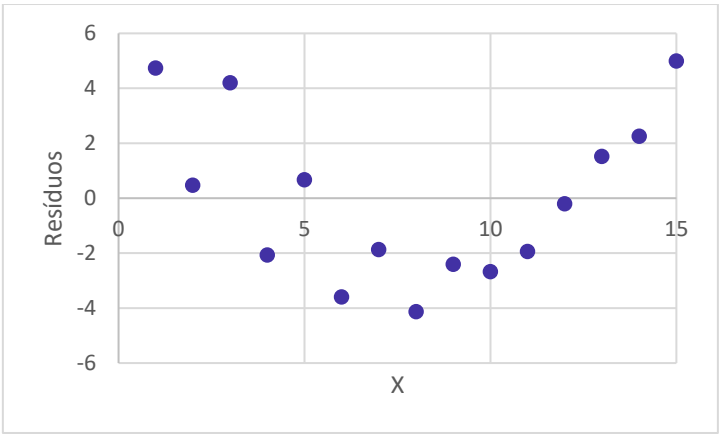
Se os resíduos aparecem como no gráfico a seguir, a variância das observações pode estar crescendo com o tempo ou com a magnitude de $\hat{y}_i$ ou x_i. **Transformações de dados na resposta y podem ser usadas para eliminar esse problema**. As transformações mais usadas para estabilizar a variância incluem o uso de $\sqrt{y}$, $ln(y)$ ou $1/y$ como a resposta.



Gráficos de resíduos contra $\hat{y}_i$ e x_i que pareçam com o próximo também indicam desigualdade de variância.



Gráficos de resíduos que pareçam com o seguinte também indicam que o modelo não é adequado; isto é, **termos de ordens maiores (quadráticos) devem ser adicionados ao modelo**, uma transformação sobre a variável x ou a variável y (ou ambas) deve ser considerada, ou outros regressores devem ser considerados.



Normalmente, ao analisarmos os resíduos da regressão, devemos empregar alguma técnica de padronização dos resíduos, pois isso os torna diretamente comparáveis. Os **resíduos padronizados** são calculados na forma:

$$d_i = \frac{e_i}{\sqrt{\hat{\sigma}^2}}$$

Se os erros forem normalmente distribuídos, então aproximadamente 95% dos resíduos padronizados devem cair no intervalo $(-2, +2)$. Além disso, os resíduos padronizados são escalonados de modo que seus desvios-padrão sejam aproximadamente iguais a 1. Consequentemente, os resíduos que estiverem fora desse intervalo podem indicar possíveis **outliers** ou **observações não usuais**.

Porém, há outras maneiras de padronizar os resíduos de regressão, outra muito popular é o resíduo na forma de Student (ou resíduo estudentizado):

$$r_i = \frac{e_i}{\sqrt{\hat{\sigma}^2(1 - h_{ii})}}$$

em que h_{ii} é o i-ésimo elemento da diagonal da matriz

$$H = X(X^T X)^{-1} X^T$$

A matriz H é chamada de matriz chapéu, uma vez que:

$$\hat{y} = X\hat{\beta} = X(X^T X)^{-1} X^T y = Hy$$

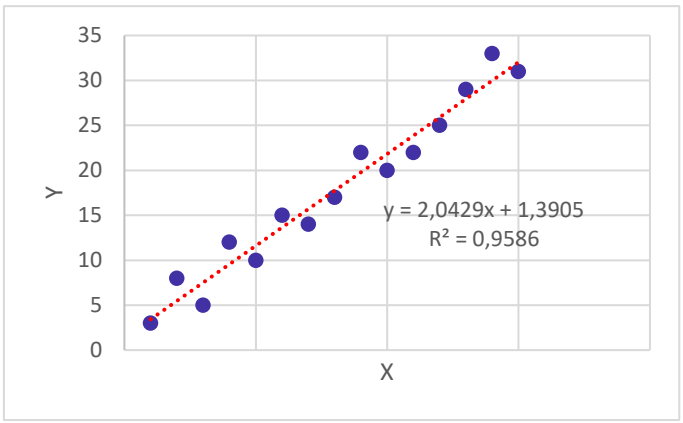
Logo, H transforma os valores observados de y em um vetor de valores ajustados $\hat{y}$.

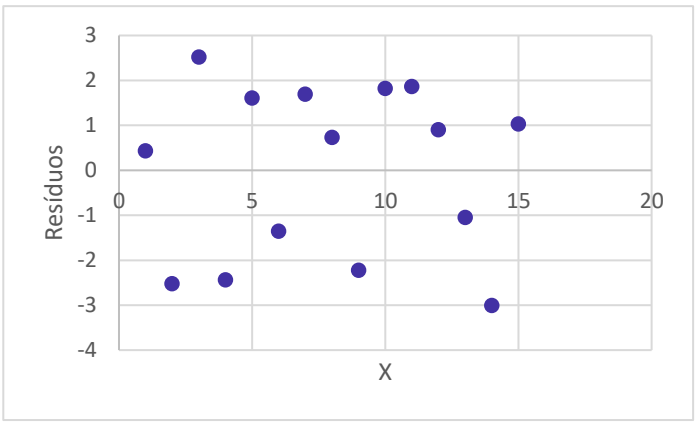
Linearidade dos Termos de Erro

A **condição de linearidade** implica que **os valores y estão todos em uma mesma linha reta**. O gráfico de dispersão de x e y deve mostrar uma tendência linear. A reta de regressão reflete a tendência de linearidade na forma de equação linear. O erro é expresso como pontos dispersos ao redor da reta ajustada.

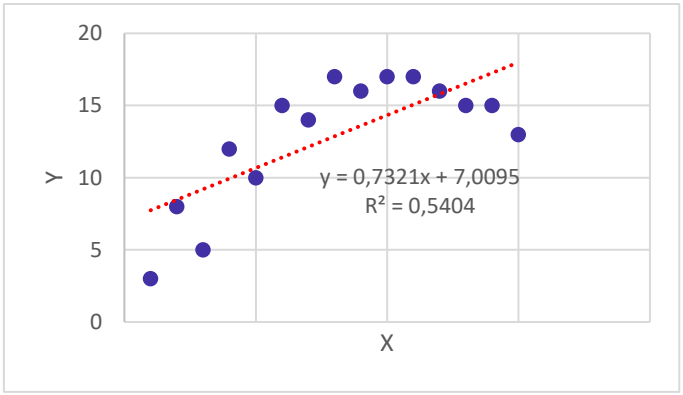
Se a relação de x e y for realmente linear, **os pontos dispersos não apresentarão nenhuma tendência, seja linear, curvilínea etc**. Significa dizer que os resíduos e a variável x não se correlacionam. Em outras palavras, os resíduos aparecem dispersos aleatoriamente em relação a x.

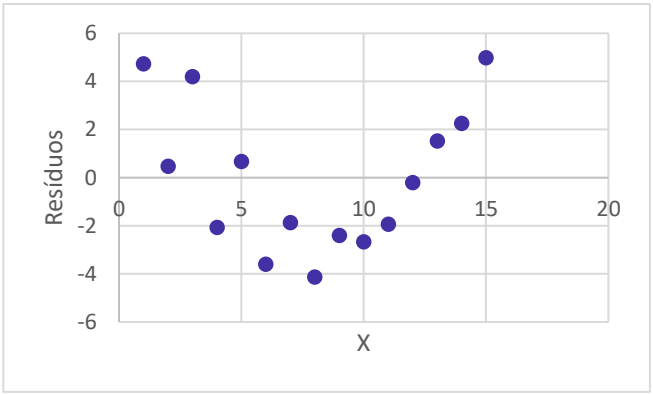
Os gráficos a seguir exibem a relação linear de x e y e os resíduos para valores de x. Os resíduos localizam-se em torno de zero e estão espalhados aleatoriamente, sem nenhum padrão. A característica aleatória em torno de zero na distribuição residual confirma que a suposição de linearidade está correta.



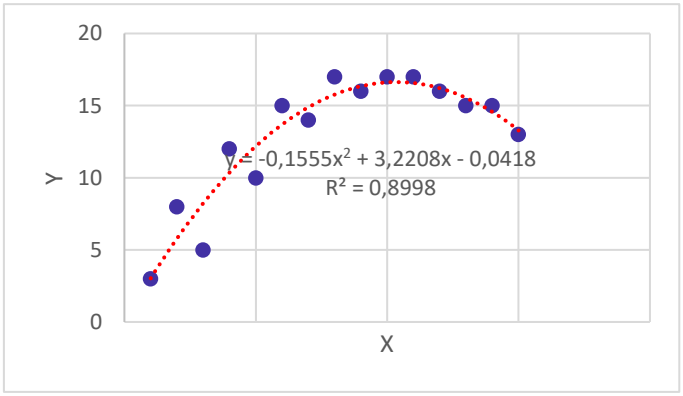


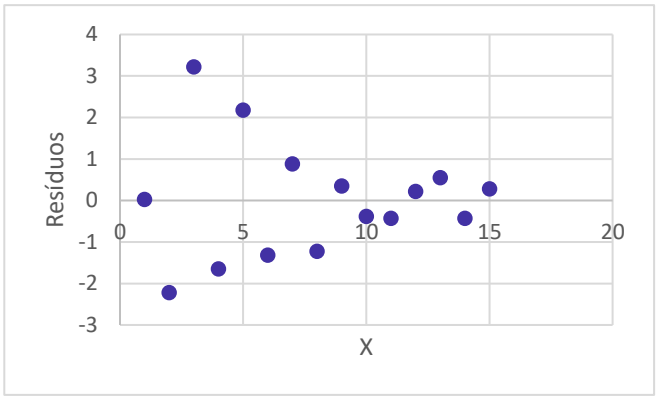
Em contraste, os gráficos seguintes mostram um padrão típico de resíduos que indicam a existência de uma relação não linear não representada no modelo. Qualquer padrão curvilíneo consistente nos resíduos indica que uma ação corretiva aumentará a precisão do modelo, bem como a validade dos coeficientes estimados:





Portanto, podemos presumir que há um termo de X^2 omitido e que precisa ser adicionado ao modelo. Depois de adicionar o termo quadrático, os resíduos finalmente são dispersos aleatoriamente sem qualquer tendência, como mostram os gráficos a seguir:





Há três ações corretivas possíveis no que diz respeito à ausência de linearidade:

a) transformações de dados (logaritmo, raiz quadrada etc.) de uma ou mais variáveis independentes para conseguir linearidade;

b) inclusão direta de relações não lineares no modelo de regressão, mediante a criação de termos polinomiais;

c) uso de métodos especializados, como a regressão não linear especificamente elaborada para acomodar os efeitos curvilíneos de variáveis independentes.

Independência dos Termos de Erro

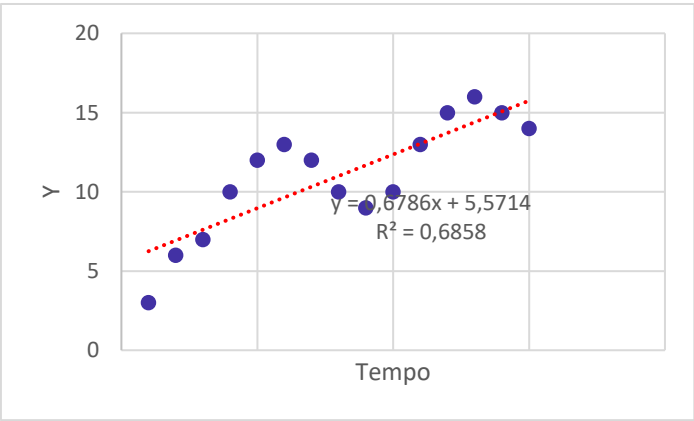
Segundo a suposição de independência, **uma observação do termo de erro não deve prever a próxima observação**, o que significa que o valor previsto não está relacionado com qualquer outra previsão. Podemos identificar melhor tal ocorrência fazendo o gráfico de resíduos contra o tempo ou em relação à ordem na qual o experimento foi realizado**.**

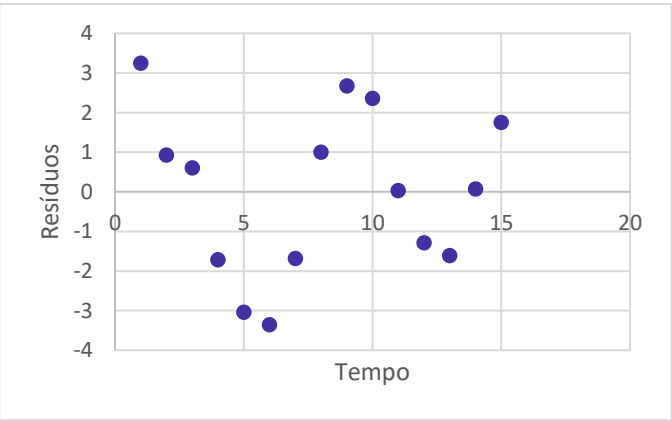
Se os resíduos forem independentes, o padrão deverá parecer aleatório e semelhante ao gráfico nulo de resíduos. Por sua vez, se houver violação dessa suposição, seremos capazes de identificar um padrão consistente nos resíduos em relação ao tempo.

Quando a suposição de independência é violada, a observação de um erro positivo pode aumentar sistematicamente a probabilidade de que o erro subsequente seja positivo **(correlação positiva)**; ou pode ser mais provável que tenha o sinal oposto **(correlação negativa)**. Esse problema acontece com mais frequência em modelos de séries temporais e são conhecidos como **correlação serial** e **autocorrelação**

Por exemplo, se as vendas forem inesperadamente altas em um dia, provavelmente serão mais altas que a média no dia seguinte. Esse tipo de correlação é comum para determinados tópicos, como taxas de inflação, PIB, desemprego e assim por diante.

Os gráficos a seguir exibem um gráfico de resíduos que mostra uma associação entre os resíduos e o tempo. Reparem no padrão oscilatório dos resíduos ao longo do tempo:





Normalidade da Distribuição dos Termos de Erro

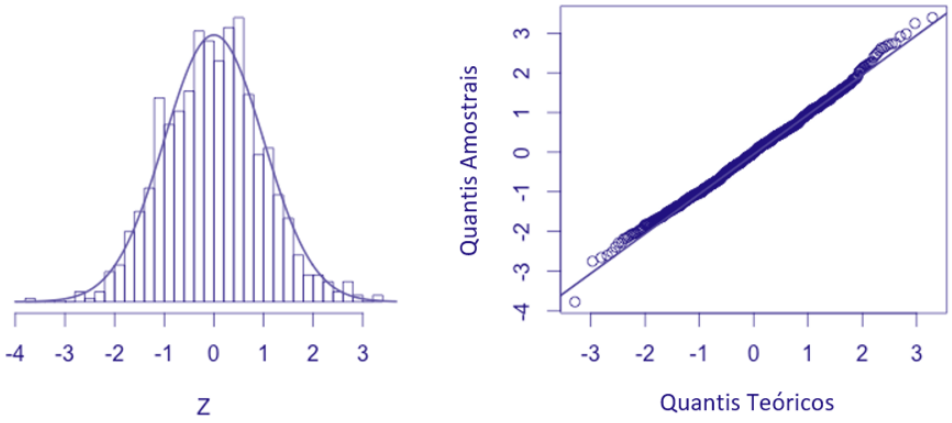
A maneira mais simples de determinar se os resíduos seguem uma distribuição normal é usando um **histograma de resíduos**, com a verificação visual para uma distribuição que se aproxima da normal. Embora muito simples, esse método pode não funcionar tão bem com amostras pequenas.

Outra ferramenta comumente usada é o gráfico de probabilidade normal. Ele difere do gráfico de resíduos no sentido de que **os resíduos padronizados são comparados com a distribuição normal**. **A distribuição normal forma uma reta diagonal**.

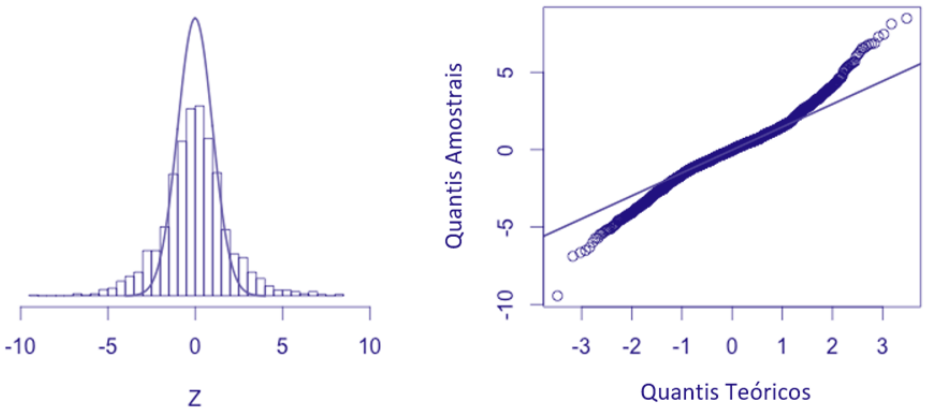
O gráfico de probabilidade normal utilizado é o **quantil-quantil normal (gráfico Q-Q)**. Também pode ser usado o gráfico **percentil-percentil normal (P-P)**. Nesses gráficos, para que uma distribuição de resíduos possa ser classificada como normal, os resíduos observados devem estar ao redor da linha diagonal

A seguir, temos exemplos de gráficos quantil-quantil para vários tipos de distribuições de resíduos:

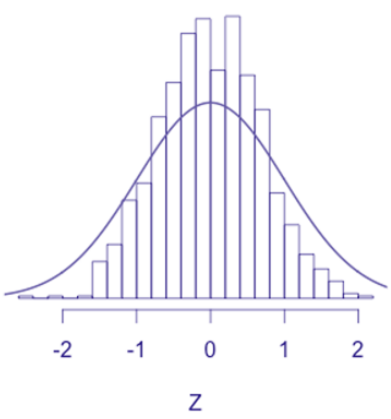
a) distribuição simétrica:

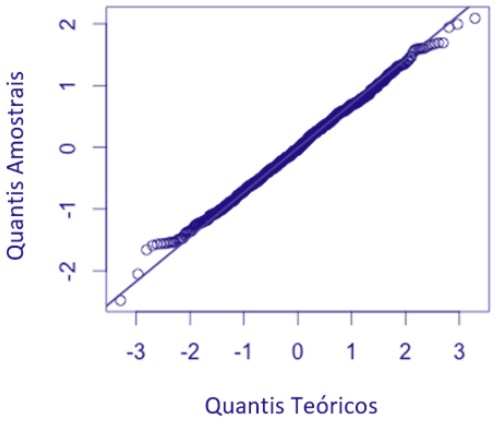


b) distribuição simétrica com caudas pesadas:

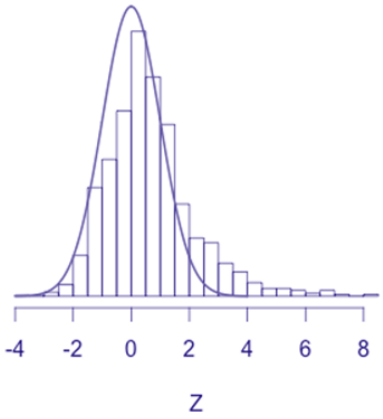


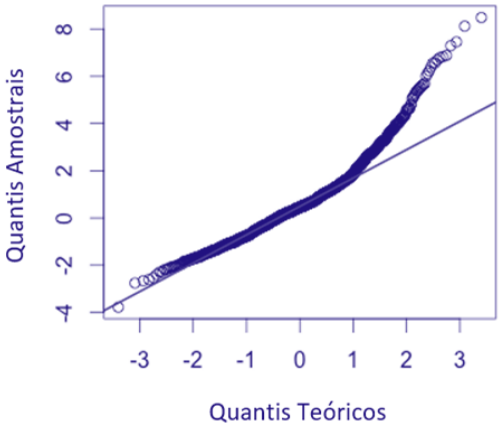
c) distribuição simétrica com caudas leves:



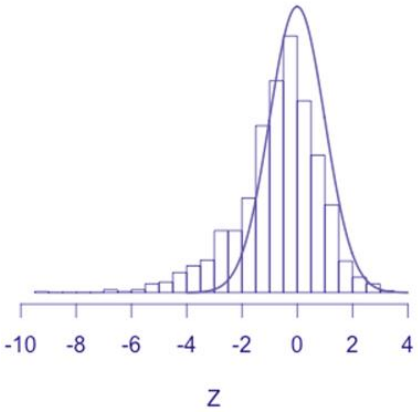


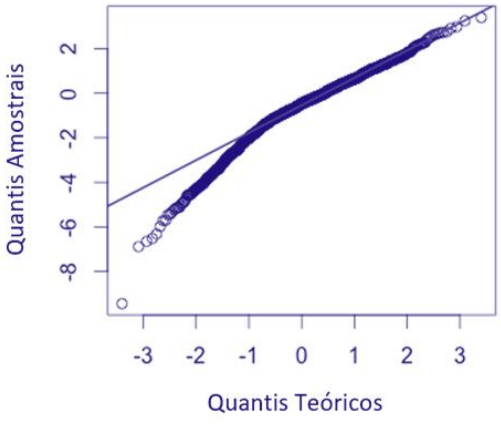
d) distribuição assimétrica à direita:





d) distribuição assimétrica à esquerda:



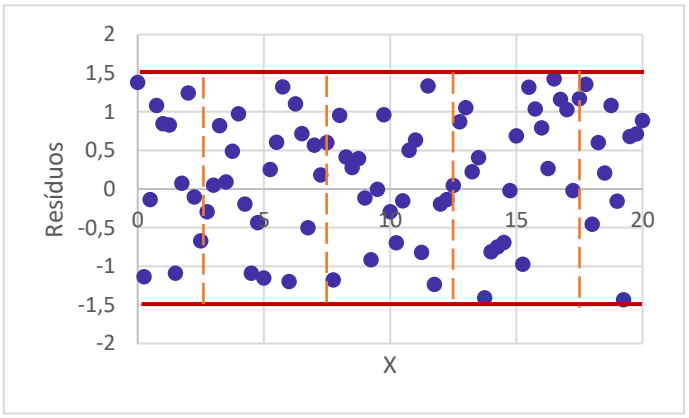


Igualdade de Variância dos Termos de Erro

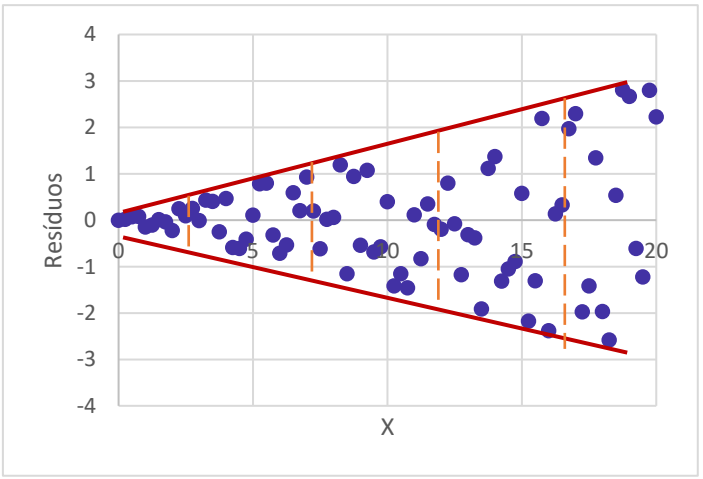
A variância dos erros deve ser consistente para todas as observações. Em outras palavras, **a variância não muda para cada observação ou para um intervalo de observações**. Esta condição preferida é conhecida como **homocedasticidade** (mesma dispersão). **Se a variância mudar**, nos referimos a isso como **heterocedasticidade** (dispersão diferente).

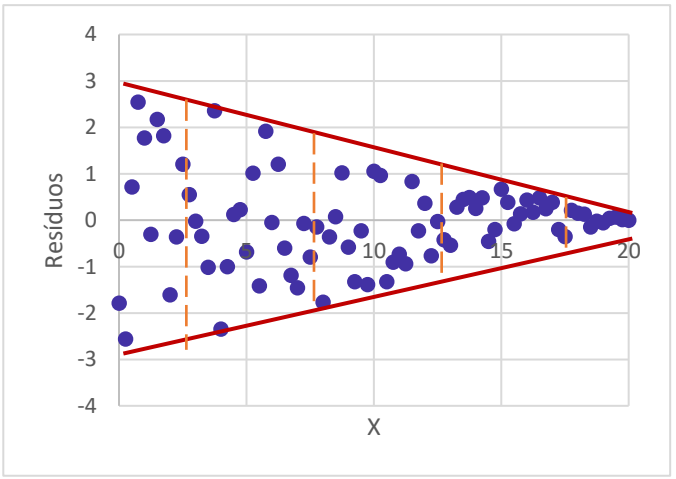
O diagnóstico é feito com gráficos de resíduos ou testes estatísticos simples. **A representação gráfica de resíduos versus os valores dependentes previstos e a sua comparação com o gráfico nulo mostra um padrão consistente se a variância não for constante**.

O gráfico a seguir não mostra nenhuma violação da igualdade de variância, porque os comprimentos das linhas tracejadas em laranja não são consideravelmente diferentes. O gráfico de resíduos dispersos contra X exibe uma forma retangular em torno de zero.

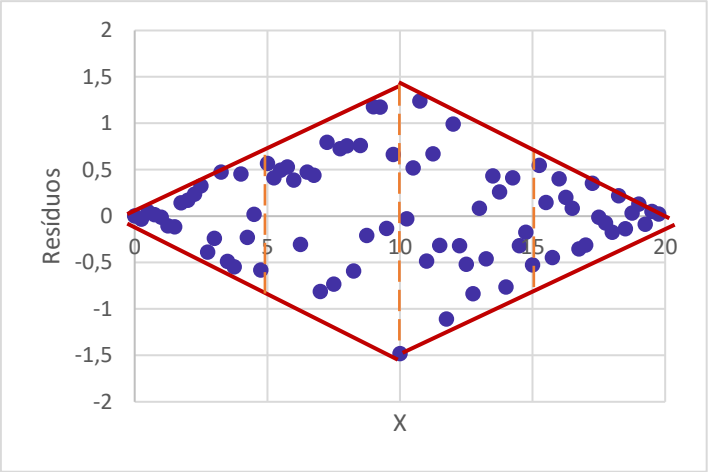


Nesse tipo de gráfico, a heterocedasticidade aparece como uma forma de triangular em que a dispersão dos resíduos aumenta em uma direção. No gráfico abaixo, a dispersão dos resíduos aumenta à medida que o valor ajustado aumenta.

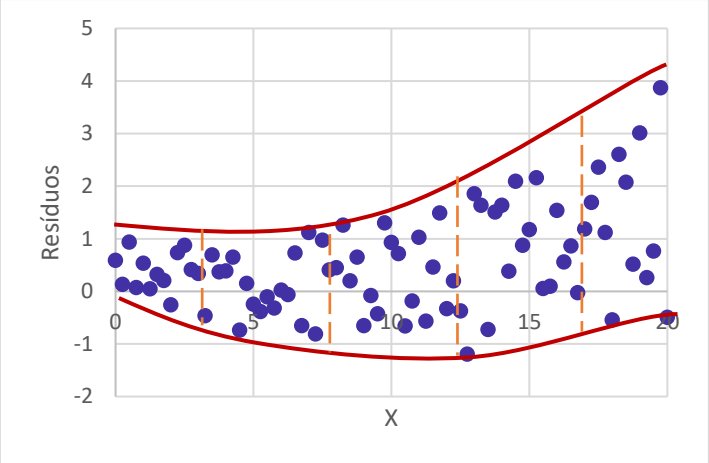




Um padrão em forma de diamante pode ser esperado no caso de percentagens nas quais se espera mais variação no meio do intervalo, em vez das bordas.



Muitas vezes, várias violações podem ocorrer simultaneamente, como não-linearidade e heteroscedasticidade.



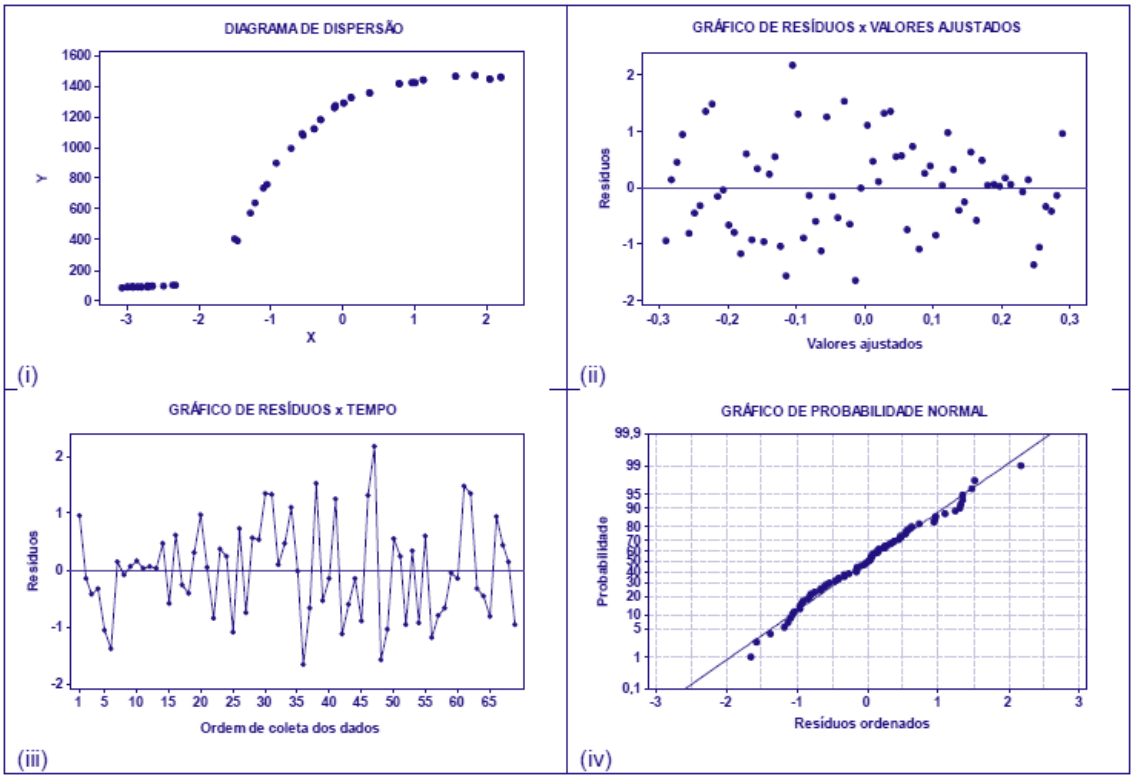
Há testes estatísticos para heteroscedasticidade. **O teste Levene para homogeneidade de variância mede a igualdade de variâncias para um par de variáveis, sendo recomendado porque é menos afetado por desvios da normalidade, outro problema comum em regressão**.

A heterocedasticidade pode ser corrigida de **duas formas**. Caso seja possível atribuir a violação a uma única variável independente por meio da análise de gráficos de resíduos, **o procedimento de mínimos quadrados ponderados (com pesos) pode ser empregado**.

Não obstante, **transformações de dados na resposta y podem ser usadas para eliminar o problema**. As transformações mais usadas para estabilizar a variância incluem o uso de $\sqrt{y}$, $ln(y)$ ou $1/y$ como a resposta.

HORA DE PRATICAR!

(FUMARC/PC MG/2013) No estudo da adequação de um modelo de regressão linear simples, foram obtidos os seguintes gráficos para análise dos resíduos:



Analisando os gráficos, é CORRETO afirmar que o modelo ajustado é inadequado porque

a) o gráfico (i) não satisfaz a suposição de que o relacionamento entre Y e X é linear.

b) o gráfico (ii) não satisfaz a suposição de que os erros são não correlacionados.

c) o gráfico (iii) não satisfaz a suposição de que a variância dos erros é constante.

d) o gráfico (iv) não satisfaz a suposição de que os erros seguem uma distribuição normal.

Comentários:

Vamos analisar cada uma das alternativas:

Alternativa A: **Correto**. Basta notarmos que o diagrama formado tem formato curvo, e não em linha reta.

Alternativa B: **Errado**. O gráfico dos resíduos versus valores ajustados (valores preditos) é uma das principais técnicas utilizadas para verificar as suposições dos resíduos. Além do diagnóstico de heterocedasticidade, esse gráfico pode indicar que existe uma relação não linear entre as variáveis explicativas com a variável resposta, por meio de alguma tendência nos pontos.

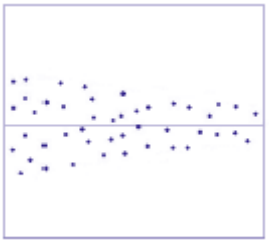
Alternativa C: **Errado**. O gráfico dos resíduos contra o tempo ou ordem de coleta é utilizado para uma suposição de independência dos erros. Ao avaliarmos o gráfico, percebemos uma tendência de os pontos se repetirem ao longo do tempo, o que indica a existência de dependência entre os resíduos.

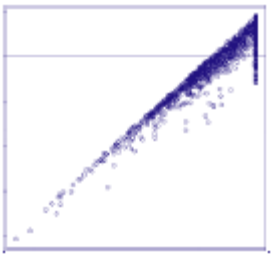
Alternativa D: **Errado**. Em um gráfico de probabilidade normal, quando os resíduos estão alinhados, a suposição de normalidade é válida; isto é, os resíduos possuem distribuição normal.

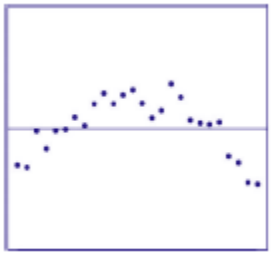
Gabarito: A.

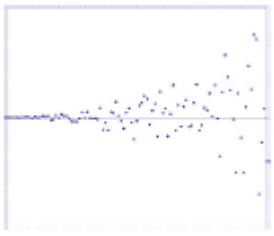
(FUNCERN/TJ RN/2016) A análise de resíduos é uma ferramenta útil para se avaliar a qualidade do ajuste de modelos de regressão aos dados em estudo.

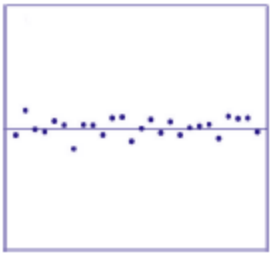
Assinale a opção que representa corretamente o gráfico de um modelo de regressão linear bem ajustado.

a)


b)


c)


d)


e) 

Comentários:

O gráfico dos resíduos versus valores ajustados (valores preditos) é uma das principais técnicas utilizadas para verificar as suposições dos resíduos. Além do diagnóstico de heterocedasticidade, esse gráfico permite verificar a existência de uma relação não linear entre as variáveis explicativas com a variável resposta.

Para a verificação da heteroscedasticidade, devemos procurar algum padrão ou tendência no gráfico de resíduos. Por isso, se os pontos estão aleatoriamente distribuídos em torno do 0, sem nenhum comportamento ou tendência, temos indícios de que a suposição de igualdade de variância foi respeitada. Já a presença de um padrão triangular indica a presença de heteroscedasticidade.

Finalmente, quando um modelo linear está bem ajustado, o gráfico de resíduos tende a se distribuir ao redor da origem.

Gabarito: E.

QUESTÕES COMENTADAS – CEBRASPE

Correlação Linear

1. (CESPE/FINEP/2024) Duas variáveis, X e Y, possuem a mesma variância; se a correlação linear de Pearson entre elas for 0,8, e se a covariância entre X e Y for 2, então a variância de X será

a) 2,50.

b) 1,25.

c) 0,20.

d) 2,00.

e) 0,40.

Comentários:

Inicialmente vamos organizar os dados do enunciado:

$\rho(X,Y) = 0{,}8$

$Cov(X,Y) = 2$

A correlação linear de Pearson é expressa por:

$$r = \frac{Cov(X,Y)}{\sqrt{Var(X) \times Var(Y)}}$$

Em que $Cov(X,Y)$,representa a covariância entre X e Y.

O enunciado nos disse que a variância de X é igual a variância de Y. Substituindo os valores dados no enunciado, temos:

$$0{,}8 = \frac{2}{\sqrt{Var(X) \times Var(Y)}}$$

$$0{,}8 = \frac{2}{Var(X)}$$

$$Var(X) \times 0{,}8 = 2$$

$$Var(X) = \frac{2}{0{,}8}$$

$$Var(X) = 2{,}5$$

Portanto, como a variância de X e Y são iguais:

$$Var(X) = Var(Y) = 2{,}5$$

Gabarito: A.

2. (CESPE/FUB/2022) Julgue o item subsequente, considerando oito pares de valores das variáveis X e Y, tais que $\Sigma X = 24$; $\Sigma Y = 49$; $\Sigma XY = 181$; $\Sigma X^2 = 100$ e $\Sigma Y^2 = 343$.

O coeficiente de correlação de Pearson para os valores apresentados será negativo, o que indica que a regressão linear será representada por uma reta decrescente.

Comentários:

Calculando o coeficiente de correlação linear de Pearson, verificamos que:

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

$$\rho(X,Y) = \frac{8 \times 181 - 24 \times 49}{\sqrt{8 \times 100 - 24^2}\sqrt{8 \times 343 - 49^2}}$$

$$\rho(X,Y) = \frac{1448 - 1176}{\sqrt{800 - 576}\sqrt{2744 - 2401}}$$

$$\rho(X,Y) = \frac{272}{\sqrt{224}\sqrt{343}}$$

$$\rho(X,Y) = \frac{272}{14{,}96 \times 18{,}52}$$

$$\rho(X,Y) = \frac{272}{277{,}06} = 0{,}98$$

Portanto, o coeficiente de correlação de Pearson para os valores apresentados será positivo.

Gabarito: Errado.

3. (CESPE/FUB/2022) Julgue o item subsequente, considerando oito pares de valores das variáveis X e Y, tais que $\Sigma X = 24$; $\Sigma Y = 49$; $\Sigma XY = 181$; $\Sigma X^2 = 100$ e $\Sigma Y^2 = 343$.

Existe uma correlação forte entre as variáveis X e Y.

Comentários:

Calculando o coeficiente de correlação linear de Pearson, verificamos que:

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

$$\rho(X,Y) = \frac{8 \times 181 - 24 \times 49}{\sqrt{8 \times 100 - 24^2}\sqrt{8 \times 343 - 49^2}}$$

$$\rho(X,Y) = \frac{1448 - 1176}{\sqrt{800 - 576}\sqrt{2744 - 2401}}$$

$$\rho(X,Y) = \frac{272}{\sqrt{224}\sqrt{343}}$$

$$\rho(X,Y) = \frac{272}{14{,}96 \times 18{,}52}$$

$$\rho(X,Y) = \frac{272}{277{,}06} = 0{,}98$$

Portanto, existe uma correlação forte (muito próxima de 1) entre as variáveis X e Y.

Gabarito: Certo.

4. (CESPE/FUB/2022) Julgue o item subsequente, considerando oito pares de valores das variáveis X e Y, tais que $\Sigma X = 24$; $\Sigma Y = 49$; $\Sigma XY = 181$; $\Sigma X^2 = 100$ e $\Sigma Y^2 = 343$.

Com base no coeficiente de correlação linear, é correto afirmar, em face dos dados apresentados, que se trata de uma correlação espúria.

Comentários:

A correlação espúria normalmente ocorre quando duas variáveis não relacionadas são percebidas como se fossem, seja pelo acaso ou por um terceiro fator. Basicamente, a correlação espúria pode ser definida como uma associação entre variáveis que, na verdade, não apresentam uma relação natural.

O presente caso não é de correlação espúria, pois as variáveis apresentam uma correlação forte entre si, conforme verificamos a seguir:

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

$$\rho(X,Y) = \frac{8 \times 181 - 24 \times 49}{\sqrt{8 \times 100 - 24^2}\sqrt{8 \times 343 - 49^2}}$$

$$\rho(X,Y) = \frac{1448 - 1176}{\sqrt{800 - 576}\sqrt{2744 - 2401}}$$

$$\rho(X,Y) = \frac{272}{\sqrt{224}\sqrt{343}}$$

$$\rho(X,Y) = \frac{272}{14{,}96 \times 18{,}52}$$

$$\rho(X,Y) = \frac{272}{277{,}06} = 0{,}98$$

Gabarito: Errado.

5. (CESPE/FUB/2022) Uma regressão linear de Y sobre X consiste em obter a equação de uma reta, ou uma função linear, como o modelo que irá melhor representar a relação entre as variáveis; a determinação dos parâmetros dessa reta é denominada ajustamento.

Considerando essas informações, julgue o seguinte item.

Para quaisquer valores das variáveis X e Y, a existência de um coeficiente de correlação diferente de zero é garantia para que haja uma relação entre X e Y.

Comentários:

O coeficiente de correlação mede a relação estatística entre duas variáveis, isto é, mede a interdependência entre duas variáveis. Contudo, mesmo eventos sem relação de causa podem ter alto grau de correlação. Quando isso ocorre, dizemos tratar-se de uma relação espúria. Há alguns exemplos famosos, como é o caso da correlação de 98% entre a redução no consumo de margarina por pessoa e da queda no número de divórcios por mil pessoas no Estado do Maine (EUA).

Gabarito: Errado.

6. (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.

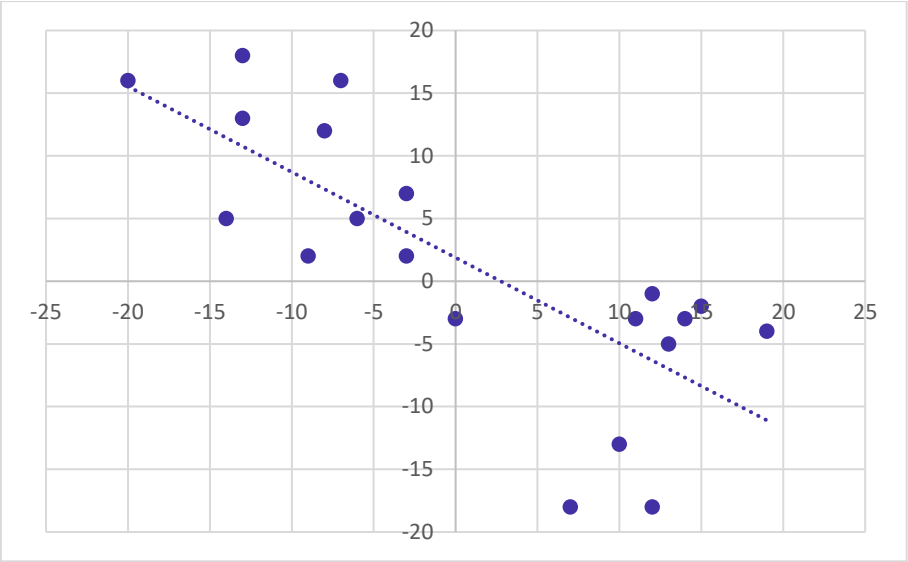
d) Poderá ser fracamente negativa.

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.

7. (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:

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

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

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.

8. (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:

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

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

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

Comentários:

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

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

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:

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

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.

9. (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.

10. (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) $0{,}38 < X < 0{,}39$.

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.

QUESTÕES COMENTADAS – CEBRASPE

Regressão Linear Simples

1. (CESPE/ANAC/2024) Mediante a aplicação do critério de mínimos quadrados ordinários, um analista deseja ajustar um modelo de regressão linear simples na forma y = a + bx + ε, com variância V, em que y representa a variável dependente, x é a variável regressora e ε denota um erro aleatório que segue distribuição normal com média zero. A partir de uma amostra aleatória simples de tamanho n = 46, o analista obteve as estatísticas descritivas mostradas na tabela a seguir.

Variável	Média Amostral	Desvio Padrão Amostral
Y	10	5
x	20	4

A partir dessas informações, e sabendo que a correlação linear de Pearson entre as variáveis y e x é igual a 0,5, julgue os próximos itens.

A estimativa do coeficiente b é igual ou superior a 0,6.

Comentários:

Para encontrar a estimativa do coeficiente $\hat{b}$ na regressão linear simples, podemos usar a fórmula:

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

Onde:

- r é a correlação linear de Pearson entre y e x (dado como 0,5);
- s_y é o desvio padrão amostral de y (dado como 5);
- s_x é o desvio padrão amostral de x (dado como 4).

Substituindo os valores conhecidos na fórmula:

$$\hat{b} = 0{,}5 \times \frac{5}{4} = 5 \times 1{,}25 = 0{,}625$$

Portanto, a estimativa do coeficiente $\hat{b}$ é de 0,625. Assim, a afirmação de que a estimativa do coeficiente $\hat{b}$ é igual ou superior a 0,6 é verdadeira.

Gabarito: Certo.

2. (CESPE/FINEP/2024)

x	0	1	2	3	4	5
y	0	1	3	13	14	25

Considerando que a tabela precedente exibe uma amostra aleatória bivariada (x,y) de tamanho 6, na qual representa uma variável dependente e denota uma variável regressora, assinale a opção que apresenta uma curva de regressão $(\hat{y})$ ajustada para esse conjunto de dados mediante aplicação do método de mínimos quadrados ordinários.

a) $\hat{y} = x^2$

b) $\hat{y} = 11{,}2$

c) $\hat{y} = 3{,}73x$

d) $\hat{y} = 0{,}5x^2 + 12{,}5$

e) $\hat{y} = 5x - 2$

Comentários:

O objetivo do método dos mínimos quadrados é minimizar o somatório dos quadrados dos erros $(\sum_{i=1}^{n} e_i^2)$. O enunciado pede para assinalarmos, dentre as alternativas, a opção que apresenta uma curva de regressão $(\hat{y})$ ajustada para esse conjunto de dados. Sendo assim, montaremos uma tabela para verificar qual alternativa minimiza o desvio quadrático:

X	Y	$\hat{Y} = X^2$	$\hat{Y} = 11,2$	$\hat{Y} = 3,73X$	$\hat{Y} = 0,5X^2 + 12,5$	$\hat{Y} = 5X - 2$
0	0	0	11,2	0	12,5	-2
1	1	1	11,2	3,73	13	3
2	3	4	11,2	7,46	14,5	8
3	13	9	11,2	11,19	17	13
4	14	16	11,2	14,92	20,5	18
5	25	25	11,2	18,65	25	23
$\bar{X} = 2,5$	$\bar{Y} = 9,33$					

Agora, calculando os desvios quadráticos para cada alternativa, temos que:

X	Y	$(Y - X^2)^2$	$(Y - 11,2)^2$	$(Y - 3,73X)^2$	$(Y - 0,5X^2 - 12,5)^2$	$(Y - 5X + 2)^2$
0	0	0	125,44	0	156,25	4
1	1	0	104,04	7,45	144	4
2	3	1	67,24	19,89	132,25	25
3	13	16	3,24	3,27	16	0
4	14	4	7,84	0,84	42,25	16
5	25	0	190,44	40,32	0	4
TOTAL		**21**	**498,24**	**71,78**	**490,75**	**53**

Portanto, dentre as opções apresentadas, a que minimiza o somatório do erro quadrático é a letra A.

Gabarito: A.

3. (CESPE/FUB/2022) Julgue o item subsequente, considerando oito pares de valores das variáveis X e Y, tais que $\Sigma X = 24$; $\Sigma Y = 49$; $\Sigma XY = 181$; $\Sigma X^2 = 100$ e $\Sigma Y^2 = 343$.

A reta dos mínimos quadrados ordinários que representa a regressão linear simples de Y em X com intercepto não nulo terá coeficiente linear aproximado de 2,48.

Comentários:

Em uma regressão linear na forma $Y_i = a + bX_i + \varepsilon_i$., os estimadores mínimos quadrados são dados pelas relações:

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

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

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

Vamos utilizar as fórmulas acima para calcular o coeficiente angular da reta:

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

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(X_i Y_i) - n \times \frac{\sum_{i=1}^{n} X_i}{n} \times \frac{\sum_{i=1}^{n} Y_i}{n}}{\sum_{i=1}^{n}(X_i^2) - n \times \left(\frac{\sum_{i=1}^{n} X_i}{n}\right)^2}$$

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(X_i Y_i) - \frac{\sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i}{n}}{\sum_{i=1}^{n}(X_i^2) - \frac{(\sum_{i=1}^{n} X_i)^2}{n}}$$

$$\hat{\beta} = \frac{181 - \frac{24 \times 49}{8}}{100 - \frac{24^2}{8}}$$

$$\hat{\beta} = \frac{181 - 3 \times 49}{100 - 3 \times 24}$$

$$\hat{\beta} = \frac{181 - 147}{100 - 72}$$

$$\hat{\beta} = \frac{34}{28}$$

Agora, vamos calcular o coeficiente linear:

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

$$\hat{\alpha} = \frac{\sum_{i=1}^{n} Y_i}{n} - \hat{\beta} \times \frac{\sum_{i=1}^{n} X_i}{n}$$

$$\hat{\alpha} = \frac{49}{8} - \frac{34}{28} \times \frac{24}{8}$$

$$\hat{\alpha} = \frac{49 \times 28 - 34 \times 24}{8 \times 28}$$

$$\hat{\alpha} = \frac{1372 - 816}{224} = \frac{556}{224} = 2{,}48$$

Gabarito: Certo.

4. (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.

5. (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.

6. (CESPE/PETROBRAS/2022)

Equação 1: $y_i = a + bX_1 + e$

Equação 2: $y_i = a + b_1X_1 + b_2X_2 + b_3X_3 + e$

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.

7. (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.

8. (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.

9. (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.

10. (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.

11. (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.

12. (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.

13. (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.

14. (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.

15. (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.

16. (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.

17. (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.

18. (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.

19. (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.

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.

20. (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, $\hat{S}_e = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$, 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.

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.

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.

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.

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 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.

23. (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.

24. (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.

25. (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 homoscedastiscidade 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 homoscedastiscidade. Portanto, a questão erra ao incluir o valor esperado dos erros nesse conceito.

Gabarito: Errado.

26. (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):

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

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.

27. (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.

28. (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.

29. (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.

30. (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 $\sum_i (x_i - \bar{x})^2 = 10$, em que $\bar{x} = (x_1 + \ldots + x_n)/n$, 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.

31. (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.

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.

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.

33. (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.

34. (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.

QUESTÕES COMENTADAS – CEBRASPE

Análise de Variância da Regressão

1. (CESPE/ANAC/2024) Mediante a aplicação do critério de mínimos quadrados ordinários, um analista deseja ajustar um modelo de regressão linear simples na forma y = a + bx + ε, com variância V, em que y representa a variável dependente, x é a variável regressora e ε denota um erro aleatório que segue distribuição normal com média zero. A partir de uma amostra aleatória simples de tamanho n = 46, o analista obteve as estatísticas descritivas mostradas na tabela a seguir.

Variável	Média Amostral	Desvio Padrão Amostral
Y	10	5
x	20	4

A partir dessas informações, e sabendo que a correlação linear de Pearson entre as variáveis y e x é igual a 0,5, julgue os próximos itens.

Estima-se que a variância V seja inferior a 15.

Comentários:

A questão quer o valor da variância do modelo. Para descobrir esse valor, vamos precisar dos conceitos vistos em análise de variância da regressão. Começaremos desenvolvendo o desvio padrão amostral da seguinte forma:

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

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

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

Como $SQT = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$, temos que:

$$(n-1) \times s_y^2 = SQT$$

$$SQT = (6-1) \times (5)^2$$

$$SQT = 5 \times (5)^2 = 125$$

A questão também nos forneceu o valor do coeficiente de correlação de Pearson. Sendo assim, podemos usar a relação abaixo para encontrar a soma dos quadrados dos erros:

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

$$SQM = R^2 \times SQT$$

$$SQM = 0{,}25 \times 125 = 31{,}25$$

O valor da variância do modelo (V) é igual ao quadrado médio do modelo (QMM):

$$\sigma^2 = QMM = \frac{SQM}{1} = \frac{31{,}25}{1} = 31{,}25$$

Gabarito: Errado.

2. (CESPE/ANAC/2024) Mediante a aplicação do critério de mínimos quadrados ordinários, um analista deseja ajustar um modelo de regressão linear simples na forma y = a + bx + ε, com variância V, em que y representa a variável dependente, x é a variável regressora e ε denota um erro aleatório que segue distribuição normal com média zero. A partir de uma amostra aleatória simples de tamanho n = 46, o analista obteve as estatísticas descritivas mostradas na tabela a seguir.

Variável	Média Amostral	Desvio Padrão Amostral
Y	10	5
x	20	4

A partir dessas informações, e sabendo que a correlação linear de Pearson entre as variáveis y e x é igual a 0,5, julgue os próximos itens.

50% da variação total de y é explicada por meio do modelo de regressão linear simples em questão.

Comentários:

Para determinar se 50% da variação total de y é explicada pelo modelo de regressão linear simples, podemos usar o coeficiente de determinação R^2.

O coeficiente de determinação é dado por:

$$R^2 = \frac{variância\ explicada}{variância\ total}$$

Onde:

- A variância explicada é a variância de y prevista pelo modelo de regressão.
- A variância total é a variância de y.

Dado que a correlação linear de Pearson entre as variáveis y e x é 0,5, o coeficiente de determinação R^2 pode ser calculado como o quadrado dessa correlação:

$$R^2 = (0{,}5)^2 = 0{,}25$$

Portanto, 25% da variação total de y é explicada pelo modelo de regressão linear simples.

Para expressar isso em termos percentuais, precisamos multiplicar por 100:

$$0{,}25 \times 100\% = 25\%$$

Assim, apenas 25% da variação total de y é explicada pelo modelo de regressão linear simples. Portanto, a afirmação de que 50% da variação total de y é explicada pelo modelo está incorreta.

Gabarito: Errado.

3. (CESPE/FUB/2022) Uma regressão linear de Y sobre X consiste em obter a equação de uma reta, ou uma função linear, como o modelo que irá melhor representar a relação entre as variáveis; a determinação dos parâmetros dessa reta é denominada ajustamento.

Considerando essas informações, julgue o seguinte item.

Suponha-se que, em uma pesquisa, o coeficiente de correlação entre duas variáveis X e Y tenha gerado um valor para o coeficiente de correlação de Pearson de 0,9200. Nesse caso, considerando-se X a variável independente e Y a variável dependente, o percentual da variância de Y explicado por X será de 84,64%.

Comentários:

O coeficiente de determinação R^2 mede a variabilidade da variável dependente (Y) explicada pela variável independente (X). Como $R^2 = (0{,}92)^2 = 84{,}64\%$, podemos concluir que o percentual da variância de Y explicado por X será de, aproximadamente, 84,64%.

Gabarito: Certo.

4. (CESPE/FUB/2022) Uma regressão linear de Y sobre X consiste em obter a equação de uma reta, ou uma função linear, como o modelo que irá melhor representar a relação entre as variáveis; a determinação dos parâmetros dessa reta é denominada ajustamento.

Considerando essas informações, julgue o seguinte item.

Um coeficiente de determinação entre as variáveis X e Y de 95% implica necessariamente a obtenção de uma reta dos mínimos quadrados crescente, ou seja, em uma correlação positiva.

Comentários:

O coeficiente de determinação mede a variabilidade da variável dependente (Y) explicada pela variável independente (X), sendo expresso por R^2. Se o coeficiente de correlação for negativo e igual a $-0{,}974$, nosso coeficiente de determinação será 95%. Logo, um coeficiente de determinação de 95% **não implica** necessariamente na obtenção de uma reta dos mínimos quadrados crescente.

Gabarito: Errado.

5. (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 de correlação entre as variáveis X e Y também pode ser expresso da seguinte forma:

$$\rho(X,Y) = \frac{Cov(X,Y)}{s_X \times s_Y}$$

Como o enunciado nos disse que $\rho(X,Y) = 0{,}8$, $s_Y = 2$ e $s_X = 4$, podemos deduzir que:

$$\rho(X,Y) = \frac{Cov(X,Y)}{s_X \times s_Y}$$

$$0{,}8 = \frac{Cov(X,Y)}{2 \times 4}$$

$$Cov(X,Y) = 6{,}4$$

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

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

$$\hat{\beta} = \frac{6{,}4}{4^2}$$

$$\hat{\beta} = 0{,}4.$$

Gabarito: A.

6. (CESPE/POLITEC RO/2022) Em relação aos procedimentos técnicos relacionados aos procedimentos de amostragem, julgue os itens a seguir.

I. Quando se adiciona variáveis explicativas no modelo de regressão linear, espera-se o incremento da estatística R^2.

II. Ao se comparar modelos com diferentes quantidades de variáveis explicativas, deve-se analisar o valor de R^2 ajustado.

III. O aumento de variáveis explicativas aumenta o R^2 ajustado.

IV. Ao se estimar um modelo com quatro variáveis explicativas e compará-lo com um modelo com três variáveis explicativas, escolhe-se o modelo que retornar o maior valor de R^2 ajustado, tudo o mais constante.

Estão corretos apenas os itens

a) I e II.

b) I e III.

c) I e IV.

d) II e III.

e) III e IV.

Comentários:

O coeficiente de determinação (R^2) mede a variabilidade total da variável dependente (Y) explicada pela variável independente (X). Esse coeficiente assume valores no intervalo de 0 a 1, ficando mais próximo de 1 quanto melhor o modelo estiver ajustado ao conjunto de dados.

O coeficiente ajustado possui a característica de penalizar a inclusão de variáveis sem poder explicativo, fazendo seu valor diminuir. Assim, ao adicionarmos uma variável independente com baixo poder explicativo, haverá uma redução no valor do R^2 ajustado, o que não ocorre no R^2 simples.

Portanto, ao compararmos modelos de regressão múltipla com diferentes quantidades de variáveis explicativas, a comparação não pode ser realizada apenas com base no coeficiente de determinação (R^2), pois ele aumenta conforme a inclusão de novas variáveis explicativas.

Agora, vamos analisar cada item:

Item I. **Correto**. De fato, ao se adicionar variáveis explicativas ao modelo de regressão linear, espera-se o incremento da estatística R^2.

Item II. **Correto**. Ao se comparar modelos com diferentes quantidades de variáveis explicativas, deve-se analisar o valor de R^2 ajustado.

Item III. **Incorreto**. O aumento de variáveis explicativas ~~aumenta~~ o R^2 ajustado. Não, o aumento de variáveis explicativas diminui o R^2 ajustado.

Item IV. **Incorreto**. Não necessariamente. Digamos que o R^2 ajustado sofra uma variação positiva muito pequena após a inclusão da nova variável. Nesse caso, embora o R^2 ajustado tenha sofrido um aumento, a nova variável não se mostra muito útil para explicar a variável dependente, pois o ganho com a sua inclusão é muito pequeno. Então, o R^2 ajustado ajuda a identificar quando variáveis adicionais estão contribuindo para o modelo.

Gabarito: A.

7. (CESPE/SECONT ES/2022) Com base no modelo clássico de regressão linear, julgue o item a seguir.

Em se tratando do modelo de regressão múltipla, ao se compararem modelos com diferentes quantidades de variáveis explicativas, o correto é analisar o valor de R^2 *ajustado.*

Comentários:

O coeficiente de determinação (R^2) mede a variabilidade total da variável dependente (Y) explicada pela variável independente (X). Esse coeficiente assume valores no intervalo de 0 a 1, ficando mais próximo de 1 quanto melhor o modelo estiver ajustado ao conjunto de dados.

O coeficiente ajustado possui a característica de penalizar a inclusão de variáveis sem poder explicativo, fazendo seu valor diminuir. Assim, ao adicionarmos uma variável independente com baixo poder explicativo, haverá uma redução no valor do R^2 ajustado, o que não ocorre no R^2 simples.

Portanto, ao compararmos modelos de regressão múltipla com diferentes quantidades de variáveis explicativas, a comparação não pode ser realizada apenas com base no coeficiente de determinação (R^2), pois ele aumenta conforme a inclusão de novas variáveis explicativas. Por isso, o correto é fazer essa comparação por meio do R^2 ajustado.

Gabarito: Certo.

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.

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.

9. (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.

10. (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.

11. (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.

12. (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.

A soma de quadrados dos resíduos é igual ou inferior a 76.

Comentários:

Com os dados fornecidos pelo enunciado, podemos utilizar o coeficiente de correlação para calcular a soma dos quadrados dos resíduos:

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

Portanto,

$$0{,}9^2 = 1 - \frac{SQR}{SQT}$$

$$0{,}81 = 1 - \frac{SQR}{SQT}$$

Como foi fornecida a variância amostral de Y, podemos chegar na soma dos quadrados total por meio da expressão:

$$s_Y^2 = \frac{SQT}{(n-1)}$$

$$SQT = (n-1) \times s_Y^2$$

Agora, jogando na expressão anterior, obtemos que:

$$0{,}81 - 1 = \frac{-SQR}{(100-1) \times 4^2}$$

$$-0{,}19 = \frac{-SQR}{99 \times 16}$$

$$0{,}19 = \frac{SQR}{1584}$$

$$SQR = 1584 \times 0{,}19$$

$$SQR = 300{,}96$$

Gabarito: Errado.

13. (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.

A estimativa de mínimos quadrados ordinários para o intercepto do modelo é igual a zero.

Comentários:

Nessa questão, temos que iniciar calculando o desvio padrão amostral da variável X. A partir dessa informação, vamos conseguir calcular a covariância entre as duas variáveis. Vejamos:

X	Freq. (f_i)	$X - \bar{X}$	$(X - \bar{X})^2$	$(X - \bar{X})^2 \times f_i$
0	55	0 - 0,45 = -0,45	$(-0,45)^2 = 0,2025$	0,2025 x 55 = 11,1375
1	45	1 - 0,45 = 0,55	$0,55^2 = 0,3025$	0,3025 x 45 = 13,8375
Total	**100**		**0,505**	**24,75**

Assim, temos que:

$$s_X = \sqrt{\frac{(X - \bar{X})^2 \times f_i}{n - 1}} = \sqrt{\frac{24,75}{99}} = \sqrt{0,25} = 0,5$$

O coeficiente de correlação entre as variáveis X e Y também pode ser expresso da seguinte forma:

$$\rho(X, Y) = \frac{Cov(X, Y)}{s_X \times s_Y}$$

Como o enunciado nos disse que $\rho(X, Y) = 0,9$ e $s_Y = 4$, podemos deduzir que:

$$\rho(X, Y) = \frac{Cov(X, Y)}{s_X \times s_Y}$$

$$0,9 = \frac{Cov(X, Y)}{0,5 \times 4}$$

$$Cov(X, Y) = 1,8$$

A partir da covariância, podemos encontrar o coeficiente angular (a) estimado pelo método dos mínimos quadrados:

$$\hat{a} = \frac{Cov(X,Y)}{s_X^2}$$

$$\hat{a} = \frac{1{,}8}{(0{,}5)^2}$$

$$\hat{a} = 7{,}2.$$

Agora, 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,

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

$$10 = 7{,}2 \times 0{,}45 + \hat{b}$$

$$\hat{b} = 10 - 3{,}24 = 6{,}76.$$

Gabarito: Errado.

14. (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.

Se $\hat{a}$ denota a estimativa de mínimos quadrados ordinários do coeficiente angular a, então $\hat{a} = 7{,}2$.

Comentários:

Nessa questão, temos que iniciar calculando o desvio padrão amostral da variável X. A partir dessa informação, vamos conseguir calcular a covariância entre as duas variáveis. Vejamos:

X	Freq. (f_i)	$X - \bar{X}$	$(X - \bar{X})^2$	$(X - \bar{X})^2 \times f_i$
0	55	0 - 0,45 = -0,45	$(-0,45)^2 = 0,2025$	0,2025 x 55 = 11,1375
1	45	1 - 0,45 = 0,55	$0,55^2 = 0,3025$	0,3025 x 45 = 13,8375
Total	**100**		**0,505**	**24,75**

Assim, temos que:

$$s_X = \sqrt{\frac{(X - \bar{X})^2 \times f_i}{n - 1}} = \sqrt{\frac{24,75}{99}} = \sqrt{0,25} = 0,5$$

O coeficiente de correlação entre as variáveis X e Y também pode ser expresso da seguinte forma:

$$\rho(X, Y) = \frac{Cov(X, Y)}{s_X \times s_Y}$$

Como o enunciado nos disse que $\rho(X, Y) = 0,9$ e $s_Y = 4$, podemos deduzir que:

$$\rho(X, Y) = \frac{Cov(X, Y)}{s_X \times s_Y}$$

$$0,9 = \frac{Cov(X, Y)}{0,5 \times 4}$$

$$Cov(X, Y) = 1,8$$

A partir da covariância, podemos encontrar o coeficiente angular (a) estimado pelo método dos mínimos quadrados:

$$\hat{a} = \frac{Cov(X, Y)}{s_X^2}$$

$$\hat{a} = \frac{1,8}{(0,5)^2}$$

$$\hat{a} = 7,2.$$

Gabarito: Certo.

15. (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.

16. (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.

17. (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.

18. (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.

19. (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.

20. (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.

21. (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.

22. (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.

23. (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.

24. (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.

25. (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.

26. (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.

27. (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.

28. (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.

29. (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.

30. (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.

31. (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.

32. (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.

33. (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.

34. (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.

35. (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.

36. (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.

37. (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.

38. (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.

QUESTÕES COMENTADAS – CEBRASPE

Análise de Resíduos

1. (CESPE/CAPES/2024) Um modelo de regressão linear entre uma variável aleatória Y (dependente) e uma variável não aleatória X (independente) é definido por $Y = \beta_0 + \beta_1 X + \epsilon$ em que ϵ, denominado erro aleatório, é uma variável aleatória independente de X com média $E(\epsilon) = 0$ e desvio padrão $Var(\epsilon) = \sigma^2$. Um modelo de regressão linear é essencialmente um modelo para a probabilidade condicional de Y com relação a X, denotada por $P(Y|X)$ ele é chamado de simples se n for uma variável aleatória gaussiana. Fixando-se valores $X_1, X_2, \cdots, X_n$ para a variável independente X, pode-se definir n variáveis aleatórias $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$, com $i = 1, \cdots, n$. Pelo método dos mínimos quadrados, é possível obter estimadores $\widehat{\beta}_0$ e $\widehat{\beta}_1$ para os parâmetros β_0 e β_1 e definir o $Y_i = \beta_0 + \beta_1 X_i$ como o estimador para Y_i.Nesse contexto, são definidos os erros, denominados resíduos, como $Y_i - \widehat{Y}_i = e_i$, a soma dos quadrados dos resíduos $SQE = \sum_i {e_i}^2$ a soma dos quadrados totais $SQT = \sum_i (Y_i - \overline{Y})^2$, e a soma dos quadrados de regressão $SQR = \sum_i (\widehat{Y}_i - \overline{Y})^2$, com $\overline{Y} = \sum_i Y_i / n$.

Com base nessas informações, julgue os próximos itens, considerando uma variável T, com média nula e desvio padrão unitário, definida por uma distribuição t de Student com 30 graus de liberdade, que tenha o seguinte intervalo com probabilidade de $0,95$: $P(-2,042 < T < 2,042) = 0,95$.

Para um modelo de regressão que não seja simples, a soma dos resíduos não será nula.

Comentários:

Em um modelo de regressão linear múltipla, o resíduo r_i é definido como a diferença entre o valor observado y_i e o valor estimado $\hat{y}_i$. Ou seja:

$$r_i = y_i - \hat{y}_i$$

O valor esperado do resíduo, $E(r_i)$, representa a média dos resíduos ao longo de todas as observações. Matematicamente, podemos definir o valor esperado do resíduo da seguinte maneira:

$$E(r_i) = E(y_i - \hat{y})$$

Devido à forma como os resíduos são calculados (diferença entre valores observados e estimados), a média dos resíduos tende a ser igual a zero. Isso significa que, em média, os resíduos se cancelam, não havendo viés na estimativa. Logo, a soma dos resíduos será nula:

$$E(r_i) = 0$$

Gabarito: Errado.

2. (CESPE/TCE-PA/2016)

Fonte de Variação	Grau 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 hipótese de normalidade exigida pelo modelo pode ser verificada a partir do gráfico dos resíduos, apesar de ser importante fazer um teste estatístico para tal fim.

Comentários:

De fato, a hipótese de normalidade pode ser verificada a partir do gráfico dos resíduos. Para isso, podemos representar os resíduos como histogramas ou como gráficos de probabilidade normal.

No gráfico de probabilidade normal, cada termo de erro é representado em função de sua respectiva probabilidade acumulada. A escala é definida de modo que, se a distribuição fosse perfeitamente normal, o gráfico produziria uma reta diagonal. Dessa forma, basta avaliar como os dados reais se comportam em relação a essa reta.

Além disso, existem testes estatísticos específicos para avaliar a normalidade. Como exemplos, podemos citar os testes de Shapiro-Wilk e Kolmogorov-Smirnov.

Gabarito: Certo.

3. (CESPE/FUB/2013) A respeito dos métodos de análise de resíduos do modelo de regressão, julgue o item subsequente.

A suposição de homocedasticidade pode ser verificada através de um gráfico de resíduos.

Comentários:

Correto. Se o gráfico de resíduos não evidencia alguma tendência de crescimento nos resíduos, significa que a suposição de igualdade de variância (homocedasticidade) está sendo respeitada. Por outro lado, se existe

uma tendência de crescimento em alguma direção, normalmente em padrão triangular, quer dizer que a suposição de homocedasticidade está sendo violada (portanto, há heterocedasticidade).

Gabarito: Certo.

4. (CESPE/FUB/2013) A respeito dos métodos de análise de resíduos do modelo de regressão, julgue o item subsequente.

Na análise de resíduos de um modelo de regressão, o diagrama de dispersão entre os resíduos do modelo ajustado e os valores preditos para a variável resposta permitem avaliar a ocorrência de heterocedasticidade.

Comentários:

Correto. Se o gráfico de resíduos evidenciar a existência de um padrão de crescimento nos resíduos, normalmente em formato triangular, quer dizer que a suposição de igualdade de variância está sendo violada. Portanto, está ocorrendo heterocedasticidade.

Gabarito: Certo.

5. (CESPE/MCom/2013)

FV	gl	SQ	QM	F
regressão			810	
resíduo	98			
total			80	

O quadro acima mostra parte de uma tabela de análise de variância (ANOVA), que resultou da regressão linear simples do tempo que um usuário permanece conectado à Internet 3G — Y, em minutos — sobre a renda — X — declarada por esse usuário. Os dados utilizados nesse ajuste pelo método de mínimos quadrados ordinários foram selecionados por amostragem aleatória simples de um cadastro de usuários. Com base nessas informações e no quadro apresentado, julgue o item seguinte acerca dos conceitos de análise de regressão, correlação e amostragem.

O gráfico dos resíduos permite diagnosticar a hipótese de sua normalidade, ou seja, com base nesse gráfico, é possível efetuar uma análise confirmatória.

Comentários:

O gráfico de resíduos realmente permite diagnosticar a hipótese de sua normalidade, mas faz isso por meio de **análise exploratória**.

Gabarito: Errado.

6. (CESPE/ALECE/2011)

Fonte de variação	Graus de liberdade	Soma dos quadrados	Média dos quadrados	Razão F
regressão	1	2.061,49	2.061,49	433,40
erro	78	371,01	4,75	
total	79	2.432,50		

Um analista deseja avaliar se o tempo — Y —, em dias, que um processo judicial leva para ser concluído está relacionado com a quantidade — X — de juízes disponíveis no tribunal em que tal processo foi julgado. O quadro acima apresenta a tabela de análise de variância (ANOVA) correspondente a essa avaliação por regressão linear simples, em que Y é a variável resposta e X é a variável regressora, com base no método de mínimos quadrados ordinários. Considerando essas informações e os conceitos de análise de regressão linear e inferência estatística, julgue o item.

Uma ferramenta descritiva para avaliação e diagnóstico do modelo é o gráfico de resíduos. Nesse gráfico, os resíduos devem apresentar-se dispostos aleatoriamente em torno do ponto zero.

Comentários:

O gráfico de resíduos realmente é uma ferramenta descritiva para avaliação e diagnóstico do modelo. Além disso, para caracterizar a situação ideal, em que todas as suposições são respeitadas, os resíduos devem se apresentar dispostos aleatoriamente em torno do ponto zero.

Gabarito: Certo.

7. (CESPE/PF/2004) Entre janeiro e novembro de 2003, foi realizado um estudo para avaliar o número mensal de ocorrências, por 1.000 habitantes, registradas em delegacias de determinada região. Para esse estudo, foi considerado o modelo de regressão linear simples na forma $Y=a+\beta X+\epsilon$, em que X é uma variável que representa os meses e assume valores discretos 0, 1, 2, ..., 10, e Y representa o número de ocorrências por 1.000 habitantes registradas no respectivo mês X. Parte do objetivo desse estudo é estimar os coeficientes a e β. O erro aleatório é representado por ϵ.

As tabelas abaixo apresentam parte dos resultados do ajuste e da análise de variância.

coeficiente	estimativa de mínimos quadrados ordinários	erro-padrão
α	50	0,05
β	0,05	0,005

fonte de variação	graus de liberdade	soma dos quadrados	quadrado médio
modelo	1	0,3	*D*
erro	9	*B*	*E*
total	*A*	*C*	*F*

Com base no texto acima, julgue o item a seguir.

Considere que, na análise dos resíduos, o estudo verificou que Y segue uma distribuição normal. Nessa situação, conclui-se que os dados são heterocedásticos.

Comentários:

O fato de a distribuição ser caracterizada como normal (suposição de normalidade) em nada tem a ver com a suposição de igualdade de variância (homocedasticidade). Por isso, o estudo verificar que Y segue uma distribuição normal não é suficiente para concluirmos que os dados são heterocedásticos.

Gabarito: Errado.

LISTA DE QUESTÕES – CEBRASPE

Correlação Linear

1. (CESPE/FINEP/2024) Duas variáveis, X e Y, possuem a mesma variância; se a correlação linear de Pearson entre elas for 0,8, e se a covariância entre X e Y for 2, então a variância de X será

a) 2,50.

b) 1,25.

c) 0,20.

d) 2,00.

e) 0,40.

2. (CESPE/FUB/2022) Julgue o item subsequente, considerando oito pares de valores das variáveis X e Y, tais que $\Sigma X = 24$; $\Sigma Y = 49$; $\Sigma XY = 181$; $\Sigma X^2 = 100$ e $\Sigma Y^2 = 343$.

O coeficiente de correlação de Pearson para os valores apresentados será negativo, o que indica que a regressão linear será representada por uma reta decrescente.

3. (CESPE/FUB/2022) Julgue o item subsequente, considerando oito pares de valores das variáveis X e Y, tais que $\Sigma X = 24$; $\Sigma Y = 49$; $\Sigma XY = 181$; $\Sigma X^2 = 100$ e $\Sigma Y^2 = 343$.

Existe uma correlação forte entre as variáveis X e Y.

4. (CESPE/FUB/2022) Julgue o item subsequente, considerando oito pares de valores das variáveis X e Y, tais que $\Sigma X = 24$; $\Sigma Y = 49$; $\Sigma XY = 181$; $\Sigma X^2 = 100$ e $\Sigma Y^2 = 343$.

Com base no coeficiente de correlação linear, é correto afirmar, em face dos dados apresentados, que se trata de uma correlação espúria.

5. (CESPE/FUB/2022) Uma regressão linear de Y sobre X consiste em obter a equação de uma reta, ou uma função linear, como o modelo que irá melhor representar a relação entre as variáveis; a determinação dos parâmetros dessa reta é denominada ajustamento.

Considerando essas informações, julgue o seguinte item.

Para quaisquer valores das variáveis X e Y, a existência de um coeficiente de correlação diferente de zero é garantia para que haja uma relação entre X e Y.

6. (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.

7. (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:

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

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

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.

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

9. (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:

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

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

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) $1,0 < X < 1,2$.

10. (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:

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

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

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

GABARITO – CEBRASPE

Correlação Linear

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

LISTA DE QUESTÕES – CEBRASPE

Regressão Linear Simples

1. (CESPE/ANAC/2024) Mediante a aplicação do critério de mínimos quadrados ordinários, um analista deseja ajustar um modelo de regressão linear simples na forma y = a + bx + ε, com variância V, em que y representa a variável dependente, x é a variável regressora e ε denota um erro aleatório que segue distribuição normal com média zero. A partir de uma amostra aleatória simples de tamanho n = 46, o analista obteve as estatísticas descritivas mostradas na tabela a seguir.

Variável	Média Amostral	Desvio Padrão Amostral
Y	10	5
x	20	4

A partir dessas informações, e sabendo que a correlação linear de Pearson entre as variáveis y e x é igual a 0,5, julgue os próximos itens.

A estimativa do coeficiente b é igual ou superior a 0,6.

2. (CESPE/FINEP/2024)

x	0	1	2	3	4	5
y	0	1	3	13	14	25

Considerando que a tabela precedente exibe uma amostra aleatória bivariada (x,y) de tamanho 6, na qual representa uma variável dependente e denota uma variável regressora, assinale a opção que apresenta uma curva de regressão $(\hat{y})$ ajustada para esse conjunto de dados mediante aplicação do método de mínimos quadrados ordinários.

a) $\hat{y} = x^2$

b) $\hat{y} = 11{,}2$

c) $\hat{y} = 3{,}73x$

d) $\hat{y} = 0{,}5x^2 + 12{,}5$

e) $\hat{y} = 5x - 2$

3. (CESPE/FUB/2022) Julgue o item subsequente, considerando oito pares de valores das variáveis X e Y, tais que $\Sigma X = 24$; $\Sigma Y = 49$; $\Sigma XY = 181$; $\Sigma X^2 = 100$ e $\Sigma Y^2 = 343$.

A reta dos mínimos quadrados ordinários que representa a regressão linear simples de Y em X com intercepto não nulo terá coeficiente linear aproximado de 2,48.

4. (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.

5. (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.

6. (CESPE/PETROBRAS/2022)

Equação 1: $y_i = a + bX_1 + e$

Equação 2: $y_i = a + b_1X_1 + b_2X_2 + b_3X_3 + e$

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 .

7. (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) $\hat{y} = 18x - 70$

8. (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 $\bar{x} = \sum_{i=1}^{6} x_i/6$.

9. (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.

10. (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.

11. (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.

12. (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.

13. (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.

14. (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.

15. (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.

16. (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.

17. (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.

18. (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 $Y = 0{,}75X + 6{,}25$.

19. (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 $\sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2 = a^2 \sum_{i=1}^{n}(x_i - \overline{x})^2$.

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.

20. (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.

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.

O intercepto do referido modelo é igual ou superior a 0,8.

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 erro aleatório ϵ segue a distribuição normal padrão.

23. (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.

24. (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.

25. (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.

26. (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.

27. (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.

28. (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.

29. (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.

30. (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 $var[\hat{b}] = 10$.

31. (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 $\bar{x} < \bar{y}$.

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.

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.

33. (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.

34. (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.

GABARITO – CEBRASPE

Regressão Linear Simples

1. CERTO
2. LETRA A
3. CERTO
4. LETRA A
5. CERTO
6. ERRADO
7. LETRA A
8. CERTO
9. ERRADO
10. ERRADO
11. LETRA C
12. ERRADO
13. LETRA A
14. ERRADO
15. CERTO
16. CERTO
17. CERTO
18. CERTO
19. CERTO
20. ERRADO
21. ERRADO
22. CERTO
23. CERTO
24. CERTO
25. ERRADO
26. CERTO
27. ERRADO
28. CERTO
29. CERTO
30. CERTO
31. CERTO
32. ERRADO
33. CERTO
34. ERRADO

LISTA DE QUESTÕES – CEBRASPE

Análise de Variância da Regressão

1. (CESPE/ANAC/2024) Mediante a aplicação do critério de mínimos quadrados ordinários, um analista deseja ajustar um modelo de regressão linear simples na forma y = a + bx + ε, com variância V, em que y representa a variável dependente, x é a variável regressora e ε denota um erro aleatório que segue distribuição normal com média zero. A partir de uma amostra aleatória simples de tamanho n = 46, o analista obteve as estatísticas descritivas mostradas na tabela a seguir.

Variável	Média Amostral	Desvio Padrão Amostral
Y	10	5
x	20	4

A partir dessas informações, e sabendo que a correlação linear de Pearson entre as variáveis y e x é igual a 0,5, julgue os próximos itens.

Estima-se que a variância V seja inferior a 15.

2. (CESPE/ANAC/2024) Mediante a aplicação do critério de mínimos quadrados ordinários, um analista deseja ajustar um modelo de regressão linear simples na forma y = a + bx + ε, com variância V, em que y representa a variável dependente, x é a variável regressora e ε denota um erro aleatório que segue distribuição normal com média zero. A partir de uma amostra aleatória simples de tamanho n = 46, o analista obteve as estatísticas descritivas mostradas na tabela a seguir.

Variável	Média Amostral	Desvio Padrão Amostral
Y	10	5
x	20	4

A partir dessas informações, e sabendo que a correlação linear de Pearson entre as variáveis y e x é igual a 0,5, julgue os próximos itens.

50% da variação total de y é explicada por meio do modelo de regressão linear simples em questão.

3. (CESPE/FUB/2022) Uma regressão linear de Y sobre X consiste em obter a equação de uma reta, ou uma função linear, como o modelo que irá melhor representar a relação entre as variáveis; a determinação dos parâmetros dessa reta é denominada ajustamento.

Considerando essas informações, julgue o seguinte item.

Suponha-se que, em uma pesquisa, o coeficiente de correlação entre duas variáveis X e Y tenha gerado um valor para o coeficiente de correlação de Pearson de 0,9200. Nesse caso, considerando-se X a variável independente e Y a variável dependente, o percentual da variância de Y explicado por X será de 84,64%.

4. (CESPE/FUB/2022) Uma regressão linear de Y sobre X consiste em obter a equação de uma reta, ou uma função linear, como o modelo que irá melhor representar a relação entre as variáveis; a determinação dos parâmetros dessa reta é denominada ajustamento.

Considerando essas informações, julgue o seguinte item.

Um coeficiente de determinação entre as variáveis X e Y de 95% implica necessariamente a obtenção de uma reta dos mínimos quadrados crescente, ou seja, em uma correlação positiva.

5. (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.

6. (CESPE/POLITEC RO/2022) Em relação aos procedimentos técnicos relacionados aos procedimentos de amostragem, julgue os itens a seguir.

I. Quando se adiciona variáveis explicativas no modelo de regressão linear, espera-se o incremento da estatística R^2.

II. Ao se comparar modelos com diferentes quantidades de variáveis explicativas, deve-se analisar o valor de R^2 ajustado.

III. O aumento de variáveis explicativas aumenta o R^2 ajustado.

IV. Ao se estimar um modelo com quatro variáveis explicativas e compará-lo com um modelo com três variáveis explicativas, escolhe-se o modelo que retornar o maior valor de R^2 ajustado, tudo o mais constante.

Estão corretos apenas os itens

a) I e II.

b) I e III.

c) I e IV.

d) II e III.

e) III e IV.

7. (CESPE/SECONT ES/2022) Com base no modelo clássico de regressão linear, julgue o item a seguir.

Em se tratando do modelo de regressão múltipla, ao se compararem modelos com diferentes quantidades de variáveis explicativas, o correto é analisar o valor de R^2 *ajustado.*

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.

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

9. (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 $\hat{y}_i = 0,9 + 0,2x_i$.

10. (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.

11. (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.

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

12. (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.

A soma de quadrados dos resíduos é igual ou inferior a 76.

13. (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.

A estimativa de mínimos quadrados ordinários para o intercepto do modelo é igual a zero.

14. (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.

Se $\hat{a}$ denota a estimativa de mínimos quadrados ordinários do coeficiente angular a, então $\hat{a} = 7{,}2$.

15. (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.

16. (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.

17. (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.

18. (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.

$\sigma^2 = 10$.

19. (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.

20. (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.

21. (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.

22. (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.

23. (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.

24. (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.

25. (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.

26. (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.

27. (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.

28. (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.

29. (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.

30. (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.

31. (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 $H_A: \beta_0 \neq 0$.

32. (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.

33. (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.

34. (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.

35. (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 $j = 1{,}2, \ldots, 26$

36. (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 $\sum(x_i - \overline{x})^2 = 17,5$ e $E(y^2) = 7,25$

O coeficiente angular é maior que 1.

37. (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 $R^2 = 0,64$.

38. (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.

GABARITO – CEBRASPE

Análise de Variância da Regressão

1. ERRADO
2. ERRADO
3. CERTO
4. ERRADO
5. LETRA A
6. LETRA A
7. CERTO
8. ERRADO
9. ERRADO
10. CERTO
11. CERTO
12. ERRADO
13. ERRADO
14. CERTO
15. CERTO
16. CERTO
17. ERRADO
18. CERTO
19. ERRADO
20. ERRADO
21. CERTO
22. ERRADO
23. LETRA C
24. CERTO
25. CERTO
26. LETRA E
27. CERTO
28. CERTO
29. ERRADO
30. ERRADO
31. ERRADO
32. ERRADO
33. CERTO
34. CERTO
35. ERRADO
36. ERRADO
37. CERTO
38. ERRADO

LISTA DE QUESTÕES – CEBRASPE

Análise de Resíduos

1. (CESPE/CAPES/2024) Um modelo de regressão linear entre uma variável aleatória Y (dependente) e uma variável não aleatória X (independente) é definido por $Y = \beta_0 + \beta_1 X + \epsilon$ em que ϵ, denominado erro aleatório, é uma variável aleatória independente de X com média $E(\epsilon) = 0$ e desvio padrão $Var(\epsilon) = \sigma^2$. Um modelo de regressão linear é essencialmente um modelo para a probabilidade condicional de Y com relação a X, denotada por $P(Y|X)$ ele é chamado de simples se n for uma variável aleatória gaussiana. Fixando-se valores $X_1, X_2, \cdots, X_n$ para a variável independente X, pode-se definir n variáveis aleatórias $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$, com $i = 1, \cdots, n$. Pelo método dos mínimos quadrados, é possível obter estimadores $\widehat{\beta}_0$ e $\widehat{\beta}_1$ para os parâmetros β_0 e β_1 e definir o $Y_i = \beta_0 + \beta_1 X_i$ como o estimador para Y_i.Nesse contexto, são definidos os erros, denominados resíduos, como $Y_i - \widehat{Y}_i = e_i$, a soma dos quadrados dos resíduos $SQE = \sum_i {e_i}^2$ a soma dos quadrados totais $SQT = \sum_i (Y_i - \overline{Y})^2$, e a soma dos quadrados de regressão $SQR = \sum_i (\widehat{Y}_i - \overline{Y})^2$, com $\overline{Y} = \sum_i Y_i / n$.

Com base nessas informações, julgue os próximos itens, considerando uma variável T, com média nula e desvio padrão unitário, definida por uma distribuição t de Student com 30 graus de liberdade, que tenha o seguinte intervalo com probabilidade de $0,95$: $P(-2,042 < T < 2,042) = 0,95$.

Para um modelo de regressão que não seja simples, a soma dos resíduos não será nula.

2. (CESPE/TCE-PA/2016)

Fonte de Variação	Grau 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 hipótese de normalidade exigida pelo modelo pode ser verificada a partir do gráfico dos resíduos, apesar de ser importante fazer um teste estatístico para tal fim.

3. (CESPE/FUB/2013) A respeito dos métodos de análise de resíduos do modelo de regressão, julgue o item subsequente.

A suposição de homocedasticidade pode ser verificada através de um gráfico de resíduos.

4. (CESPE/FUB/2013) A respeito dos métodos de análise de resíduos do modelo de regressão, julgue o item subsequente.

Na análise de resíduos de um modelo de regressão, o diagrama de dispersão entre os resíduos do modelo ajustado e os valores preditos para a variável resposta permitem avaliar a ocorrência de heterocedasticidade.

5. (CESPE/MCom/2013)

FV	gl	SQ	QM	F
regressão			810	
resíduo	98			
total			80	

O quadro acima mostra parte de uma tabela de análise de variância (ANOVA), que resultou da regressão linear simples do tempo que um usuário permanece conectado à Internet 3G — Y, em minutos — sobre a renda — X — declarada por esse usuário. Os dados utilizados nesse ajuste pelo método de mínimos quadrados ordinários foram selecionados por amostragem aleatória simples de um cadastro de usuários. Com base nessas informações e no quadro apresentado, julgue o item seguinte acerca dos conceitos de análise de regressão, correlação e amostragem.

O gráfico dos resíduos permite diagnosticar a hipótese de sua normalidade, ou seja, com base nesse gráfico, é possível efetuar uma análise confirmatória.

6. (CESPE/ALECE/2011)

Fonte de variação	Graus de liberdade	Soma dos quadrados	Média dos quadrados	Razão F
regressão	1	2.061,49	2.061,49	433,40
erro	78	371,01	4,75	
total	79	2.432,50		

Um analista deseja avaliar se o tempo — Y —, em dias, que um processo judicial leva para ser concluído está relacionado com a quantidade — X — de juízes disponíveis no tribunal em que tal processo foi julgado. O quadro acima apresenta a tabela de análise de variância (ANOVA) correspondente a essa avaliação por regressão linear simples, em que Y é a variável resposta e X é a variável regressora, com base no método de mínimos quadrados ordinários. Considerando essas informações e os conceitos de análise de regressão linear e inferência estatística, julgue o item.

Uma ferramenta descritiva para avaliação e diagnóstico do modelo é o gráfico de resíduos. Nesse gráfico, os resíduos devem apresentar-se dispostos aleatoriamente em torno do ponto zero.

7. (CESPE/PF/2004) Entre janeiro e novembro de 2003, foi realizado um estudo para avaliar o número mensal de ocorrências, por 1.000 habitantes, registradas em delegacias de determinada região. Para esse estudo, foi considerado o modelo de regressão linear simples na forma $Y=a+\beta X+\epsilon$, em que X é uma variável que representa os meses e assume valores discretos 0, 1, 2, ..., 10, e Y representa o número de ocorrências por 1.000 habitantes registradas no respectivo mês X. Parte do objetivo desse estudo é estimar os coeficientes a e β. O erro aleatório é representado por ϵ.

As tabelas abaixo apresentam parte dos resultados do ajuste e da análise de variância.

coeficiente	estimativa de mínimos quadrados ordinários	erro-padrão
α	50	0,05
β	0,05	0,005

fonte de variação	graus de liberdade	soma dos quadrados	quadrado médio
modelo	1	0,3	*D*
erro	9	*B*	*E*
total	*A*	*C*	*F*

Com base no texto acima, julgue o item a seguir.

Considere que, na análise dos resíduos, o estudo verificou que Y segue uma distribuição normal. Nessa situação, conclui-se que os dados são heterocedásticos.

GABARITO – CEBRASPE

Análise de Resíduos

1. ERRADO
2. CERTO
3. CERTO
4. CERTO
5. ERRADO
6. CERTO
7. ERRADO

ESSA LEI TODO MUNDO CONHECE:
PIRATARIA É CRIME.

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

1. 

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

2. 

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

3. 

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

4. 

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

5. 

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

6. 

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

7. 

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

8. 

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

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