Regressão linear múltipla com duas variáveis

Um exercício resolvido

Um estudo foi realizado em um caminhão de cargas leves movido a diesel para verificar se a umidade (em %) e a temperatura (em ºC) do ar influenciam a emissão de óxido nitroso (em ppm). As medições das emissões foram tomadas em momentos diferentes, com condições experimentais variadas. Os dados observados estão na tabela abaixo.

Estação (j)	Vazão mínima média (y)	Declividade de drenagem (x_1)	Densidade de drenagem (x_2)
1	2,60	2,69	0,098
2	1,49	3,94	0,079
3	1,43	7,20	0,119
4	3,44	3,18	0,102
5	1,37	2,44	0,123
6	2,53	1,25	0,136
7	15,12	1,81	0,121
8	16,21	1,59	0,137
9	21,16	1,21	0,134
10	30,26	1,08	0,018
11	28,53	1,00	0,141
12	1,33	4,52	0,064
13	0,43	10,27	0,131
14	39,12	0,66	0,143
15	45,00	0,60	0,133

a) Assumindo que o modelo de regressão linear múltipla é adequado para descrever a relação entre y, x_1 e x_2, estime os parâmetros do modelo e ajuste a equação do plano.
b) Efetue a **análise da variância** para testar a hipótese geral de linearidade da relação entre y, x_1 e x_2.
c) Calcule o coeficiente de determinação corrigido.
d) Teste as hipóteses parciais sobre o coeficiente de regressão, usando α = 0,05. (Relacione todos os passos do teste de hipóteses.)

Tabela auxiliar

j	y	x_1	x_2	y^2	x_1^2	x_2^2	x_1y	x_2y	x_1x_2
1	2,6	2,69	0,098	6,760	7,236	0,00960	6,994	0,2548	0,2636
2	1,49	3,94	0,079	2,220	15,524	0,00624	5,871	0,11771	0,3113
3	1,43	7,20	0,119	2,045	51,840	0,01416	10,296	0,17017	0,8568
4	3,44	3,18	0,102	11,834	10,112	0,01040	10,939	0,35088	0,3244
5	1,37	2,44	0,123	1,877	5,954	0,01513	3,343	0,16851	0,3001
6	2,53	1,25	0,136	6,401	1,563	0,01850	3,163	0,34408	0,1700
7	15,12	1,81	0,121	228,614	3,276	0,01464	27,367	1,82952	0,2190
8	16,21	1,59	0,137	262,764	2,528	0,01877	25,774	2,22077	0,2178
9	21,16	1,21	0,134	447,746	1,464	0,01796	25,604	2,83544	0,1621
10	30,26	1,08	0,018	915,668	1,166	0,00032	32,681	0,54468	0,0194
11	28,53	1,00	0,141	813,961	1,000	0,01988	28,530	4,02273	0,1410
12	1,33	4,52	0,064	1,769	20,430	0,00410	6,012	0,08512	0,2893
13	0,43	10,27	0,131	0,185	105,473	0,01716	4,416	0,05633	1,3454
14	39,12	0,66	0,143	1530,374	0,436	0,02045	25,819	5,59416	0,0944
15	45,00	0,60	0,133	2025,000	0,360	0,01769	27,000	5,985	0,0798
Soma	210,02	43,44	1,679	6257,217	228,362	0,20500	243,808	24,580	4,7944
Média	14	2,896	0,1119						

Cálculos iniciais

$$SQY = \sum y_j^2 - n\bar{y}^2 = 6257{,}217 - 15 \times 14^2 = \mathbf{3317{,}22}$$

$$SQX_1 = \sum x_{1j}^2 - n\bar{x}_1^2 = 228{,}362 - 15 \times 2{,}896^2 = \mathbf{102{,}56}$$

$$SQX_2 = \sum x_{2j}^2 - n\bar{x}_2^2 = 0{,}2050 - 15 \times 0{,}1119^2 = \mathbf{0{,}01716}$$

$$SPX_1Y = \sum x_{1j}y_j - n\bar{x}_1\bar{y} = 243{,}808 - 15 \times 14 \times 2{,}896 = -\mathbf{364{,}35}$$

$$SPX_2Y = \sum x_{2j}y_j - n\bar{x}_2\bar{y} = 24{,}580 - 15 \times 14 \times 0{,}1119 = \mathbf{1{,}081}$$

$$SPX_1X_2 = \sum x_{1j}x_{2j} - n\bar{x}_1\bar{x}_2 = 4{,}7944 - 15 \times 2{,}896 \times 0{,}1119 = -\mathbf{0{,}06654}$$

Cálculos iniciais

$$SQY = \sum y_j^2 - n\bar{y}^2 = 6257{,}217 - 15 \times 14^2 = \mathbf{3317{,}22}$$

$$SQX_1 = \sum x_{1j}^2 - n\bar{x}_1^2 = 228{,}362 - 15 \times 2{,}896^2 = \mathbf{102{,}56}$$

$$SQX_2 = \sum x_{2j}^2 - n\bar{x}_2^2 = 0{,}2050 - 15 \times 0{,}1119^2 = \mathbf{0{,}01716}$$

$$SPX_1Y = \sum x_{1j}y_j - n\bar{x}_1\bar{y} = 243{,}808 - 15 \times 14 \times 2{,}896 = -\mathbf{364{,}35}$$

$$SPX_2Y = \sum x_{2j}y_j - n\bar{x}_2\bar{y} = 24{,}580 - 15 \times 14 \times 0{,}1119 = \mathbf{1{,}081}$$

$$SPX_1X_2 = \sum x_{1j}x_{2j} - n\bar{x}_1\bar{x}_2 = 4{,}7944 - 15 \times 2{,}896 \times 0{,}1119 = -\mathbf{0{,}06654}$$

Estimação dos parâmetros do modelo

$$\begin{cases} \hat{\beta}_1 SQX_1 + \hat{\beta}_2 SPX_1X_2 = SPX_1Y \\ \hat{\beta}_1 SPX_1X_2 + \hat{\beta}_2 SQX_2 = SPX_2Y \end{cases}$$

$$\begin{cases} 102{,}56\,\hat{\beta}_1 - 0{,}06654\,\hat{\beta}_2 = -364{,}35 \\ -0{,}06654\,\hat{\beta}_1 + 0{,}01716\,\hat{\beta}_2 = 1{,}081 \end{cases}$$

$$\hat{\beta}_1 = -3{,}5207$$

$$\hat{\beta}_2 = 49{,}3$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}_1 - \hat{\beta}_2\bar{x}_2$$

$$\hat{\beta}_0 = 14 + 3{,}52 \times 2{,}896 - 49{,}3 \times 0{,}1119$$

$$\hat{\beta}_0 = 18{,}677$$

Equação do plano ajustada: $\hat{\mu} = 18{,}68 - 3{,}52\,x_1 + 49{,}3\,x_2$

Teste da hipótese de linearidade da relação entre as variáveis

$$\begin{cases} H_0 : \beta_i = 0, \text{ sendo } i = 1, 2 \\ H_1 : \beta_i \neq 0, \text{ para pelo menos um } i \ (i = 1 \text{ e/ou } 2) \end{cases}$$

Tabela da análise da variância

Fonte de variação	GL (ν)	SQ	QM (S^2)	F
Regressão	2	1335,81	667,905	4,045
Resíduo	12	1981,41	165,12	-
Total	14	3317,22	-	-

$f_{\alpha(2;\ 12)} = 3,89$

Rejeitamos H_0

Obtenção das somas de quadrados:

$$SQ_{Total} = \sum (y_j - \bar{y})^2 = SQY = \mathbf{3317,22}$$

$$SQ_{Reg} = \sum (\hat{\mu}_i - \bar{y})^2 = \hat{\beta}_1 SPX_1Y + \hat{\beta}_2 SPX_2Y = -3,52 \times (-364,35) + 49,3 \times 1,081 = \mathbf{1335,81}$$

$$SQ_{Res} = \sum (y_j - \hat{\mu}_j)^2 = \sum \hat{e}_j^2$$ (obtido por diferença)

Tabela da análise da variância

Fonte de variação	GL (ν)	SQ	QM (S^2)	F
Regressão	2	1335,81	667,905	4,045
Resíduo	12	1981,41	165,12	-
Total	14	3317,22	-	-

Coeficiente de determinação (r^2)

$$r^2 = \frac{SQ_{Reg}}{SQ_{Total}} = \frac{1335,81}{3317,22} = 0,4027$$

Coeficiente de determinação corrigido

$$r_C^2 = r^2 - \frac{2}{n-3}\left(1-r^2\right) = 0,4027 - \frac{2}{15-3}\left(1-0,4027\right) = 0,3032$$

Testes das hipóteses parciais

$$\begin{cases} H_0^1 : \beta_1 = 0 \\ H_1^1 : \beta_1 \neq 0 \end{cases} \quad e \quad \begin{cases} H_0^2 : \beta_2 = 0 \\ H_1^2 : \beta_2 \neq 0 \end{cases}$$

$$T = \frac{\hat{\beta}_i}{S(\hat{\beta}_i)} = \frac{\hat{\beta}_i}{\sqrt{S^2(\hat{\beta}_i)}} \sim t(\nu = n-3)$$

$$S^2(\hat{\beta}_1) = \frac{SQX_2}{SQX_1.SQX_2 - (SPX_1X_2)^2} \left(\frac{\sum_{j=1}^{n} \hat{e}_j^2}{n-3} \right)$$

$$S^2(\hat{\beta}_2) = \frac{SQX_1}{SQX_1.SQX_2 - (SPX_1X_2)^2} \left(\frac{\sum_{j=1}^{n} \hat{e}_j^2}{n-3} \right)$$

Variância do resíduo (buscar na tabela da análise da variância)

Cálculos iniciais

$$SQY = \sum y_j^2 - n\bar{y}^2 = 6257{,}217 - 15 \times 14^2 = 3317{,}22$$

$$SQX_1 = \sum x_{1j}^2 - n\bar{x}_1^2 = 228{,}362 - 15 \times 2{,}896^2 = 102{,}56$$

$$SQX_2 = \sum x_{2j}^2 - n\bar{x}_2^2 = 0{,}2050 - 15 \times 0{,}1119^2 = 0{,}01716$$

$$SPX_1Y = \sum x_{1j}y_j - n\bar{x}_1\bar{y} = 243{,}808 - 15 \times 14 \times 2{,}896 = -364{,}35$$

$$SPX_2Y = \sum x_{2j}y_j - n\bar{x}_2\bar{y} = 24{,}580 - 15 \times 14 \times 0{,}1119 = 1{,}081$$

$$SPX_1X_2 = \sum x_{1j}x_{2j} - n\bar{x}_1\bar{x}_2 = 4{,}7944 - 15 \times 2{,}896 \times 0{,}1119 = -0{,}06654$$

Estimação das variâncias dos estimadores dos coeficientes de regressão parciais

$$S^2(\hat{\beta}_1) = \frac{SQX_2}{SQX_1.SQX_2 - (SPX_1X_2)^2} \left(\frac{\sum_{j=1}^{n} \hat{e}_j^2}{n-3} \right) = \frac{0{,}01716}{102{,}56 \times 0{,}01716 - (-0{,}06654)^2} 165{,}12 = 1{,}614$$

$$S^2(\hat{\beta}_2) = \frac{SQX_1}{SQX_1.SQX_2 - (SPX_1X_2)^2} \left(\frac{\sum_{j=1}^{n} \hat{e}_j^2}{n-3} \right) = \frac{102{,}56}{102{,}56 \times 0{,}01716 - (-0{,}06654)^2} 165{,}12 = 9646{,}65$$

Testes das hipóteses parciais

Hipótese estatística

$$\begin{cases} H_0^1 : \beta_1 = 0 \\ H_1^1 : \beta_1 \neq 0 \end{cases}$$

Estatística do teste

$$t = \frac{\hat{\beta}_1}{s(\hat{\beta}_1)} = \frac{\hat{\beta}_1}{\sqrt{s^2(\hat{\beta}_1)}} = \frac{-3{,}52}{\sqrt{1{,}614}} = -2{,}771$$

Decisão e conclusão

$|t = -2{,}771| > t_{\alpha/2(12)} = 2{,}179$

Rejeitamos H_0

Hipótese estatística

$$\begin{cases} H_0^2 : \beta_2 = 0 \\ H_1^2 : \beta_2 \neq 0 \end{cases}$$

Estatística do teste

$$t = \frac{\hat{\beta}_2}{s(\hat{\beta}_2)} = \frac{\hat{\beta}_2}{\sqrt{s^2(\hat{\beta}_2)}} = \frac{49{,}3}{\sqrt{9646{,}65}} = 0{,}5019$$

Decisão e conclusão

$|t = 0{,}5019| < t_{\alpha/2(12)} = 2{,}179$

Não rejeitamos H_0

Concluímos, ao nível de 5% de significância, que a variável X_1 tem efeito significativo sobre Y, mas a variável X_2 não. Com base neste resultado, deve-se proceder a análise de regressão novamente, excluindo a variável X_2 do modelo.

Parametro	Estimativa	ErrPadrao	T	p	Inf95	Sup95
Intercep	24.29128	4.80486	5.0556	0.00022028	13.91101	34.67155
X1	-3.553159	1.231444	-2.8854	0.012761	-6.213532	-0.8927863

Fontes	GL	SQ	QM	F	p
Regressão	1	1294.8082	1294.8082	8.325303	0.01276
Resíduo	13	2021.849	155.52685	-	-
Total	14	3316.6572	-	-	-

(1)

Estat: Estatísticas Auxiliares

Estat	DesvPadr	CoefDet	CDetAjust
Valor	12.471	0.3904	0.3435

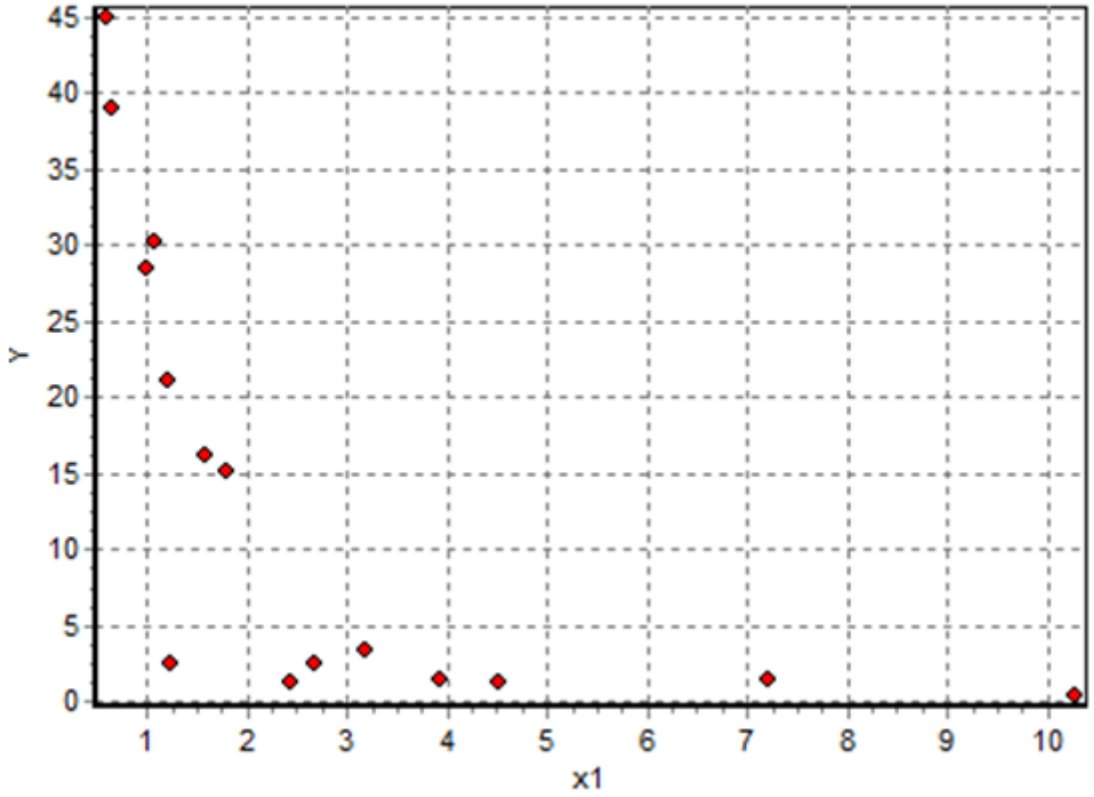
(2)

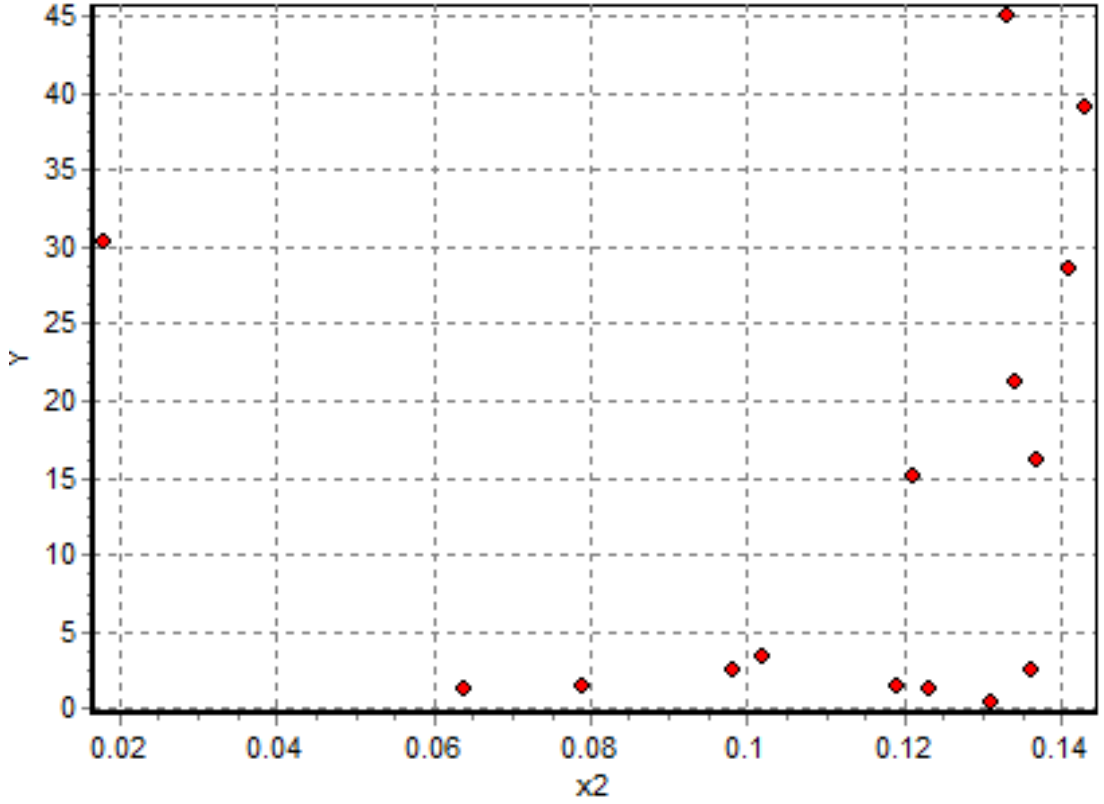
Dois critérios devem ser considerados na escolha do modelo:

(1) A significância do efeito linear de X_i sobre Y;

(2) O coeficiente de determinação (r^2).

Mesmo que o efeito linear de X_i sobre Y seja significativo, se o r^2 não for elevado, o modelo não deve ser adotado. Neste caso, é possível que a relação entre as variáveis seja melhor explicada por um modelo não linear.





Variaveis: Correlações entre as variáveis do modelo

Variaveis	X1	X2	Y
X1	1	-0.046335	-0.62482
X2	-0.046335	1	0.13054
Y	-0.62482	0.13054	1

Os diagramas de dispersão e os coeficientes de correlação linear entre Y e X_1 e X_2 indicam que a relação entre as variáveis pode ser não linear. Assim, é recomendável testar modelos não lineares para expressar a relação entre essas variáveis.