

## Gabarito da Lista 3 – Tópicos 3 e 4

1.

a) Redução da DQO: variável numérica contínua, escala de razão

Redução de ST: variável numérica contínua, escala de razão

b) Tabela auxiliar.

| i     | x     | y     | $x^2$ | $y^2$ | xy    |
|-------|-------|-------|-------|-------|-------|
| 1     | 3     | 5     | 9     | 25    | 15    |
| 2     | 7     | 11    | 49    | 121   | 77    |
| 3     | 11    | 21    | 121   | 441   | 231   |
| 4     | 15    | 16    | 225   | 256   | 240   |
| 5     | 18    | 16    | 324   | 256   | 288   |
| 6     | 27    | 28    | 729   | 784   | 756   |
| 7     | 29    | 27    | 841   | 729   | 783   |
| 8     | 30    | 25    | 900   | 625   | 750   |
| 9     | 30    | 35    | 900   | 1225  | 1050  |
| 10    | 31    | 30    | 961   | 900   | 930   |
| 11    | 31    | 40    | 961   | 1600  | 1240  |
| 12    | 32    | 32    | 1024  | 1024  | 1024  |
| 13    | 33    | 34    | 1089  | 1156  | 1122  |
| 14    | 33    | 32    | 1089  | 1024  | 1056  |
| 15    | 34    | 34    | 1156  | 1156  | 1156  |
| 16    | 36    | 37    | 1296  | 1369  | 1332  |
| 17    | 36    | 38    | 1296  | 1444  | 1368  |
| 18    | 36    | 34    | 1296  | 1156  | 1224  |
| 19    | 37    | 36    | 1369  | 1296  | 1332  |
| 20    | 38    | 38    | 1444  | 1444  | 1444  |
| 21    | 39    | 37    | 1521  | 1369  | 1443  |
| 22    | 39    | 36    | 1521  | 1296  | 1404  |
| 23    | 39    | 45    | 1521  | 2025  | 1755  |
| 24    | 40    | 39    | 1600  | 1521  | 1560  |
| 25    | 41    | 41    | 1681  | 1681  | 1681  |
| 26    | 42    | 40    | 1764  | 1600  | 1680  |
| 27    | 42    | 44    | 1764  | 1936  | 1848  |
| 28    | 43    | 37    | 1849  | 1369  | 1591  |
| 29    | 44    | 44    | 1936  | 1936  | 1936  |
| 30    | 45    | 46    | 2025  | 2116  | 2070  |
| 31    | 46    | 46    | 2116  | 2116  | 2116  |
| 32    | 47    | 49    | 2209  | 2401  | 2303  |
| 33    | 50    | 51    | 2500  | 2601  | 2550  |
| Soma  | 1104  | 1124  | 41086 | 41998 | 41355 |
| Média | 33,45 | 34,06 |       |       |       |

c)  $r = 0,9555$

Interpretação: Correlação forte e positiva entre as variáveis redução da DQO e redução de ST, ou seja, existe uma forte tendência de valores altos de redução da DQO estarem associados a valores altos de redução de ST, e vice-versa.

d)  $y_i = \beta_0 + \beta_1 x_i + e_i$

onde:

$y_i$  : redução da DQO é a variável resposta;

$x_i$  : redução de ST é a variável preditora;

$\beta_0$  : é o intercepto, ou seja, o valor redução da DQO quando a redução de ST é zero;

$\beta_1$  : é o coeficiente de regressão, ou seja, a taxa de variação da redução da DQO para cada unidade de redução de ST que aumenta;

$e_i$ : é o erro (variação aleatória não controlável)

Se Y é uma variável aleatória, então, está sujeita a um erro de observação. Este erro ( $e_i$ ) deverá ser adicionado ao modelo, desde que se admitam como verdadeiras as seguintes pressuposições:

1. Os erros são aleatórios, têm média zero e variância constante, ou seja,  $E(e_i)=0$  e  $V(e_i)=\sigma^2$ .
2. Os erros têm distribuição normal e são independentes entre si.
3. O modelo é adequado para todas as observações, não podendo haver nenhum valor de X que produza um valor de Y discrepante dos demais.
4. A variável X é fixa (não aleatória).

e) É o método de estimação de parâmetros que obtém estimativas de tal forma que a soma de quadrados dos erros seja o menor valor possível. Os estimadores dos coeficientes do modelo de regressão linear simples são obtidos por este método.

f)  $\hat{\beta}_1 = 0,9036$

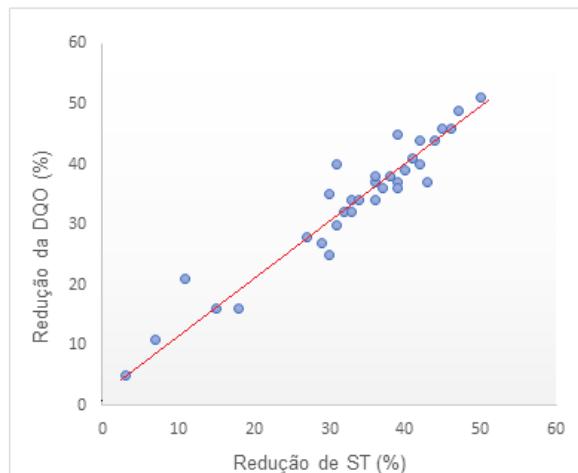
$\hat{\beta}_0 = 3,830$

g) Equação da reta ajustada:  $\hat{y}_i = 3,83 + 0,9036 x_i$

$\hat{y}_5 = 3,83 + 0,9036 \times 18 = 20,09$

$\hat{e}_5 = 16 - 20,09 = -4,09$

h)



- i) No modelo de regressão simples, o coeficiente de regressão é **proporcional** e tem o **mesmo sinal** do coeficiente de correlação.

$$\hat{\beta}_1 = r_{xy} \frac{\sqrt{SQX \cdot SQY}}{SQX}$$

- j) 1. Hipóteses estatísticas

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

2. Taxa de erro:  $\alpha = 0,05$

3. Cálculo das somas de quadrados

$$SQ_{\text{Total}} = SQY = 3713,88$$

$$SQ_{\text{Reg}} = \beta^2 \cdot SQX = 0,9036^2 \times 4152,18 = 3390,55$$

$$SQ_{\text{Resíduo}} = SQ_{\text{Total}} - SQ_{\text{Reg}} = 3713,88 - 3390,55 = 323,33$$

Tabela da análise da variação:

| Fontes    | GL | SQ      | QM      | F      |
|-----------|----|---------|---------|--------|
| Regressão | 1  | 3390,55 | 3390,55 | 325,08 |
| Resíduo   | 31 | 323,33  | 10,43   | -      |
| Total     | 32 | 3713,88 | -       | -      |

Como  $f_{\alpha(1; 31)} = 2,042 < f = 325,08$ , rejeitamos  $H_0$ .

Concluímos ao nível de 5% de significância que o coeficiente de regressão populacional difere de zero. Portanto, existe relação linear significativa entre a redução de ST e a redução da DQO.

- k) O coeficiente de determinação, denotado por  $r^2$ , expressa a proporção da variação total de Y que é explicada pela regressão, ou seja, pelo efeito linear de X sobre Y. Assim, o  $r^2$  informa sobre a qualidade do ajustamento do modelo aos dados observados. O coeficiente de determinação é o quadrado do coeficiente de correlação linear de Pearson.

- l)  $r^2 = 3390,55 / 3713,88 = 0,9129$

Verificamos que 91,29% da variação a redução da DQO é explicada pela variação da redução de ST. Portanto, o modelo de regressão apresenta um ótimo ajustamento aos dados observados.

- m)  $IC(\beta_0; 0,95): [0,22; 7,44]$

Concluímos, com 95% de confiança, que o intervalo de 0,22 a 7,44 contém o intercepto populacional, ou seja, quando a redução de ST é zero a redução da DQO populacional está entre 0,22 e 7,44.

- n)  $IC(\beta_1; 0,95): [0,80; 1,00]$

Concluímos, com 95% de confiança, que o coeficiente de regressão populacional é coberto pelos limites 0,80 e 1,00. Ou seja, a taxa de redução da DQO para cada unidade que aumenta na redução de ST está no intervalo de 0,8 a 1,00.

- o)  $IC(\mu_{y|x=30}; 0,95): 30,94 \pm 1,20$

$$IC(\mu_{y|x=30}; 0,95): [29,74; 32,14]$$

Concluímos, com 95% de confiança, que o intervalo de 29,74 a 32,14 contém a média de Y para X=30. Isto significa que temos 95% de confiança de que quando a redução de ST é igual a 30, a redução média de DQO está no intervalo de 29,74 a 32,14.

- p)  $IC(y_{x=30}; 0,95): 30,94 \pm 6,7$

$$IC(y_{x=30}; 0,95): [24,24; 37,64]$$

Concluímos, com 95% de confiança, que o intervalo de 24,24 a 37,64 contém uma observação de Y para X=30. Isto significa que temos 95% de confiança de que para uma redução de ST igual a 30 a redução de DQO para uma observação qualquer está entre 24,24 e 37,64.

- q) A amplitude do intervalo na predição é menor que na previsão, ou seja, a precisão é maior na predição.

- r) A extrapolação da equação de regressão para além dos limites dos dados amostrais utilizados na estimativa dos parâmetros do modelo de regressão linear não é recomendada por dois motivos: (1) a amplitude do intervalo de confiança sobre a linha de regressão aumenta à medida que os valores da variável X se afastam da média e (2) a relação entre as variáveis X e Y pode não ser linear para valores que extrapolam os dados utilizados na regressão.

## 2. Cálculos intermediários:

$$SQY = 2715,76$$

$$SQX_1 = 415,23$$

$$SQX_2 = 2905,69$$

$$SPYX_1 = 775,96$$

$$SPYX_2 = 2292,95$$

$$SPX_1X_2 = 251,077$$

- a)  $r_{yx1} = 0,7307$

$$r_{yx2} = 0,8163$$

- b)

$$y_j = \beta_0 + \beta_1 x_{1j} + \beta_2 x_{2j} + e_j,$$

onde:

$y_j$  é a quantidade de calor por grama de cimento;

$x_{1j}$  é a quantidade de aluminato tricálcico;

$x_{2j}$  é a quantidade de silicato tricálcico;

$\beta_0$  é quantidade de calor do cimento, quando as quantidades de aluminato tricálcico e silicato tricálcico são iguais a zero ( $X_1=0$  e  $X_2=0$ );

$\beta_1$  é a taxa de variação a quantidade de calor, para cada unidade que se acrescenta na quantidade de aluminato tricálcico, numa quantidade fixa qualquer de silicato tricálcico;  
 $\beta_2$  é a taxa de variação a quantidade de calor, para cada unidade que se acrescenta na quantidade de silicato tricálcico, numa quantidade fixa qualquer de aluminato tricálcico;  
 $e_j$  é o erro (variação aleatória) associado à observação j.

- c) Se Y é uma variável aleatória, então, está sujeita a um erro de observação. Este erro ( $e_i$ ) deverá ser adicionado ao modelo, desde que se admitam como verdadeiras as seguintes pressuposições:

1. As variáveis  $X_i$  são fixas, isto é, observados sem erro.
2. Os erros são aleatórios, têm média zero e variância constante, ou seja,  $E(e_i)=0$  e  $V(e_i)=\sigma^2$ .
3. Os erros têm distribuição normal e são independentes entre si.
4. O modelo é adequado para todas as observações, não podendo haver nenhum valor de X que produza um valor de Y discrepante dos demais.

d)

$$\hat{\beta}_0 = 52,58$$

$$\hat{\beta}_1 = 1,468$$

$$\hat{\beta}_2 = 0,6623$$

$$\hat{\mu} = 52,58 + 1,468 X_1 + 0,6623 X_2$$

e)

1. Hipótese estatística

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

2. Taxa de erro:  $\alpha = 0,05$

3. Cálculo das somas de quadrados

$$SQ_{Total} = SQY = 2715,76$$

$$SQ_{Reg} = \beta_1 \cdot SPYX_1 + \beta_2 \cdot SPYX_2 = 2657,86$$

$$SQ_{Resíduo} = SQTotal - SQReg = 2715,76 - 2657,86 = 57,904$$

Tabela da análise da variação:

| Fontes    | GL | SQ       | QM       | F      |
|-----------|----|----------|----------|--------|
| Régressão | 2  | 2657,86  | 1328,929 | 229,50 |
| Resíduo   | 10 | 57,904   | 5,7904   | -      |
| Total     | 12 | 2715,763 | -        | -      |

Como  $f_{\alpha(2; 10)} = 4,10 < f = 229,50$ , rejeitamos  $H_0$ .

Concluímos ao nível de 5% de significância que pelo menos um dos coeficientes de regressão parciais difere de zero. Portanto, existe relação linear significativa entre a quantidade de calor e pelo menos uma das variáveis quantidade de aluminato tricálcico e quantidade de silicato tricálcico.

f)

$$r^2_c = 0,9744$$

Interpretação: O modelo tem um ótimo aos dados observados, pois 97% da variação que pelo menos um dos coeficientes de regressão parciais difere de zero. Portanto, existe relação linear significativa entre a quantidade de calor e pelo menos uma das variáveis quantidade de aluminato tricálcico e quantidade de silicato tricálcico..

g)

1. Hipótese estatística

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

2. Taxa de erro:  $\alpha = 0,05$

3. Estatística do teste

$$t = 12,11$$

4. Decisão e conclusão

Como  $t_{\alpha/2(10)} = 2,228 < t = 12,11$ , rejeitamos  $H_0$ .

Concluímos ao nível de 5% de significância o coeficiente de regressão  $\beta_1$  difere de zero. Portanto, existe efeito significativo da quantidade de aluminato tricálcico sobre a quantidade de calor, adicional ao efeito da quantidade de silicato tricálcico.

1. Hipótese estatística

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

2. Taxa de erro:  $\alpha = 0,05$

3. Estatística do teste

$$t = 14,44$$

4. Decisão e conclusão

Como  $t_{\alpha/2(10)} = 2,228 < t = 14,44$ , rejeitamos  $H_0$ .

Concluímos ao nível de 5% de significância o coeficiente de regressão  $\beta_2$  difere de zero. Portanto, existe efeito significativo da quantidade de silicato tricálcico sobre a quantidade de calor, adicional ao efeito da quantidade de aluminato tricálcico.

- h) A relação linear entre Y e  $(X_1, X_2)$  foi significativa, as contribuições adicionais das variáveis  $X_1$  e  $X_2$  para a explicação da variação de Y foram significativas. Assim, segundo os testes efetuados, o "melhor" modelo para exprimir a relação linear entre Y e  $(X_1, X_2)$  é o modelo de regressão linear múltipla com duas variáveis.
- i) Valor esperado de y ( $x_1=15$  e  $x_2=30$ ) = 94,47.
- j) Descreva resumidamente os principais métodos de seleção de variáveis.
- ⇒ Inclusão ascendente (*forward selection*): inicia-se com um modelo que possui somente o intercepto e, de acordo com o critério fixado, as variáveis preditoras são incluída no modelo, uma a uma. Uma vez incluída no modelo, a variável não sai mais.
- ⇒ Seleção descendente (*backward elimination*): começa com o modelo completo e, de acordo com o critério fixado, vai excluindo, uma a uma, as variáveis de menor contribuição não significativa, na presença das demais variáveis no modelo.
- ⇒ Seleção ascendente-descendente (*stepwise selection*) é uma aplicação conjunta dos critérios de inclusão e exclusão. O procedimento inicia do mesmo modo que a seleção ascendente, mas em cada passo verifica se, na presença das outras variáveis do modelo, alguma variável não agrega contribuição significativa à explicação da variação da resposta. Dentre as que não estão contribuindo significativamente, a de menor f parcial é eliminada. Por outro lado, uma variável que já foi excluída poderá retornar em um passo posterior.