Cap. 8 – Regressão linear simples

8.1. Introdução

A análise de regressão consiste na realização de uma análise estatística com o objetivo de verificar a existência de uma relação funcional entre uma variável dependente (Y) com uma ou mais variáveis independentes (X). Em outras palavras, consiste na obtenção de uma equação que tenta explicar a variação da variável dependente pela variação dos níveis da(s) variável(is) independente(s). As variáveis independentes são classificadas como quantitativas, cujos níveis representam diferentes quantidades de um mesmo fator.

- Desempenho de coelhos e diferentes níveis (7%, 9%, 11%, 13%) de fibra.

- Rendimento da cultura e densidade de plantio.

- Peso dos animais e idade.

- Teor de gordura no leite e densidade.

- Teor de açúcar no iogurte (6, 8, 10, 12, 14%) e aceitação do produto (1, 2, ... , 9).

8.2 – Escolha do modelo

Para tentar estabelecer uma equação que representa o fenômeno em estudo, pode-se plotar um diagrama de dispersão de Y em função de X.

O comportamento de Y em relação a X, pode se apresentar de diversas maneiras: linear, quadrático, cúbico, exponencial etc.

Pontos não se ajustam perfeitamente a curva do modelo matemático. Pois os fenômenos estudados não são matemáticos e sim um fenômeno biológico, químico etc. Que estão sujeitos a influências que acontecem ao acaso.

- O modelo escolhido deve ser coerente com o que acontece na prática;

- O modelo selecionado deve ser condizente tanto no grau como no aspecto da curva, para representar em termos práticos, o fenômeno em estudo;

- Conter apenas as variáveis que são relevantes para explicar o fenômeno.

8.2 – Método para obtenção da equação estimada

Os pontos do diagrama de dispersão ficam um pouco distantes da curva do modelo matemático escolhido. Um dos métodos que se pode utilizar para obter a relação funcional, se baseia na obtenção de uma equação estimada de tal forma que as distâncias entre os pontos do diagrama e os pontos da curva do modelo matemático, no todo, sejam as menores possíveis. Este método é denominado de Método dos Mínimos Quadrados (MMQ). Em resumo, por este método a soma de quadrados das distâncias entre os pontos do diagrama e os respectivos pontos na curva da equação estimada é minimizada, obtendo-se, desta forma, uma relação funcional entre X e Y, para o modelo escolhido, com um mínimo de desvio possível.

8.3 – Modelo linear de 1° grau

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

Obtenção da equação estimada – MMQ:

$$\varepsilon_i = Y_i - \beta_0 - \beta_1 X_i$$

$$(\varepsilon_i)^2 = (Y_i - \beta_0 - \beta_1 X_i)^2$$

$$\sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2$$

$$\begin{cases} \dfrac{\delta \sum_{i=1}^{n} \varepsilon_i^2}{\delta \beta_0} = 0 \Rightarrow 2\sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)(-1) = 0 \Rightarrow \sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) = 0 \\ \dfrac{\delta \sum_{i=1}^{n} \varepsilon_i^2}{\delta \beta_1} = 0 \Rightarrow 2\sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)(-X_i) = 0 \Rightarrow \sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)(X_i) = 0 \end{cases}$$

$$\begin{cases} \sum_{i=1}^{n} Y_i - \sum_{i=1}^{n} \hat{\beta}_0 - \sum_{i=1}^{n} \hat{\beta}_1 X_i = 0 \Rightarrow \sum_{i=1}^{n} Y_i - n\hat{\beta}_0 - \hat{\beta}_1 \sum_{i=1}^{n} X_i = 0 \\ \sum_{i=1}^{n} Y_i X_i - \sum_{i=1}^{n} \hat{\beta}_0 X_i - \sum_{i=1}^{n} \hat{\beta}_1 X_i^2 = 0 \Rightarrow \sum_{i=1}^{n} Y_i X_i - \hat{\beta}_0 \sum_{i=1}^{n} X_i - \hat{\beta}_1 \sum_{i=1}^{n} X_i^2 = 0 \end{cases}$$

$$\begin{cases} \sum_{i=1}^{n} Y_i = n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} X_i \\ \sum_{i=1}^{n} Y_i X_i = \hat{\beta}_0 \sum_{i=1}^{n} X_i + \hat{\beta}_1 \sum_{i=1}^{n} X_i^2 \end{cases}$$

$$\begin{cases} \hat{\beta}_1 = \dfrac{\sum_{i=1}^{n} X_i Y_i - \dfrac{\sum_{i=1}^{n} X_i \cdot \sum_{i=1}^{n} Y_i}{n}}{\sum_{i=1}^{n} X_i^2 - \dfrac{\left(\sum_{i=1}^{n} X_i\right)^2}{n}} \\ \hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X} \end{cases}$$

Equação estimada $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$

8.5 – Análise de variância da regressão

$$H_0 : \beta_1 = \beta_2 = ... = \beta_p = 0 \ vs \ H_a : \beta_i \neq 0$$

FV	GL	SQ	QM	F
Regressão	p	SQReg	QMReg	
Desvio da Regres	n-p-1	SQDes	QMDes	
Total	n-1	SQTotal		

$$SQT_{otal} = \sum_{i=1}^{n} Y_i^2 - \frac{\left(\sum_{i=1}^{n} Y_i \right)^2}{n}$$

$$\mathrm{SQDes} = \mathrm{SQT_{otal}} - \mathrm{SQ\,Re\,g}$$

$$SQ\,\mathrm{Re}\,g = \hat{\beta}_0 \sum_{i=1}^{n} Y_i + \hat{\beta}_1 \sum_{i=1}^{n} Y_i X_i - \frac{\left(\sum_{i=1}^{n} Y_i \right)^2}{n}$$

Linear grau 1

$$SQ\,\mathrm{Re}\,g = \hat{\beta}_0 \sum_{i=1}^{n} Y_i + \hat{\beta}_1 \sum_{i=1}^{n} Y_i X_i + \hat{\beta}_2 \sum_{i=1}^{n} Y_i X_i^2 - \frac{\left(\sum_{i=1}^{n} Y_i \right)^2}{n}$$

Linear grau 2

8.6.1 – Coeficiente de determinação

$$R^2 = \frac{SQ\,\mathrm{Re}\,g}{SQTotal}$$

8.6.2 – Coeficiente de determinação ajustado para grau de liberdade

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

8.7 – Exemplo 1: As produções médias de leite de um grupo de vacas tratadas com diferentes níveis de proteínas na ração foram as seguintes:

X	10	12	14	16	18	20	22	24	26	28
Y	11,8	12,0	12,1	13,2	14,1	14,4	15,6	16,0	16,4	17,0

X – Proteína na ração em %; Y – produção de leite em kg

Modelo proposto → $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$

8.7 – Exemplo 1: As produções médias de leite de um grupo de vacas tratadas com diferentes níveis de proteínas na ração foram as seguintes:

X	10	12	14	16	18	20	22	24	26	28
Y	11,8	12,0	12,1	13,2	14,1	14,4	15,6	16,0	16,4	17,0

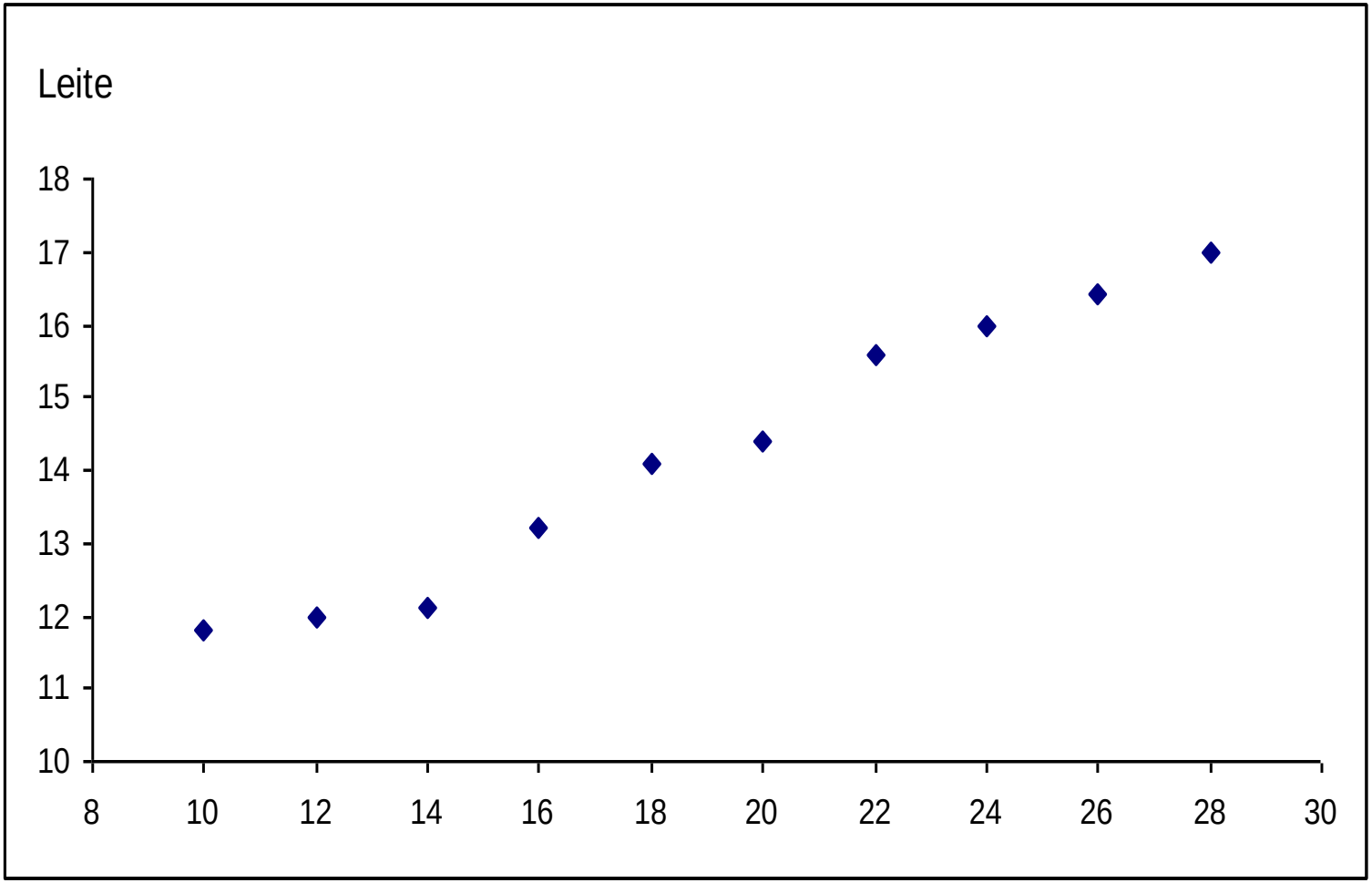
X – Proteína na ração em %; Y – produção de leite em kg

$$\sum_{i=1}^{10} X_i = 10+12+...+28 = 190{,}0 \qquad \sum_{i=1}^{10} X_i^2 = 10^2+12^2+...+28^2 = 3.940{,}0 \qquad \sum_{i=1}^{10} Y_i = 11{,}8+12{,}0+...+17{,}0 = 142{,}6$$

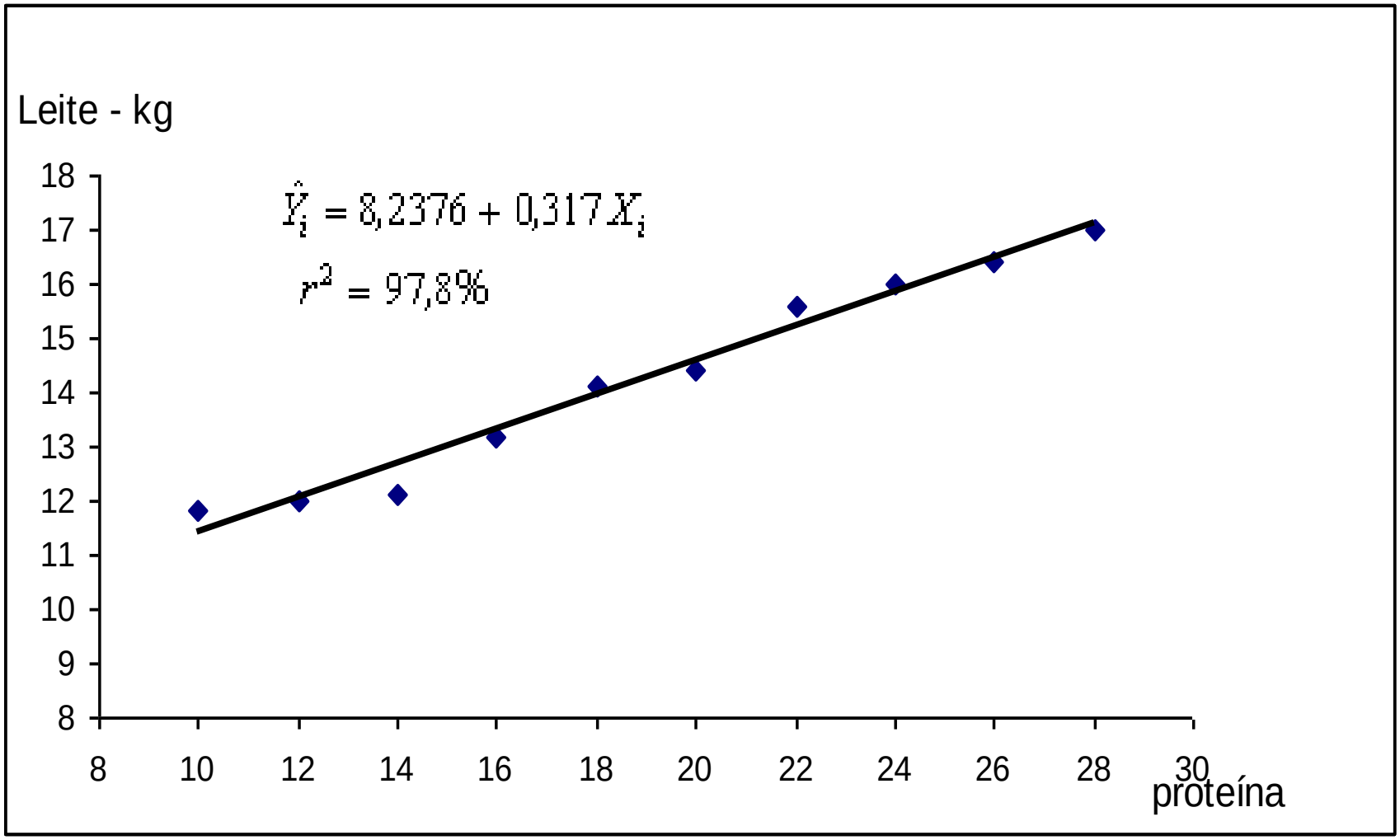
$$\sum_{i=1}^{10} Y_i^2 = 11{,}8^2+12{,}0^2+...+17{,}0^2 = 2.067{,}38 \qquad \sum_{i=1}^{10} X_i Y_i = 10x11{,}8+12x12{,}0+...+28x17{,}0 = 2.814{,}0$$

$$\begin{cases} \sum_{i=1}^{n} Y_i = n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} X_i \\ \sum_{i=1}^{n} Y_i X_i = \hat{\beta}_0 \sum_{i=1}^{n} X_i + \hat{\beta}_1 \sum_{i=1}^{n} X_i^2 \end{cases} \qquad \begin{cases} 142{,}6 = 10\hat{\beta}_0 + 190\hat{\beta}_1 \\ 2.814 = 190\hat{\beta}_0 + 3.940\hat{\beta}_1 \end{cases} \qquad \begin{cases} \hat{\beta}_0 = 8{,}237575 \\ \hat{\beta}_1 = 0{,}316970 \end{cases}$$

Equação estimada => $\hat{Y}_i = 8{,}237575 + 0{,}316970 X_i$


Leite
18
17
16
15
14
13
12
11
10
8
10
12
14
16
18
20
22
24
26
28
30

Representação gráfica



Análise de variância da regressão

$$H_0 : \beta_1 = 0 \ vs \ H_a : \beta_1 \neq 0$$

FV	GL	SQ	QM	F	Signif.
Regressão	1	33,1558	33,1558	354,5	0,000
Desvio da Regres	8	0,7482	0,0935		
Total	9	33,9040			

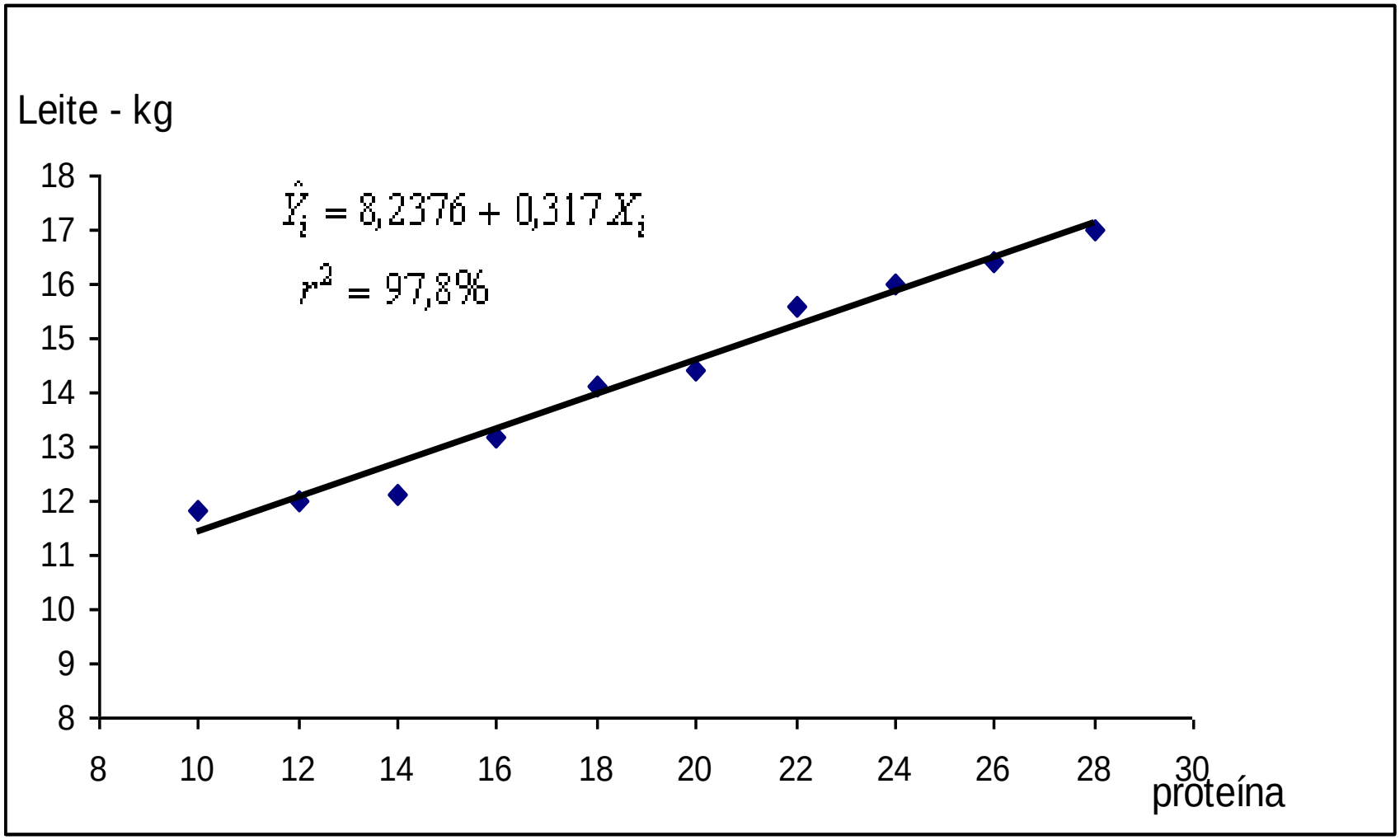
$$SQT_{otal} = \sum_{i=1}^{n} Y_i^2 - \frac{\left(\sum_{i=1}^{n} Y_i\right)^2}{n} = 2.067{,}38 - \frac{142{,}6^2}{10} = 33{,}9040$$

$$SQ\,\mathrm{Re}\,g = \hat{\beta}_0 \sum_{i=1}^{n} Y_i + \hat{\beta}_1 \sum_{i=1}^{n} Y_i X_i - \frac{\left(\sum_{i=1}^{n} Y_i\right)^2}{n} = 8{,}237575x142{,}6 + 0{,}31697x2.814{,}0 - \frac{142{,}6^2}{10} = 33{,}1558$$

$$\mathrm{SQDes} = \mathrm{SQT}_{\mathrm{otal}} - \mathrm{SQ\,Re\,g} = 33{,}9040 - 33{,}1558 = 0{,}7482$$

$$F_{calc} = \frac{QM\,\mathrm{Re}\,g}{QM\,\mathrm{Re}\,s} = \frac{33{,}1558}{0{,}0935} = 354{,}51 \qquad F_{tab} = F_{1\%(1,8)} = 11{,}26$$

Representação gráfica

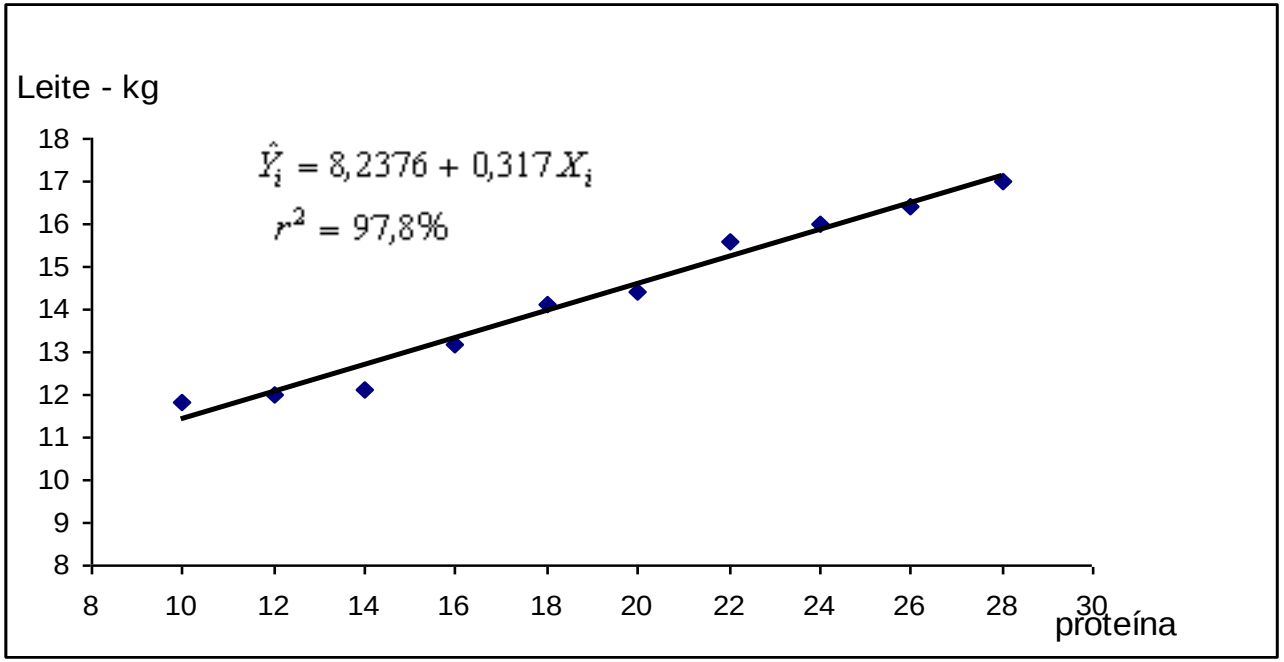


Coeficiente de determinação

$$R^2 = \frac{SQ\,\mathrm{Re}\,g}{SQTotal} = \frac{33{,}1558}{33{,}9040} = 0{,}9779 = 97{,}79\%$$

Coeficiente de determinação ajustado para grau de liberdade

$$\overline{R}^2 = R^2 - \frac{p}{n-p-1}\left(1-R^2\right) = 0{,}9779 - \frac{1}{10-1-1}\left(1-0{,}9779\right) = 0{,}9751 = 97{,}51\%$$



Exemplo 2: Considere um experimento em que se testou seis níveis de um antioxidante. Os valores de X são os níveis do antioxidante expressos em porcentagem e os valores de Y são os escores para rancidez.

X	0	2	4	6	8	10
Y	0,5	2,5	3,0	4,8	5,0	4,7

X –níveis de antioxidantes;

Y – escores para rancidez.

Modelo proposto → $Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \varepsilon_i$

Exemplo 2: Considere um experimento em que se testou seis níveis de um antioxidante. Os valores de X são os níveis do antioxidante expressos em porcentagem e os valores de Y são os escores para rancidez.

X	0	2	4	6	8	10
Y	0,5	2,5	3,0	4,8	5,0	4,7

X –níveis de antioxidantes;

Y – escores para rancidez.

Modelo proposto → $Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \varepsilon_i$

$\sum_{i=1}^{6} X_i = 30{,}0$ $\sum_{i=1}^{6} X_i^2 = 220{,}0$ $\sum_{i=1}^{6} X_i^3 = 1.800$ $\sum_{i=1}^{6} X_i^4 = 15.664$ $\sum_{i=1}^{6} Y_i = 20{,}5$

$\sum_{i=1}^{6} Y_i = 20{,}5$ $\sum_{i=1}^{6} Y_i^2 = 85{,}63$ $\sum_{i=1}^{6} X_i Y_i = 132{,}8$ $\sum_{i=1}^{6} X_i^2 Y_i = 1.020{,}8$

$$\begin{cases} \sum_{i=1}^{n} Y_i = n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} X_i + \hat{\beta}_2 \sum_{i=1}^{n} X_i^2 \\ \sum_{i=1}^{n} Y_i X_i = \hat{\beta}_0 \sum_{i=1}^{n} X_i + \hat{\beta}_1 \sum_{i=1}^{n} X_i^2 + \hat{\beta}_2 \sum_{i=1}^{n} X_i^3 \\ \sum_{i=1}^{n} Y_i X_i^2 = \hat{\beta}_0 \sum_{i=1}^{n} X_i^2 + \hat{\beta}_1 \sum_{i=1}^{n} X_i^3 + \hat{\beta}_2 \sum_{i=1}^{n} X_i^4 \end{cases}$$

$$\begin{cases} 20{,}5 = 6\hat{\beta}_0 + 30{,}0\hat{\beta}_1 + 220{,}0\hat{\beta}_2 \\ 132{,}8 = 30{,}0\hat{\beta}_0 + 220{,}0\hat{\beta}_1 + 1800\hat{\beta}_2 \\ 1.020{,}8 = 220{,}0\hat{\beta}_0 + 1.800\hat{\beta}_1 + 15.664\hat{\beta}_2 \end{cases}$$

Equação estimada => $\hat{Y}_i = 0{,}4964 + 0{,}9998\, X_i - 0{,}0567\, X_i^2$

Análise de variância da regressão

$$H_0 : \beta_1 = \beta_2 = 0 \ vs \ H_a : não \ H_0$$

FV	GL	SQ	QM	F	Probab.
Regressão	2	15,0286	7,5143	40,28	0,0068
Desvio da Regres	3	0,5597	0,1866		
Total	5	15,5883			

$$SQT_{otal} = \sum_{i=1}^{n} Y_i^2 - \frac{\left(\sum_{i=1}^{n} Y_i\right)^2}{n} = 85{,}630 - 70{,}042 = 15{,}5883$$

$$SQDes = SQT_{otal} - SQ\text{Re } g = 15{,}5883 - 15{,}0286 = 0{,}5597$$

$$SQ\text{Re } g = \hat{\beta}_0 \sum_{i=1}^{n} Y_i + \hat{\beta}_1 \sum_{i=1}^{n} Y_i X_i + \hat{\beta}_2 \sum_{i=1}^{n} Y_i X_i^2 - \frac{\left(\sum_{i=1}^{n} Y_i\right)^2}{n}$$

$$SQ\text{Re } g = 0{,}4964 * 20{,}5 + 0{,}9998 * 132{,}8 - 0{,}0567 * 1.020{,}8 - \frac{20{,}5^2}{6} = 15{,}0286$$

$$F_{calc} = \frac{QM\text{ Re } g}{QM\text{ Re } s} = \frac{7{,}5143}{0{,}1866} = 40{,}28$$

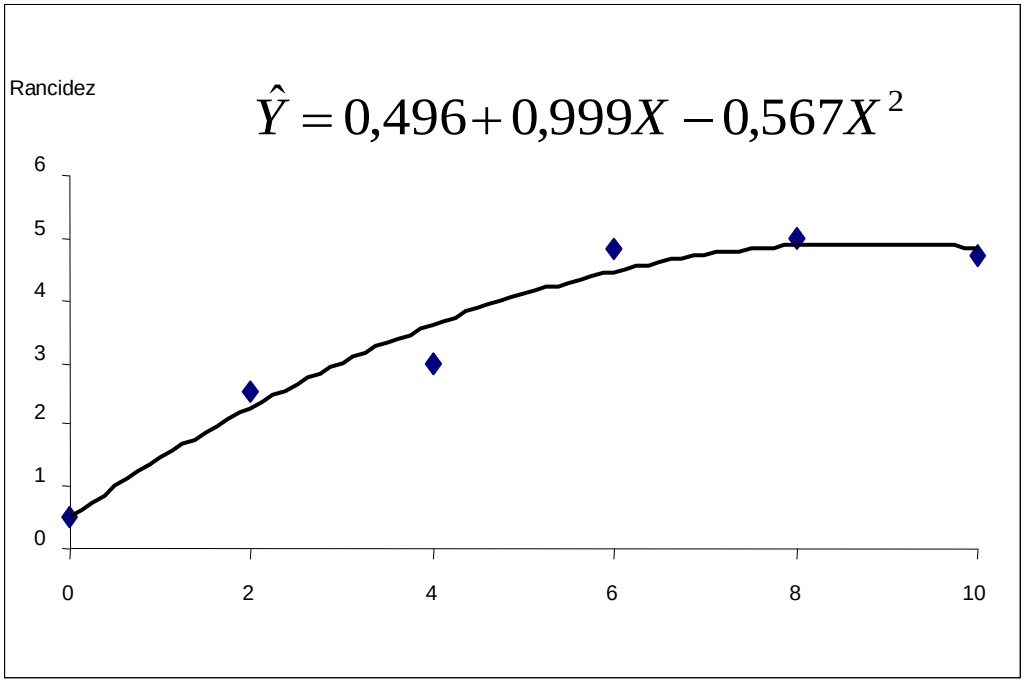
$$F_{tab} = F_{5\%(2,3)} = 9{,}55$$

Coeficiente de determinação

$$R^2 = \frac{SQ\text{Re } g}{SQTotal} = \frac{15{,}0286}{15{,}5883} = 0{,}9645 = 96{,}45\%$$

Coeficiente de determinação ajustado para grau de liberdade

$$\overline{R}^2 = R^2 - \frac{p}{n-p-1}\left(1-R^2\right) = 0{,}9645 - \frac{2}{6-2-1}\left(1-0{,}9645\right) = 0{,}9409 = 94{,}09\%$$



Ponto de máximo:

$$\frac{\delta\hat{Y}}{\delta X} = 0{,}9998 - 2x0{,}0567X$$

$$X_{\max} = \frac{0{,}9998}{2x0{,}0567} = 8{,}8$$

Resposta no PM:

$$\hat{Y}_i = 0{,}4964 + 0{,}9998x8{,}8 - 0{,}0567x8{,}8^2 = 4{,}90$$

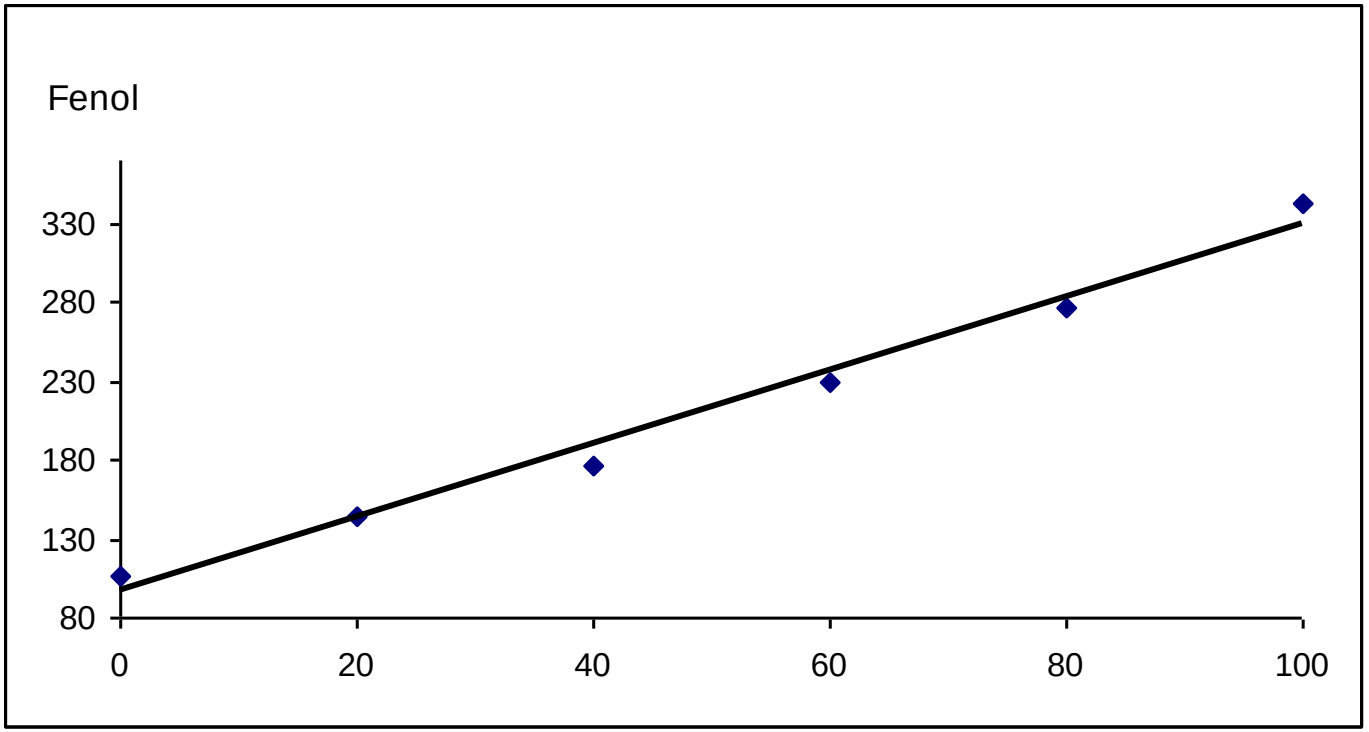
Exemplo 3: Uma pesquisadora realizou um ensaio para avaliar a extração de fenóis presente na especiaria açafrão. Para tanto utilizou diferentes concentrações de etanol (0, 20, 40, 60, 80 e 100%) diluído em água. Organizou o trabalho num DIC com quatro repetições.

Etanol (%)	Repetições				Totais Trat	Médias Trat
	R_1	R_2	R_3	R_4		
0	110,0	112,0	125,0	81,0	428,0	107,00
20	145,0	170,0	135,0	130,0	580,0	145,00
40	195,0	165,0	170,0	176,0	706,0	176,50
60	225,0	235,0	215,0	245,0	920,0	230,00
80	285,0	275,0	265,0	287,0	1112,0	278,00
100	360,0	385,0	292,0	335,0	1372,0	343,00

$$\begin{cases} \sum_{i=1}^{n} Y_i = n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} X_i \\ \sum_{i=1}^{n} Y_i X_i = \hat{\beta}_0 \sum_{i=1}^{n} X_i + \hat{\beta}_1 \sum_{i=1}^{n} X_i^2 \end{cases} \qquad \begin{cases} 5.118 = 24\hat{\beta}_0 + 1.200\hat{\beta}_1 \\ 321.200 = 1.200\hat{\beta}_0 + 88.000\hat{\beta}_1 \end{cases} \qquad \begin{cases} \hat{\beta}_0 = 96{,}643 \\ \hat{\beta}_1 = 2{,}332 \end{cases}$$

Equação estimada => $\hat{Y}_i = 96{,}643 + 2{,}332X_i$

Representação gráfica



Análise de variância da regressão

FV	GL	SQ	QM	F	P (F > f) %
Regressão	1	152.288,9	152.288,9	341,5	0,000
Falta de ajuste	4	2.134,60	533,65	1,20	0,346
(Tratamentos)	[5]	[154.423,0]	30.884,70	69,26	0,000
Resíduo	18	8.027,00	445,94		
Total	23	162.450,5			

$$SQT_{otal} = \sum_{i=1}^{n} Y_i^2 - C = 1.253.864 - C = 162.450,5$$

$$SQTrat = \sum_{i=1}^{6} \frac{T_i^2}{r_i} - C = 1.245.837 - C = 154.423,5$$

$$C = \frac{\left(\sum_{i=1}^{n} Y_i\right)^2}{n} = \frac{5.118^2}{24} = 1.091.413,5$$

$$SQ\text{Re } s = SQT_{otal} - SQTrat = 8.027,0$$

$$SQ\text{Re } g = \hat{\beta}_0 \sum_{i=1}^{n} Y_i + \hat{\beta}_1 \sum_{i=1}^{n} Y_i X_i - C = 96,643x5.118 + 2,332x321.200 - C = 152.288,90$$

$$SQFaltaaj. = SQTrat - SQ\text{Re } g = 2.134,6$$

Avaliação da precisão experimental:

$$C.V.(\%) = \frac{\sqrt{QM\,\text{Re}\,s}}{\hat{m}}\,x100 = \frac{\sqrt{445{,}94}}{213{,}25}\,x100 = 9{,}90\%$$

Teste para falta de ajuste:

$$F_{calc} = \frac{QMFaltaaj.}{QM\,\text{Re}\,s} = \frac{533{,}65}{445{,}94} = 1{,}20 \qquad F_{tab} = F_{5\%(\ 4{,}18)} = 2{,}93$$

$$H_0 : \beta_1 = 0 \ \ vs \ H_a : \beta_1 \neq 0$$

$$F_{calc} = \frac{QM\,\text{Re}\,g}{QM\,\text{Re}\,s} = \frac{152.288{,}9}{445{,}94} = 341{,}50 \qquad F_{tab} = F_{5\%(\ 1{,}18)} = 4{,}41$$

Coef. de determinação:

$$R^2 = \frac{SQ\text{Re}\,g}{SQTotal} = \frac{152.288{,}9}{162.450{,}5} = 0{,}9374 = 93{,}74\%$$

Coef. de determinação em SQTrat:

$$R^2 = \frac{SQ\text{Re}\,g}{SQTratl} = \frac{152.288{,}9}{154.423{,}5} = 0{,}9862 = 98{,}62\%$$

Exemplo 4: Uma pesquisadora realizou um ensaio para avaliar a extração de fenóis presente nas especiarias cebola. Para tanto utilizou diferentes concentrações de etanol (0, 20, 40, 60, 80 e 100%) diluído em água. Organizou o trabalho num DIC com quatro repetições.

Etanol (%)	Repetições				Totais Trat	Médias Trat
	R_1	R_2	R_3	R_4		
0	70	60	55	35	220	55
20	195	170	165	206	736	184
40	195	235	215	243	888	222
60	305	275	315	297	1192	298
80	285	315	325	327	1252	313
100	330	315	285	278	1208	302

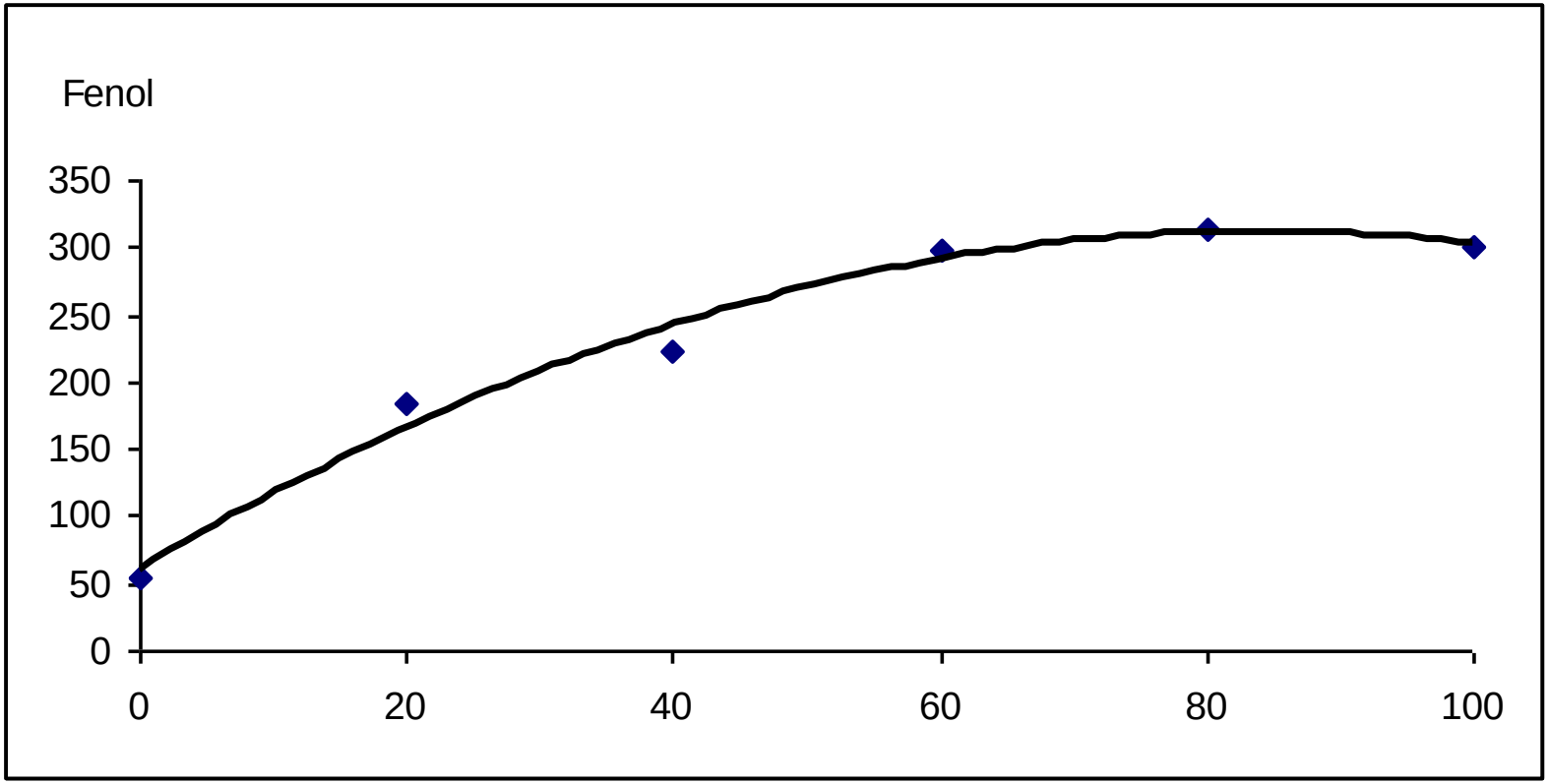
Modelo proposto =>

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \varepsilon_i$$

Equação estimada =>

$$\hat{Y}_i = 60{,}571 + 5{,}961 X_i - 0{,}0354 X_i^2$$

Representação gráfica



Análise de variância da regressão

FV	GL	SQ	QM	F	P (F > f) %
Regressão	2	194.624,21	97.312,11	249,73	0,000
Falta de ajuste	3	3.359,79	1.119,93	2,87	0,065
(Tratamentos)	[5]	[197.984,00]	39.596,80	101,62	0,000
Resíduo	18	7.014,00	389,67		
Total	23	204.998,00			

Avaliação da precisão experimental:

$$C.V.(\%) = \frac{\sqrt{QM\,\text{Re}\,s}}{\hat{m}}\,x100 = \frac{\sqrt{389{,}67}}{229{,}0}\,x100 = 8{,}62\%$$

Teste para falta de ajuste:

$$F_{calc} = \frac{QMFaltaaj.}{QM\,\text{Re}\,s} = \frac{1.119{,}93}{389{,}67} = 2{,}87 \qquad F_{tab} = F_{5\%(\ 4{,}18)} = 2{,}93$$

$$H_0 : \beta_1 = \beta_2 = 0 \ \ vs\ H_a : não\ H_0$$

$$F_{calc} = \frac{QM\,\text{Re}\,g}{QM\,\text{Re}\,s} = \frac{97.312{,}11}{389{,}67} = 249{,}73 \qquad F_{tab} = F_{5\%(\ 1{,}18)} = 4{,}41$$

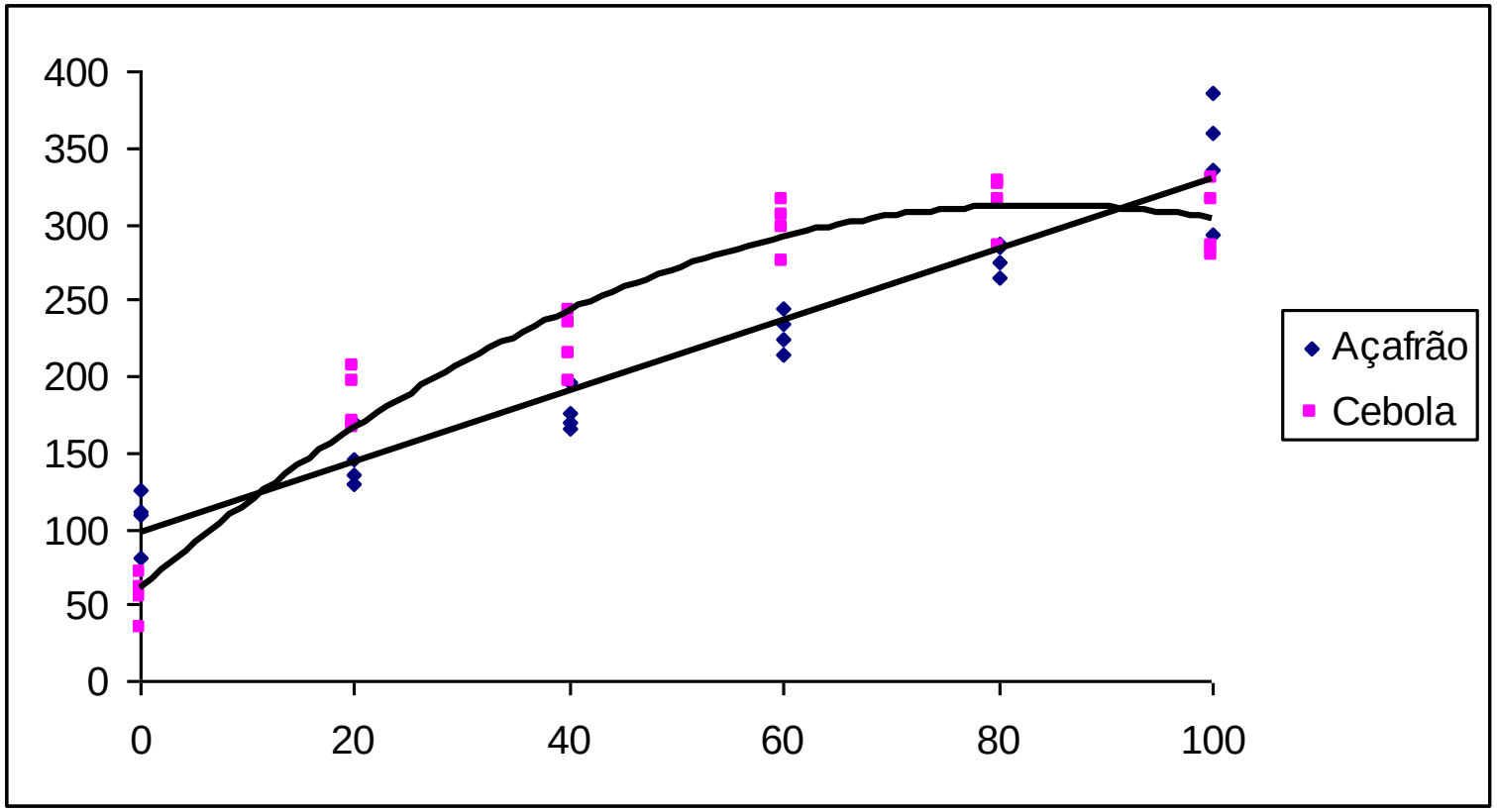
Coef. de determinação:

$$R^2 = \frac{SQ\text{Re}\,g}{SQTotal} = \frac{194.624{,}21}{204.998{,}00} = 0{,}9494 = 94{,}94\%$$

Coef. de determinação em SQTrat:

$$R^2 = \frac{SQ\text{Re}\,g}{SQTrat} = \frac{194.624{,}21}{197.984{,}0} = 0{,}9830 = 98{,}30\%$$

Representação gráfica para açafrão e cebola



REGRESSÂO MÚLTIPLA

O Modelo

Um modelo de regressão múltipla é um modelo em que uma resposta variável $Y(aleatória)$ está ligada a p$\geq$ 1 variáveis explanatórias (fixadas) ou regressoras $X_1, X_2, \dots, X_p$ e um termo aleatório (erro ou resíduo)

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \epsilon_i \quad (i = 1, \dots, n)$$

$\varepsilon_i \sim$ N(0; σ^2) – independentes (hipótese de normalidade necessária para inferência sobre $\beta_j - coeficientes\ de\ regressão$)

É conveniente:

$$y_i = \beta_0 x_{i0} + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \epsilon_i = \sum_{j=0}^{p} \beta_j x_{ij} + \epsilon_i \quad (i = 1, \dots, n)$$

tal que ($x_{i0} = 1$ para todo i)

$$Y = (y_1, y_2, \dots, y_n)' \ (n \times 1)$$

$$\mathrm{X} = \begin{pmatrix} 1 & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ 1 & \cdots & x_{np} \end{pmatrix} \quad (n \times [p+1]); \quad (B = \beta_0, \beta_1, \dots, \beta_p)' \quad ([p+1] \times 1)$$

$$\mathrm{E} = (\epsilon_1, \epsilon_2, \dots, \epsilon_n)'$$

$$Y = \mathrm{X}\, B + \mathrm{E} \qquad E(Y) = \mathrm{X}\, B \quad e \quad cov(Y) = \sigma^2 I$$

REGRESSÂO MÚLTIPLA(cont.)

Ajustamento do Modelo (método dos mínimos quadrados)

$$S(B) = \sum_{i=1}^{n} \varepsilon_i{}^2 = \varepsilon' \varepsilon = (Y - \mathrm{X}\, B)' (Y - \mathrm{X}\, B)$$
$$= Y'Y - 2B'X'Y + B'X'XB.$$

Substituindo B por $\hat{B}$ e derivando $S(\hat{B})$ em relação a cada β_j $e\ igualando\ a\ zero$ temos:

$\frac{\partial}{\partial \hat{\mathrm{B}}}(\varepsilon' \varepsilon) = -2X'Y + 2X'X\hat{B} = 0 \Rightarrow\ X'X\hat{B} = X'Y$,daí,

$\hat{B} = (X'X)^{-1}X'Y$ (estimador de mínimos quadrados de B).

(a) $E(\hat{B}) = B$
(b) $cov(\hat{B}) = \sigma^2 (X'X)^{-1}$

• Variâncias dos $\widehat{\beta_j}$ − os elementos da diagonal de $cov(\hat{B})$

• Covariâncias entre os pares $\widehat{\beta_j}$, $\widehat{\beta_k}$ − os elementos fora da diagonal

• Erro padrão dos $\widehat{\beta_j}$ − a raiz quadrada dos elementos da diagonal

• Se os elementos fora da diagonal de $(X'X)$ são todos zeros ⇒ os elementos da inversa são zeros fora da diagonal e os da diagonal o inverso dos elementos da diagonal de ⇒ os $\widehat{\beta_j}$ são não correlacionados e as regressoras (explanatórias) são ortogonais.

• benefícios adicionais surgem quando as variáveis são ortogonais.

AVALIANDO A REGRESSÂO

$Y = (y_1, y_2, \dots, y_n)'$ – valores observados

$\hat{Y} = (\widehat{y_1}, \widehat{y_2}, \dots, \widehat{y_n})'$ – valores ajustados $\Rightarrow \hat{Y} = X\hat{B} = X(X'X)^{-1}X'Y$

$\hat{Y} = HY \cdots H = X(X'X)^{-1}X'$

Vetor dos resíduos : $e = (e_1, e_2, \dots, e_n)' \Rightarrow$

$e = Y - \hat{Y}$

$= Y - X\hat{B}$

$= Y - X(X'X)^{-1}X'Y$

$= Y - HY$

$= (I - H)Y$

• Variação de $y(variável\ resposta)$ em torno da sua média (Soma de quadrado total)

$$\boldsymbol{S_{YY} = \sum(y_i - \bar{y})^2 = \sum y_i^2 - n\bar{y}^2 = Y'Y - n\bar{y}^2}$$

• Soma de quadrados devido a regressão

$\hat{Y}'\hat{Y} = \hat{B}'X'\ X\hat{B} = \hat{B}'X'\ X((X'X)^{-1}X'Y) = \hat{B}'X'Y$

$$\boldsymbol{SS_R = \sum(\hat{y}_i - \bar{y})^2 = \hat{B}'X'Y - n\bar{y}^2 \cdots \hat{Y}'\hat{Y} = \hat{B}'X'Y}$$

• Soma de quadrados dos resíduos

$$\boldsymbol{SS_E = \sum(y_i - \hat{y}_i)^2 = Y'Y - \hat{B}'X'Y}$$

• decomposição da soma de quadrado total

$$\boldsymbol{S_{YY} = Y'Y - n\bar{y}^2 = (\hat{B}'X'Y - n\bar{y}^2) + (Y'Y - \hat{B}'X'Y)}$$

$$\boldsymbol{S_{YY} = SS_R + SS_E}$$

AVALIANDO A REGRESSÂO (cont.)

Sumarizando:

Fonte de Variação	Soma de Quadrados	Graus de Liberdade	Quadrado Médio
Regressão	$\hat{B}'X'Y - n\bar{y}^2$	p	$MS_R = \frac{SS_R}{P}$
Residual	$Y'Y - \hat{B}'X'Y$	$n-p-1$	$MS_E = \frac{SS_E}{(n-p-1)}$
Total	$\boldsymbol{Y'Y - n\bar{y}^2}$	$\boldsymbol{n-1}$	

• Teste de Hipóteses

$H_0: \beta_1 = \beta_2 = \cdots \beta_p = 0$

$$\Rrightarrow \qquad \frac{MS_R}{MS_E} = F_{(p,(n-p-1))}$$

H_a : *no mínimo um* $\beta_j \neq 0$

$R^2 = \frac{\hat{B}'X'Y - n\bar{y}^2}{Y'Y - n\bar{y}^2}$ (correlação entre y_i e $\hat{y}_i$)

INFERÊNCIAS SOBRE OS PARÂMETROS INDIVIDUAIS DA REGRESSÂO

$E(\hat{B}) = B \qquad cov(\hat{B}) = \sigma^2 (X'X)^{-1}$

Var($\widehat{\beta_j}$) = $\sigma^2 c_{jj}$, onde c_{jj} é o elemento na $((j+1, j+1) -$ ésima posição de $(X'X)^{-1}$ para $j = 0,1, \dots, p$.

Assumindo a normalidade de ϵ_i no modelo $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \epsilon_i$ $(i = 1, \dots, n)$ segue que y_i e $\widehat{\beta_j}$ são também normalmente distribuídos, isto é:

$\widehat{\beta_j} \sim N(\beta_j, \sigma^2 c_{jj})$.

INFERÊNCIAS SOBRE OS PARÂMETROS INDIVIDUAIS DA REGRESSÂO (cont.)

Assim

$\frac{\widehat{\beta_j}-\beta_j}{\sqrt{\sigma^2 c_{jj}}} \sim N(0,1)$ e $\frac{\widehat{\beta_j}-\beta_j}{\sqrt{s^2 c_{jj}}} \sim t_{n-p-1}$ onde $\sqrt{s^2 c_{jj}} = s\sqrt{c_{jj}}$ é o erro padrão de $\widehat{\beta_j}$

Sob as hipóteses: $H_0: \beta_j = 0$ (X_j *não é importante*) e $H_1: \beta_j \neq 0$ (X_j *é importante*) nos podemos usar o teste estatístico

$$T_j = \frac{\widehat{\beta_j}}{\sqrt{s^2 c_{jj}}}$$

e rejeitar H_0 em favor de H_1 no nível de $100\alpha\%$ se o valor calculado de T_j encontra-se fora do intervalo $(-t_{\frac{\alpha}{2}.n-p-1}; t_{\frac{\alpha}{2}.n-p-1})$, daí, o intervalo de confiança a 100 (1-α)% para β_j será

$$IC_{\beta_j} = (\beta_j \pm t_{\frac{\alpha}{2}.n-p-1} s\sqrt{c_{jj}})$$

PREDIÇÃO

Para uma determinação conhecida $(x_{a1}, x_{a2}, \ldots, x_{ap})$ das variáveis explicativas podemos escrever:

$\hat{y}=\hat{\mu} = x_a^{'} \hat{\beta}$ (estimativa pontual $(\hat{\mu})$ da média de todos os valores de $y(\mu)$ para um particular vetor (determinação) das variáveis regressoras ou uma previsão individual de y $(\hat{y})$.

Demonstra-se que:

Estimativa para var $(\hat{\mu})$ = $s^2 x_a^{'} (X^{'}X)^{-1} x_a \Rightarrow$ ErroPadrão $(\hat{\mu})$ =$s\sqrt{(x_a^{'} (X^{'}X)^{-1} x_a}$, daí,

$$IC_{\mu} = (\hat{\mu} \pm t_{\frac{\alpha}{2}.n-p-1} s\sqrt{(x_a^{'} (X^{'}X)^{-1} x_a})$$

$$IC_{\hat{y}} = (\hat{y} \pm t_{\frac{\alpha}{2}.n-p-1} s\sqrt{1 + (x_a^{'} (X^{'}X)^{-1} x_a}$$

ALGUNS POSSÍVEIS PROBLEMAS

$X'X$ não é inversível - sigular (alguma dependência linear entre as variáveis regressoras)

Suprimir algumas variáveis regressoras para chegar a não singularidade de $X'X$

$X'X$ pode ser invertida mas é quase singular(relação aproximadamente linear entre algumas ou todas as variáveis regressoras (multicolinearidade). Os principais problemas:

• Erro padrão dos coeficientes muito grande, portanto não são confiáveis como estimadores do modelo;

• Instabilidade no modelo ajustado (o que significa que uma pequeno erro numa observação, ou supressão de uma observação do conjunto de dados irá produzir um modelo ajustado muito diferente).

• Dificuldades surgem na seleção de variáveis (verificar técnicas de seleção de variáveis)

Para obter o diagnóstico desta condição regredimos cada X_j sobre todos os outros X_s (a variável que usada como resposta pode ser suprimida).

Um problema relacionado é quando as variáveis regressoras tem unidades de medida que variam muito, daí podemos ter pouca precisão no cálculo de $(X'X)^{-1}$, neste caso é dita ser mal condicionada(solução-padronizar as variáveis)

PRESSUPOSTOS DO MODELO

• Para verificar a hipótese de normalidade dos resíduos, plotar os resíduos padronizados contra os valores acumulados da inversa da distribuição normal e verificar se os mesmos encontram-se próximo a faixa linear.

• Para verificar a hipótese de variância constante, plotar os resíduos padronizados (modelo ajustado) que devem aparecer de forma aleatória em torno de zero.

• Para verificar a hipótese de independência, plotar os resíduos padronizados contra as obsevações na ordem em que foram tomadas. Mais uma vez, dispersão aliatória de pontos, indica que o pressuposto é válido, mas uma tendência visível sugere que ela é violada.