ANÁLISE DA VARIÂNCIA DA REGRESSÃO

PROCEDIMENTO GERAL DE REGRESSÃO

Em um modelo de análise de variância, como no DIA, o fator em estudo pode ser *quantitativo* ou *qualitativo*.

FATOR QUANTITATIVO: é aquele cujos níveis podem ser associados com pontos em uma escala numérica, tais como Temperatura, Pressão, Tempo, Doses (adubo, medicamento, etc.) .

FATOR QUALITATIVO: é aquele cujos níveis não podem ser colocados em ordem de magnitude, tais como Variedade, Raça, Linhagem, Material. Não existe razão para ordena-los em qualquer ordem numérica particular.

Ambos os tipos de fatores são tratados identicamente na análise de variância dos dados de um ensaio. O pesquisador está interessado em determinar as diferenças, se alguma, entre os níveis dos fatores.

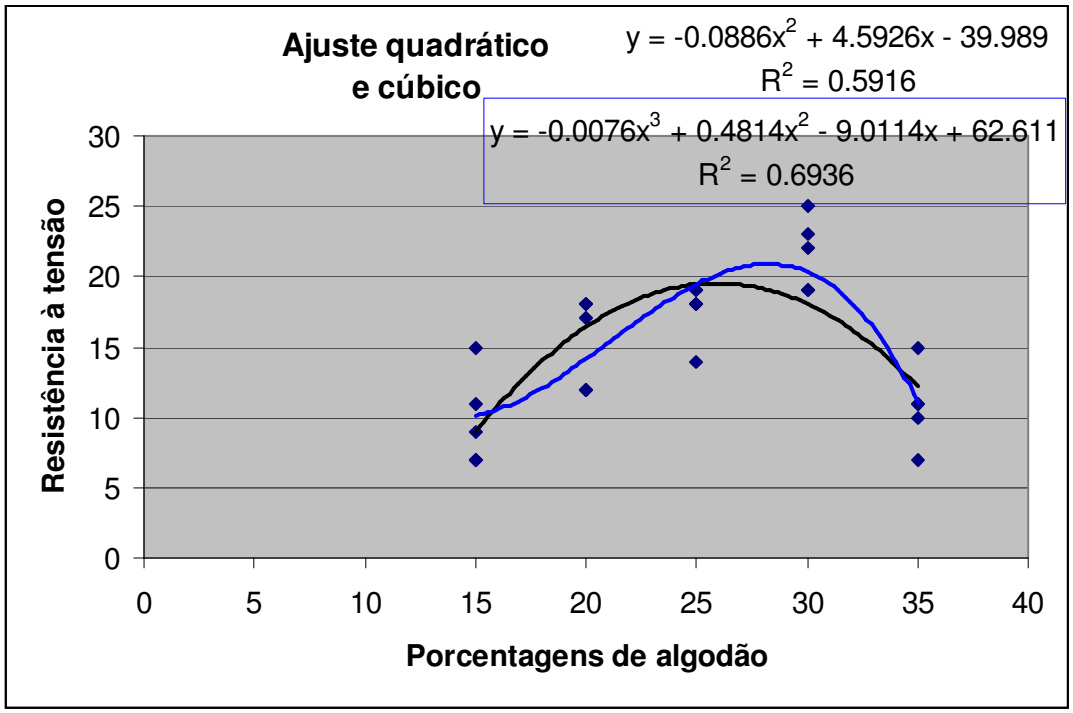
Se o fator é qualitativo, como Variedade, não tem sentido considerar a resposta para um nível intermediário do fator. Entretanto, com um fator quantitativo, como Dose, o pesquisador está usualmente interessado na amplitude de valores usados como níveis desse fator. Particularmente, na resposta de um nível intermediário do fator. Por exemplo, se o fator é Tempo, e os níveis 1.0h, 2.0h e 3.0h são usados no ensaio, o pesquisador pode está interessado na resposta a 2.5h. Então, o pesquisador está freqüentemente interessado em desenvolver uma equação de interpolação a partir dos dados.

Exemplo:

Um pesquisador está interessado em determinar se variando o conteúdo de algodão em uma fibra sintética, afeta a tensão de resistência, e ele executou um ensaio no DIA com cinco níveis de algodão (porcentagens) e cinco repetições.

	Dados de Resistência à Tensão (lb/pol^2)						
Porcentagem	Repetições						
De Algodão	1	2	3	4	5	Totais	Médias
15	7	7	15	11	9	49	9,8

20	12	17	12	18	18	77	15,4
25	14	18	18	19	19	88	17,6
30	19	25	22	19	23	108	21,6
35	7	10	11	15	11	54	10,8
						G=376	$\hat{\mu} = 15{,}04$



Do exame desse gráfico, fica claro que a relação entre Resistência à Tensão e porcentagem de algodão não é linear. Como uma primeira tentativa pode-se ajustar uma equação quadrática aos dados, digamos

$$y=a_o+a_1x+a_2x^2+e$$

onde a_0, a_1, e a_2 são parâmetros a serem estimados e *e* é o resíduo de regressão (tudo que não é explicado pela equação ajustada). O ajuste de mínimos quadrados forneceu a primeira equação. Esse ajuste parece não ser muito satisfatório porque ele dramaticamente subestima a resposta em x=30% de algodão e superestima a resposta em x=25% de algodão. Talvez, uma melhora possa ser obtida adicionando um termo cúbico em x. O ajuste cúbico é mostrado no gráfico, na segunda equação (em azul). O modelo cúbico parece ser superior ao modelo quadrático pois ele fornece um melhor ajuste em x=25 e x=30% de algodão.

Em geral, procura-se ajustar uma equação polinomial de menor ordem (modelos parcimoniosos) que adequadamente descreva os dados. Neste exemplo, o modelo cúbico polinomial parece ajustar melhor que o quadrático, e assim uma complexidade extra do modelo cúbico é justificada.

Selecionar a ordem do polinômio não é sempre fácil, entretanto, é relativamente fácil superajustar, isto é, adicionar alta ordem polinomial mas que não melhora o ajuste significativamente, aumentando a complexidade do modelo e prejudicando a sua utilidade como um preditor ou equação de interpolação.

POLINÔMIOS ORTOGONAIS

Na situação onde os níveis dos fatores são igualmente espaçados, o ajuste de modelos polinomiais pelo método de mínimos quadrados é grandemente simplificado. O procedimento faz uso dos coeficientes para contrates ortogonais (v. Tabela X do apêndice de Montgomery(1991) ou Gomes(2000) ou Neter e Wasserman(1974)). Em adição ao ajuste da equação polinomial de mínimos quadrados, obtém-se o efeito e Soma de quadrados: linear, quadrático, cúbico, quarta ordem etc. para o fator (Tratamento). Isto permite estimar a contribuição de cada termo para o polinômio a ser testado. É possível extrair o efeito polinomial até a ordem I-1 (gl do fator ou do tratamento) se existem I níveis do fator envolvido no experimento.

Exemplo: O processo é ilustrado na tabela abaixo usando os dados de Algodão.

		Coeficientes (c_i) para contraste ortogonal(caso de 5 níveis)			
Porcentagem de Algodão	Totais de Tratamentos	Linear	Quadrático	Cúbico	Quarta ordem
15	49	-2	2	-1	1
20	77	-1	-1	2	-4
25	88	0	-2	0	6
30	108	1	-1	-2	-4
35	54	2	2	1	1
Efeitos: $\left(\sum_{i=1}^{I} c_i T_i\right)$		41	-155	-57	-109

S.Q.: $\left[\frac{\left(\sum_{i=1}^{I} c_i T_i\right)^2}{J\sum_{i=1}^{I} c_i^2}\right]$	$\frac{(41)^2}{5(10)} =$ 33,62	$\frac{(-155)^2}{5(14)} =$ 343,21	$\frac{(-57)^2}{5(10)} =$ 64,98	$\frac{(-109)^2}{5(70)} =$ 33,95

Para esses dados, o fator independente "Porcentagem de Algodão", é igualmente espaçado nos cinco níveis. As somas de quadrados para os efeitos: Linear, Quadrático, Cúbico e Quarta Ordem do fator, formam uma partição ortogonal da Soma de Quadrados de Tratamentos e pode ser incorporada na ANVA. Cada efeito tem um grau de liberdade (pois é um contraste) e pode ser testado comparando suas respectivas Soma de Quadrados ao Quadrado Médio do Resíduo.

Análise da variância para os dados de Algodão

Fonte de Variação	Soma de Quadrados	Graus de Liberdade	Quadrado Médio	F
Tratamentos	475,76	4	118,94	14,76*
(Linear)	(33,62)	1	33,62	4,17
(Quadrático)	(343,21)	1	343,21	42,58**
(Cúbico)	(64,98)	1	64,98	8,06*
(Quarta Ordem)	(33,95)	1	33,95	4,21
Resíduo	161,20	20	8,06	
Total	636,96	24	----------	

Do quadro da ANVA acima, pode-se ver que os efeitos Quadrático e Cúbico da % de Algodão são estatisticamente significantes quando comparados a $F_{(1;20;\,0,05)}$=4,35. Dessa forma, deve-se ajustar aos dados um polinômio do terceiro grau, como

$$y = a_0 + a_1P_1(x) + a_2P_2(x) + a_3P_3(x) + \varepsilon$$

onde $P_u(x)$ é a u-ésima ordem do polinômio ortogonal, o qual implica que se existem I níveis de x têm-se $\sum_{j=1}^{I} P_u(x_j)P_s(x_j) = 0$. Os cinco primeiros polinômios ortogonais são:

$P_0(x)=1$

$$P_1(x) = \lambda_1\left[\frac{(x-\bar{x})}{d}\right]$$

$$P_2(x) = \lambda_2\left[\left(\frac{x-\bar{x}}{d}\right)^2 - \left(\frac{I^2-1}{12}\right)\right]$$

$$P_3(x) = \lambda_3\left[\left(\frac{x-\bar{x}}{d}\right)^3 - \left(\frac{x-\bar{x}}{d}\right)\left(\frac{3I^2-7}{20}\right)\right]$$

$$P_4(x) = \lambda_4\left[\left(\frac{x-\bar{x}}{d}\right)^4 - \left(\frac{x-\bar{x}}{d}\right)^2\left(\frac{3I^2-13}{14}\right) + \frac{3(I^2-1)(I^2-9)}{560}\right]$$

onde d é a distância entre os níveis de x, I é o número de níveis e $\{\lambda_i\}$ são constantes tais que os polinômios tenham valores inteiros. Ver Tabela X no apêndice de Montgomery (1991), onde é listado os coeficientes dos polinômios ortogonais e os valores de λ_i, para $I \leq 10$.

Para os dados do exemplo, têm-se as estimativas de mínimos quadrados dos parâmetros do modelo polinomial:

$$\hat{a}_i = \frac{\sum yP_i(x)}{\sum J[P_i(x)]^2} \qquad i=0,1,...,I-1, \qquad \text{ou } \hat{a}_i = \frac{\sum c_i T_i}{J\sum c_i^2}$$

Assim,

$$\hat{a}_0 = \frac{\sum yP_0(x)}{\sum J[P_0(x)]^2} = \frac{\sum y}{25} = \frac{376}{5(5)} = 15{,}0400$$

$$\hat{a}_1 = \frac{\sum yP_1(x)}{\sum J[P_1(x)]^2} = \frac{41}{5(10)} = 0{,}8200$$

$$\hat{a}_2 = \frac{\sum yP_2(x)}{\sum J[P_2(x)]^2} = \frac{-155}{5(14)} = -2{,}2143$$

$$\hat{a}_3 = \frac{\sum yP_3(x)}{\sum J[P_3(x)]^2} = \frac{-57}{5(10)} = -1{,}1400$$

Caso se queira adicionar ou retirar termos do modelo, não é necessário recalcular os parâmetros que já estão no modelo devido a propriedade de ortogonalidade dos polinômios.

Desde que I=5 níveis de x e a distância entre eles é d=5, o modelo polinomial ortogonal fica:

$$\hat{y} = 15{,}04 + 0{,}82(1)\left(\frac{x-25}{5}\right) - 2{,}2143(1)\left[\left(\frac{x-25}{5}\right)^2 - \left(\frac{5^2-1}{12}\right)\right] - 1{,}14(5/6)\left[\left(\frac{x-25}{5}\right)^3 - \left(\frac{x-25}{5}\right)\left(\frac{3(5)^2-7}{20}\right)\right]$$

onde $\lambda_1 = \lambda_2 = 1$ e $\lambda_3 = 5/6$ é obtido da Tabela X. Esta equação pode ser simplificada para

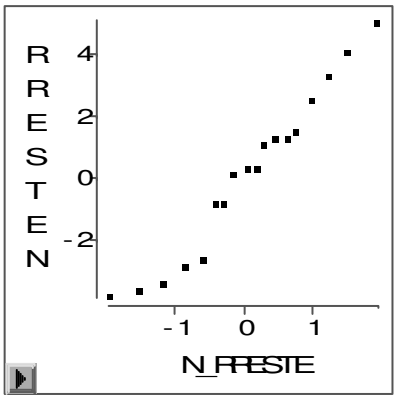
$$\hat{y} = 62{,}6111 - 9{,}01x + 0{,}4814x^2 - 0{,}00786x^3$$

a qual é exatamente a mesma equação encontrada anteriormente usando métodos de regressão geral pelo Excel ou SAS, considerando as aproximações computacionais.

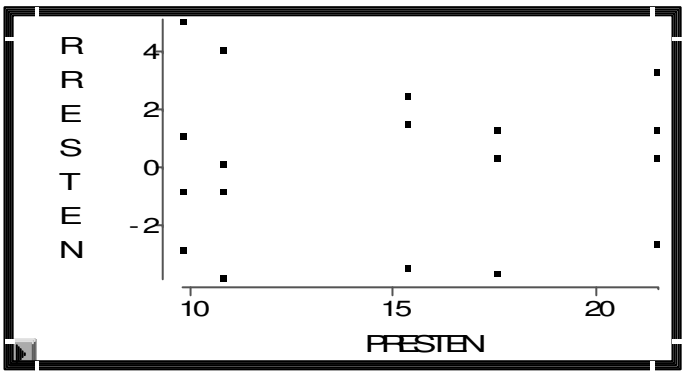
CHECANDO A ADEQUACIDADE DO MODELO DE REGRESSÃO

Análise de resíduo

Como no ajuste de qualquer modelo linear, análise dos resíduos de um modelo de regressão é necessário para determinar a adequacidade do ajuste de mínimos quadrados. É útil examinar um gráfico de probabilidade normal um gráfico de resíduo versus valores ajustados, e um gráfico de resíduos versus cada variável regressora. Em adição, se existem variáveis não incluídas no modelo que são de potencial interesse, então os resíduos devem ser representados contra esses fatores omitidos. Qualquer estrutura na qual uma representação indicaria que o modelo poderia ser melhorado pela adição daquele fator. Um gráfico de probabilidade normal dos resíduos do exemplo do algodão ajustado para os efeitos Linear, Quadrático Cúbico e de Quarta ordem é mostrado a seguir. Esse gráfico não indica qualquer séria violação da suposição de normalidade para o resíduo da análise.



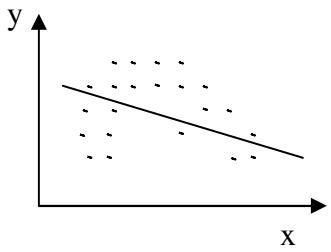
Os resíduos $\hat{e}_i$ são representados versus os valores ajustados $\hat{y}_i$. Esse gráfico não revela qualquer problema, tal que concluímos que a análise da variância da regressão com os efeitos linear, quadrático, cúbico e de quarta ordem é um ajuste adequado.



TESTE PARA FALTA DE AJUSTE DO MODELO DE REGRESSÃO

Modelos de regressão são frequentemente ajustados aos dados quando a relação verdadeira é conhecida. Naturalmente, gostaríamos de saber se a ordem do modelo assumido por tentativa está correto.

O perigo de usar um modelo de regressão é quando ele é uma aproximação pobre da verdadeira relação funcional, como mostrado na figura a seguir.



Obviamente, um modelo polinomial de grau dois ou maior deve ser usado para essa situação hipotética. O resultado é que um modelo muito pobre foi obtido para ajustar os dados.

Um teste para "qualidade do ajuste" de um modelo de regressão será aquele que testa a hipóteses:

H_o: O modelo adequadamente ajusta os dados;

H_1: O modelo não ajusta os dados.

O teste envolve a partição da Soma de Quadrados do Resíduo nos seguintes dois componentes: $SQ_{RESÍDUO} = SQ_{PURO} + SQ_{FALTA\ DE\ AJUSTE}$,

em que SQ_{PURO} é a soma de quadrados atribuída ao erro experimental "puro", e $SQ_{FALTA\ DE\ AJUSTE}$ é a soma de quadrados atribuída à falta do ajuste do modelo. Para calcular a SQ_{PURO} é preciso observações sobre y para no mínimo um nível de x. Suponha que temos n observações tais que:

$$y_{11}, y_{12}, \ldots, y_{1J1} = \text{observações em } x_1$$

$$y_{21}, y_{22}, \ldots, y_{2J2} = \text{observações em } x_2$$

$$\vdots$$

$$y_{I1}, y_{I2}, \ldots, y_{IJ1} = \text{observações em } x_I$$

Vemos que existe I distintos níveis de x. A contribuição para a Soma de Quadrados do Erro Puro no nível x_1, por exemplo, é

$$\sum_{j=1}^{J_1} \left(y_{1j} - \bar{y}_1\right)^2$$

A Soma de Quadrados Total para o Erro Puro será obtido somando a Equação anterior sobre todos os níveis de x. como:

$$SQ_{PURO} = \sum_{i=1}^{I} \sum_{j=1}^{J_i} \left(y_{ij} - \bar{y}_i\right)^2$$

Existem $n_e = \sum_{i=1}^{I} (J_i - 1) = n - I$ graus de liberdade associados com a Soma de Quadrados do Erro Puro, em que $n=J_1+J_2+...+J_I$. A Soma de Quadrados para Falta de Ajuste é simplesmente

$$SQ_{FALTA\ DE\ AJUSTE} = SQ_{RESÍDUO} - SQ_{PURO},$$

com $n - p - n_e = I - p$ graus de liberdade, em que p é o número de parâmetros no modelo que está sendo ajustado. O teste estatístico para falta de ajuste é então,

$$F_o = \frac{SQ_{FALTA\ DE\ AJUSTE}/(I-p)}{SQ_{ERRO\ PURO}/(n-I)} = \frac{QM_{FALTA\ DE\ AJUSTE}}{QM_{ERRO\ PURO}}$$

E rejeitamos a hipótese de adequacidade do modelo se $F_o > F_{\alpha,I-p,n-I}$.

Este procedimento teste pode ser facilmente introduzido dentro da análise de variância conduzida para a regressão. Se a hipótese nula da adequacidade do modelo é rejeitada, então o modelo deve ser abandonado e tentativas devem ser feitas para encontrar um modelo mais apropriado. Se H_o não é rejeitada, então não existe razão aparente para

duvidar da adequacidade do modelo, e $QM_{ERRO\ PURO}$ e $QM_{FALTA\ DE\ AJUSTE}$ são combinados para estimar σ^2.

EXEMPLO: no caso dos dados de Algodão, temos:

Fonte de variação	GL	Soma de Quadrados	F	Pr > F
Modelo	4	475.76000000	14.76	0.0001
LINEAR	(1)	33.62000000	4.17	0.0545
FALTAJU	(3)	442.14000000	18.29	0.0001
Error	20	161.20000000		
Total	24	636.96000000		

Fonte de variação	GL	Soma de Quadrados	F	Pr > F
Modelo	4	475.76000000	14.76	0.0001
LINEAR	(1)	33.62000000	4.17	0.0545
QUADRAT	(1)	343.21428571	42.58	0.0001
FALTAJU	(2)	98.92571429	6.14	0.0084
Error	20	161.20000000		
Total	24	636.96000000		

Fonte de variação	GL	Soma de Quadrados	F	Pr > F
Modelo	4	475.76000000	14.76	0.0001
LINEAR	(1)	33.62000000	4.17	0.0545
QUADRAT	(1)	343.21428571	42.58	0.0001
CUBICO	(1)	64.98000000	8.06	0.0101
FALTAJU	(1)	33.94571429	4.21	0.0535
Error	20	161.20000000		
Total	24	636.96000000		

```
data ALGODAO;
input
PORCENT     RESTEN;
LINEAR=PORCENT;
QUADRAT=PORCENT**2;
CUBICO=PORCENT**3;
QUARTA=PORCENT**4;
cards;
15    7
15    7
15    15
15    11
15    9
20    12
20    17
20    12
20    18
20    18
25    14
25    18
25    18
25    19
25    19
30    19
30    25
```

```
30    22
30    19
30    23
35    7
35    10
35    11
35    15
35    11
;
proc glm data=ALGODAO;
class PORCENT;
model RESTEN=PORCENT/ss3;
title 'ANVA DO DIA';

proc glm data=ALGODAO;
model RESTEN=LINEAR QUADRAT CUBICO QUARTA/ss1;
title 'ANVA DA REGRESSÂO POR POLINÔMIOS ORTOGONAIS';

proc reg data=ALGODAO;
model RESTEN=LINEAR QUADRAT CUBICO;
title 'AJUSTE DO MODELO POLINOMIAL CÚBICO';
run;

proc gplot data=ALGODAO;
plot RESTEN*PORCENT;
symbol1 v=dot i=rc c=blue;
run;
```

```
/*COM TESTE PARA FALTA DE AJUSTE*/
data ALGODAO;
input
PORCENT RESTEN;
LINEAR=PORCENT;
QUADRAT=PORCENT**2;
CUBICO=PORCENT**3;
QUARTA=PORCENT**4;
FALTAJU=PORCENT;
cards;
15      7
15      7
15      15
15      11
15      9
20      12
20      17
20      12
20      18
20      18
25      14
25      18
25      18
25      19
25      19
30      19
30      25
30      22
30      19
30      23
```

```
35      7
35      10
35      11
35      15
35      11
;
proc glm data=ALGODAO;
class PORCENT;
model RESTEN=PORCENT/ss3;
title 'ANVA DO DIA';

proc glm data=ALGODAO;
model RESTEN=LINEAR QUADRAT CUBICO QUARTA/ss1;
output out=RES r=RRESTEN p=PRESTEN;
title 'ANVA DA REGRESSÂO POR POLINÔMIOS ORTOGONAIS';
run;

proc reg data=ALGODAO;
model RESTEN=LINEAR QUADRAT CUBICO;
title 'AJUSTE DO MODELO POLINOMIAL CÚBICO';
run;


proc gplot data=ALGODAO;
plot RESTEN*PORCENT;
symbol1 v=dot i=rc c=blue;
run;

proc glm data=ALGODAO;
class FALTAJU;
model RESTEN=LINEAR FALTAJU/ss1;
title 'TESTE PARA FALTA DE AJUSTE LINEAR';
run;

proc glm data=ALGODAO;
class FALTAJU;
model RESTEN=LINEAR QUADRAT FALTAJU/ss1;
title 'TESTE PARA FALTA DE AJUSTE LINEAR+QUADRAT';
run;

proc glm data=ALGODAO;
class FALTAJU;
model RESTEN=LINEAR QUADRAT CUBICO FALTAJU/ss1;
title 'TESTE PARA FALTA DE AJUSTE LINEAR+QUADRAT+CUBICO';
run;
```

==========

O Coeficiente de Determinação

A quantidade

$$R^2 = \frac{SQ_{REGRESSÃO}}{SQ_{TOTAL}}$$

É chamado de coeficiente de determinação, e é frequentemente usado para julgar a adequacidade do modelo de regressão. Claramente $0<R^2 \leq 1$. Frequentemente nos referimos

ao R^2 como a proporção da variabilidade nos dados explicada pelo modelo de regressão. Se a regressora x é uma variável aleatória tal que x e y pode ser vista como variáveis aleatórias conjuntamente distribuídas, então R é exatamente a correlação simples entre y e x. Entretanto, se x não é uma variável aleatória, como no caso de ensaios com níveis quantitativos para um fator, então o conceito de correlação entre y e x fica indefinido.

A estatística R^2 deve ser usada com cautela desde que é sempre possível fazer R2 igual a unidade simplesmente adicionando termos suficientes ao modelo. Por exemplo, podemos obter um "perfeito" ajuste para n pontos de dados com um polinômio de grau (n-1). Também, R^2 sempre aumentará se adicionarmos uma variável ao modelo, porém isto não necessariamente significa que o novo modelo é superior ao anterior. A menos que a Soma de quadrados do resíduo no novo modelo seja reduzida por uma quantidade igual ao Quadrado Médio do Resíduo, o novo modelo terá uma Soma de Quadrados do resíduo maior do que o modelo antigo por causa da perda de um grau de liberdade do resíduo. Assim, o novo modelo será realmente pior do que o antigo.

EXERCÍCIO: Considere os dados de altura (cm) de plantas de alface (Lactuca sativa L.) em relação aos níveis de adubação orgânica (kg de esterco de boi/3,6m^2,) Silva e Fereira 1985, adaptado de Fereira 1991.

Tratamentos	1	2	3	4	5	6	Totais de Tratamentos
10	8,07	12,69	6,65	7,68	8,34	8,07	51,50
20	8,17	12,96	6,85	7,61	7,60	10,84	54,03
30	13,80	8,00	9,80	9,58	8,63	10,11	59,90
40	13,27	12,71	9,52	12,10	10,60	12,21	70,11

Pede-se:

a) Fazer a análise da variância da regressão por polinômios ortogonais.
b) Obter a equação polinomial que melhor se ajuste aos dados.
c) Faça o teste para falta de ajuste.
d) Calcule o coeficiente de determinação R^2, e interprete-o.