10 Regressão linear múltipla

José Luis Duarte Ribeiro
Carla ten Caten

Muitos problemas de regressão envolvem mais de uma variável regressora. Por exemplo, a qualidade de um processo químico pode depender da *temperatura*, *pressão* e *taxa de agitação*. Nesse caso há três variáveis regressoras.

O MODELO DA REGRESSÃO LINEAR MÚLTIPLA

O modelo geral da regressão linear múltipla é:

Eq 194 : $$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k + \varepsilon$$

O problema então é estimar o valor dos coeficientes β_i a partir de um conjunto de dados, conforme o esquema apresentado na Tabela 14.

Y	X_2	X_1	....	X_k
y_1	x_{12}	x_{11}	....	x_{1k}
y_2	x_{22}	x_{21}	....	x_{2k}
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
y_n	x_{n2}	x_{n1}	....	x_{nk}

Tabela 14 - Apresentação de um conjunto de dados.

Novamente, o método dos mínimos quadrados é usado para minimizar:

Eq 195: $$L = \sum \left[y_j - \left(b_0 + b_1 x_{1j} + ... + b_k x_{kj} \right) \right]^2$$

Observa-se que a aplicação do método dos mínimos quadrados fica simplificada se o modelo da Eq 194 é escrito como:

Eq 196: $$Y = \beta_0' + \beta_1 \left(X_1 - \bar{x}_1 \right) + ... + \beta_k \left(X_k - \bar{x}_k \right) + \varepsilon$$

nesse caso é fácil demostrar que:

Eq 197: $$\beta_0' = \beta_0 + \beta_1 \bar{x}_1 + ... + \beta_k \bar{x}_k$$

enquanto que os demais coeficientes $\beta_1,...,\beta_k$ ficam inalterados. O que está sendo feito é simplesmente eliminar o valor médio das variáveis regressoras. Além de simplificar a estimativa dos coeficientes, o uso do

modelo da **Eq 196** também facilita outras tarefas associadas a inferências.

Usando a **Eq 196** , a função a ser minimizada é:

Eq 198: $$L = \sum \left[y_i - \left(b_0' + b_1 \left(x_{1j} - \bar{x}_1 \right) + ... + b_k \left(x_{kj} - \bar{x}_k \right) \right) \right]^2$$

NOTAÇÃO MATRICIAL

Para lidar com o problema de regressão linear múltipla, é mais conveniente usar notação matricial, pois assim tem-se uma apresentação muito compacta dos dados, do modelo e dos resultados.

Em notação matricial o modelo da **Eq 196** aparece representado como:

Eq 199: $$Y = X\beta + \varepsilon$$

onde:

Eq 200: $$Y = \begin{bmatrix} Y_1 \\ . \\ . \\ . \\ Y_n \end{bmatrix}; \; X = \begin{bmatrix} 1 & (x_{11} - \bar{x}_1) & ... & (x_{k1} - \bar{x}_k) \\ . & . & & . \\ . & . & & . \\ . & . & & . \\ 1 & (x_{1n} - \bar{x}_1) & ... & (x_{kn} - \bar{x}_k) \end{bmatrix}; \; \beta = \begin{bmatrix} \beta_0' \\ . \\ . \\ . \\ \beta_k \end{bmatrix}; \; \varepsilon = \begin{bmatrix} \varepsilon_1 \\ . \\ . \\ . \\ \varepsilon_n \end{bmatrix}$$

Genericamente, tem-se que Y é o vetor n x 1 das observações, X é a matriz n x p com os níveis das variáveis regressoras, β é o vetor p x 1 com os coeficientes da regressão e ε é o vetor n x 1 com os erros aleatórios. (Sendo $p = k + 1$).

ESTIMATIVA DOS COEFICIENTES

Pode ser demonstrado que a aplicação do método dos mínimos quadrados conduz a seguinte solução:

Eq 201: $$b = \left(X'X \right)^{-1} X'Y$$

onde b é o vetor p x 1 com as estimativas dos coeficientes β. A solução da Eq 199 irá existir sempre que as variáveis regressoras forem linearmente independentes.

(*Nota*: as variáveis regressoras não serão independentes quando uma coluna da matriz X for uma combinação linear de outras colunas).

Exemplo 10.1

ver (Montgomery (1984)) Um distribuidor de cerveja está analisando seu sistema de distribuição. Especificamente ele está interessado em prever o tempo requerido para atender um ponto de venda. O engenheiro industrial acredita que os dois fatores mais importantes são o número de caixas de cerveja fornecidas e a distância do depósito ao posto de venda. Os dados coletados aparecem na Tabela 15.

X1: Nº de caixas	X2: Distância	Y: Tempo
10	30	24
15	25	27
10	40	29
20	18	31
25	22	25
18	31	33
12	26	26
14	34	28
16	29	31
22	37	39
24	20	33
17	25	30
13	27	25
30	23	42
24	33	40

Tabela 15 - Exemplo do distribuidor de cervejas.

Solução:

Escolhemos ajustar o seguinte modelo a esses dados:

Eq 202: $$Y = \beta_0' + \beta_1\left(X_1 - \bar{x}\right) + \beta_2\left(X_2 - \bar{x}\right) + \varepsilon$$

Desde que $\bar{x}_1 = 18$ e $\bar{x}_2 = 28$, esse modelo em notação matricial é:

$$\begin{bmatrix}24\\27\\29\\31\\25\\33\\26\\28\\31\\39\\33\\30\\25\\42\\40\end{bmatrix} = \begin{bmatrix}1 & -8 & 2\\1 & -3 & -3\\1 & -8 & 12\\1 & 2 & -10\\1 & 7 & -6\\1 & 0 & 3\\1 & -6 & -2\\1 & -4 & 6\\1 & -2 & 1\\1 & 4 & 9\\1 & 6 & -8\\1 & -1 & -3\\1 & -5 & -1\\1 & 12 & -5\\1 & 6 & 5\end{bmatrix} \begin{bmatrix}\beta_0'\\\beta_1\\\beta_2\end{bmatrix} + \begin{bmatrix}\varepsilon_1\\\varepsilon_2\\\varepsilon_3\\\varepsilon_4\\\varepsilon_5\\\varepsilon_6\\\varepsilon_7\\\varepsilon_8\\\varepsilon_9\\\varepsilon_{10}\\\varepsilon_{11}\\\varepsilon_{12}\\\varepsilon_{13}\\\varepsilon_{14}\\\varepsilon_{15}\end{bmatrix}$$

E usando as regras para produto e inversão de matriz, obtemos:

$$X'X = \begin{bmatrix}15 & 0 & 0\\0 & 504 & -213\\0 & -213 & 548\end{bmatrix}; \qquad X'Y = \begin{bmatrix}463\\345\\63\end{bmatrix}$$

e

$$(X'X)^{-1} = \begin{bmatrix}0{,}06667 & 0 & 0\\0 & 0{,}002374 & 0{,}0009228\\0 & 0{,}0009228 & 0{,}002183\end{bmatrix}$$

De forma que o vetor das estimativas dos coeficientes resulta:

$$b = \begin{bmatrix}b_0'\\b_1\\b_2\end{bmatrix} = (X'X)^{-1}X'Y = \begin{bmatrix}30{,}87\\0{,}8772\\0{,}4559\end{bmatrix}$$

E o modelo de regressão é:

$$\hat{Y} = 30{,}87 + 0{,}8772(X_1 - 18) + 0{,}4559(X_2 - 28)$$

ou

$$\hat{Y} = 2{,}315 + 0{,}8772X_1 + 0{,}4559X_2$$

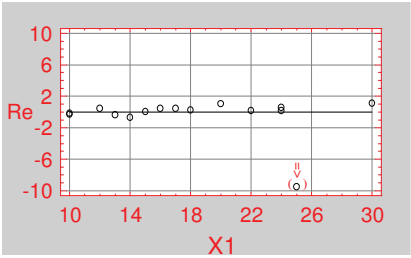
A tabela a seguir apresenta os valores observados, os valores previstos pelo modelo e os respectivos resíduos $r_j = Y_j - \hat{Y}_j$.

Y_j	$\hat{Y}_j$	$r_j = Y_j - \hat{Y}_j$
24	24,76	-0,76
27	26,87	0,13
29	29,32	-0,32
31	28,06	2,94
25	34,27	-9,27
33	32,23	0,77
26	24,69	1,31
28	30,09	-2,09
31	29,57	1,43
39	38,48	0,52
38	32,48	0,52
30	28,62	1,38
25	26,02	-1,02
42	39,11	2,89
40	38,41	1,59

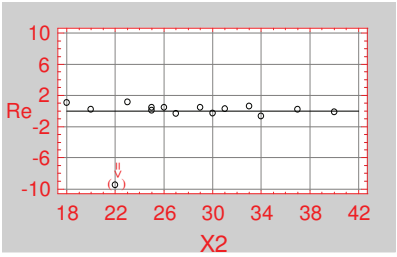
Tabela 16 - Valores observados, valores previstos e resíduos.

Para testar se o ajuste é adequado, os resíduos poderiam ser plotados em função de $\hat{Y}$, em função de X_1 ou em função de X_2. Os resíduos também poderiam ser plotados em papel de probabilidade, para testar a suposição de normalidade.

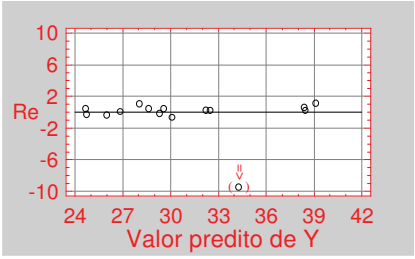
Resíduos x X_1



Resíduos x X_2



Resíduos x $\hat{Y}$



Papel de Probabilidade

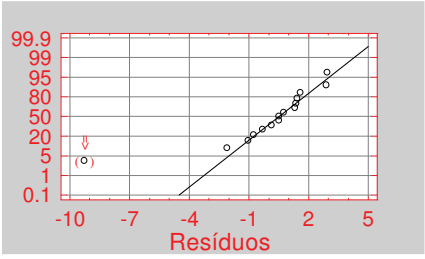


Figura 45 - Gráficos do distribuidor de cervejas.

Qualquer um desses gráficos iria evidenciar que a observação da linha 5 é, sem dúvida, um dado atípico.

Se houver registro de alguma causa especial que tenha afetado esta entrega em particular, essa observação poderia ser eliminada do conjunto e a análise poderia ser refeita, possivelmente fornecendo um modelo mais preciso.

Exemplo 10.2

(ver Montgomery (1984)) Esse exemplo ilustra o uso da Análise de Regressão em conjunto com Projeto de Experimentos.

O ganho em um processo químico está sendo estudado. O engenheiro escolheu 3 fatores (temperatura, pressão e concentração) e rodou um experimento fixando cada um desses fatores a dois níveis.

Os dados aparecem a seguir. Vejam que os níveis dos fatores foram codificados como -1 (nível baixo) e +1 (nível alto).

Ganho %	X_1 (Temp.)	X_2 Pressão)	X_3 (Concent.)
32	-1	-1	-1
36	-1	-1	1
57	-1	1	-1
46	1	-1	-1
65	1	1	-1
57	-1	1	1
48	1	-1	1
68	1	1	1

Tabela 17 - Valores observados em um processo químico.

Solução:

Escolhemos ajustar o seguinte modelo ($\overline{X}_i = 0$)

$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$

As matrizes $X'X$ e $X'Y$ resultam:

$$X'X = \begin{bmatrix} 8 & 0 & 0 & 0 \\ 0 & 8 & 0 & 0 \\ 0 & 0 & 8 & 0 \\ 0 & 0 & 0 & 8 \end{bmatrix} = 8\, I_4 \;; \qquad X'Y = \begin{bmatrix} 409 \\ 45 \\ 85 \\ 9 \end{bmatrix}$$

E como $X'X$ é diagonal, a sua inversa $(X'X)^{-1}=(1/8)I_4$. Assim as estimativas dos coeficientes resultam:

$$b = \begin{bmatrix} b_0 \\ b_1 \\ b_2 \\ b_3 \end{bmatrix} = \begin{bmatrix} 51{,}125 \\ 5{,}625 \\ 10{,}625 \\ 1{,}125 \end{bmatrix}$$

E o modelo de regressão é:

$$\hat{Y} = 51{,}125 + 5{,}625X_1 + 10{,}625X_2 + 1{,}125X_3$$

Nesse exemplo a matriz inversa é fácil de obter porque $X'X$ é diagonal. Há várias vantagens quando $X'X$ é diagonal. Os cálculos são mais fáceis e a estimativa dos coeficientes está livre de qualquer correlação *[Cov (b_i,b_j) = 0]*.

Se nós podemos escolher os níveis de X_i é vantajoso fazer essa escolha de modo a obter $X'X$ diagonal. Projetos de Experimentos que apresentam essa propriedade são chamados de *projetos ortogonais*. Um exemplo de projetos desse tipo é a classe dos projetos 2^k. Esses projetos têm sido usados com freqüência crescente no meio industrial.

MATRIZ DE VARIÂNCIAS E COVARIÂNCIAS

A matriz $(X'X)^{-1}$ é chamada de matriz de variâncias e covariâncias. É uma matriz simétrica de ordem p x p e seus elementos são usados na determinação das variâncias S_{ij}^2.

Usando a notação:

Eq 203:
$$(X'X)^{-1} = \begin{bmatrix} C_{00} & C_{01} & \dots & C_{0k} \\ C_{10} & C_{11} & \dots & C_{1k} \\ . & . & & . \\ . & . & & . \\ . & . & & . \\ C_{k0} & C_{k1} & \dots & C_{kk} \end{bmatrix}$$

É possível demonstrar que:

Eq 204: $\mathrm{Var}(b_i) = C_{ii}\, S^2 \qquad i = 0,\dots,k$

Eq 205: $\mathrm{Covar}(b_i,b_j) = C_{ij}\, S^2 \qquad i,j = 0,\dots,k$

onde S^2 é a variância residual, associada com os desvios em relação ao hiperplano do modelo de regressão:

Eq 206: $$S^2 = \sum \left(Y_j - \hat{Y}_j\right)^2 \Big/ (n-k-1) \;; j = 1,n$$

A partir da matriz de variâncias e covariâncias também é possível encontrar a matriz de correlação, uma vez que têm-se:

Eq 207: $$r_{ij} = C_{ij} \Big/ \sqrt{C_{ii} C_{jj}} \;; i,j = 0,\dots,k$$

onde, naturalmente, para $i = j$ tem-se $r_{ii} = 1$.

A matriz de correlações também é simétrica, de ordem p x p :

Eq 208:
$$K = \begin{bmatrix} 1 & r_{01} & \dots & r_{0k} \\ r_{10} & 1 & \dots & r_{1k} \\ . & . & & . \\ . & . & & . \\ . & . & & . \\ r_{k0} & r_{k1} & \dots & 1 \end{bmatrix}$$

A matriz de correlações R é útil para detectar problemas de multicolinearidade. Se um coeficiente r_{ij} qualquer fora da diagonal tiver módulo ≅ 1,0 teremos uma dependência entre as variáveis *independentes* i e j.

Nesse caso, a estimativa dos coeficientes associados às variáveis i e j estará comprometida. (Não é possível distinguir se o efeito sobre a variável de resposta se deve a variável regressora i ou j, uma vez que elas estão variando sempre no mesmo sentido).

O ideal é que a matriz de correlações seja diagonal, com zeros ou valores próximos de zeros nas posições fora da diagonal. Isso assegura estimativas *não-confundidas* dos diversos coeficientes β_i.

TESTES DE HIPÓTESE

Para construir os testes de hipótese relativos a regressão múltipla, vamos supor que os resíduos ε_j sigam o modelo normal com média 0 e variância S^2.

Há dois tipos de teste que podem ser feitos: testes individuais sobre a significância de cada parâmetro b_j e um teste global para o modelo.

Significância de cada parâmetro

Se os resíduos seguem o modelo normal, os parâmetros b_j também irão seguir esse modelo, ou seja:

Eq 209:
$$b_j \rightarrow N\left(\beta_j, \sigma_{bj}^2\right)$$

De modo que para testar as hipóteses

H_0: $\beta_j = 0$

H_1: $\beta_j \neq 0$

Usamos a distribuição de Student, calculando

Eq 210:
$$t_j = b_j / S_{bj}$$

Como sempre, a hipótese nula será rejeitada se

Eq 211:
$$\left|t_j\right| > t_{\alpha/2}\ ,\ n-k-1$$

Significância do modelo de regressão

Para testar a significância do modelo de regressão múltipla, usaremos o teste F. Os desvios $\left(Y_j - \overline{Y}\right)$ podem ser escritos na forma:

Eq 212: $\left(Y_j - \overline{Y}\right) = \left(Y_j - \hat{Y}_j\right) + \left(\hat{Y}_j - \overline{Y}\right)$

elevando ao quadrado e somando, obtemos:

Eq 213: $\sum\left(Y_j - \overline{Y}\right)^2 = \sum\left(Y_j - \hat{Y}_j\right)^2 + \sum\left(Y_j - \overline{Y}\right)^2$

uma vez que pode ser demonstrado que o produto cruzado é nulo. Dessa forma temos:

Eq 214: $S_{YY} = SQR + SQR_{eg}$

onde os correspondentes *GDL* valem:

Eq 215: (n-1) = (n-k-1) + (k)

de forma que as médias quadradas resultam:

Eq 216: MQR = SQR / (n-k-1)

Eq 217: $MQR_{eg} = SQR_{eg} / k$

e usamos,

Eq 218: $F = MQR / MQR_{eg}$

para testar a significância do modelo. A hipótese (inexistência de relação entre X e Y) deve ser rejeitada se resultar,

$F > F_{\alpha/2,\ k,\ n-k-1}$

para o cálculo das somas quadradas as seguintes fórmulas práticas podem ser usadas:

Eq 219: $$S_{YY} = \sum_{j=1}^{n} y_j^2 - \left(\sum_{j=1}^{n} y_j\right)^2 \Big/ n$$

Eq 220: $$SQR = S_{YY} - \sum_{i=1}^{k} b_i S_{iy}$$

Eq 221: $$SQR_{eg} = \sum_{i=1}^{k} b_i S_{iy}$$

Eq 222: onde os valores S_{iy} aparecem no vetor $X'Y$, ou seja,

$$Eq\ 223: \quad X'Y = \begin{bmatrix} \sum Y_j \\ S_{1y} \\ . \\ . \\ . \\ S_{ky} \end{bmatrix}$$

COEFICIENTES DE DETERMINAÇÃO PARA O MODELO DE REGRESSÃO MÚLTIPLA

A fórmula para o cálculo do coeficiente de determinação r^2 é a mesma apresentada ao final do capítulo 9, ou seja:

$$Eq\ 224: \quad r^2 = \frac{SQR_{eg}}{S_{YY}}$$

O coeficiente r^2 indica a percentagem da variabilidade total que é explicada pelo modelo de regressão. Se $r^2 = 1$, todas as observações estarão sobre o hiperplano definido pelo modelo. Se $r^2 = 0$, não há nenhuma relação entre a variável de resposta e as variáveis regressoras.

Exemplo 10.3

Para o problema da distribuição das caixas de cerveja, pede-se:

Apresente a matriz de variâncias e covariâncias e a matriz de correlação;

Calcule a variância residual $\boldsymbol{S^2}$ e a variância de $\boldsymbol{b_1}$, $\boldsymbol{S_{b1}^2}$;

Teste de significância de $\boldsymbol{b_1}$;

Teste a significância do modelo;

Calcule o coeficiente de determinação;

Solução:

A matriz de variâncias e covariâncias é a matriz $(X'X)^{-1}$, enquanto que a matriz de correlações é obtida dividindo os termos da matriz $X'X$ pelos correspondentes termos da diagonal. Assim,

$$(X'X)^{-1} = \begin{bmatrix} 0{,}06667 & 0 & 0 \\ 0 & 0{,}002374 & 0{,}0009228 \\ 0 & 0{,}0009228 & 0{,}002183 \end{bmatrix}$$

$$r = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0{,}405 \\ 0 & 0{,}405 & 1 \end{bmatrix}$$

Para calcular S^2 e S_{b1}^2, usamos:

$$S^2 = \sum \left(Y_i - \hat{Y}\right)^2 \Big/ (n-k-1) = 118{,}37/12 = 9{,}86$$

$$S_{b1}^2 = S^2 C_{11} = 9{,}86(0{,}002374) = 0{,}0234\,; \quad S_{b1} = 0{,}153$$

o teste de significância para b_1 é:

$t_1 = b_1 / S_{b1} = 0{,}8772 / 0{,}153 = 5{,}73$

$t_1 = 5{,}73 > t_{0{,}025;12} = 2{,}179 \Rightarrow$ rejeita-se a hipótese nula

O teste de significância para o modelo é feito usando a tabela ANOVA.

$S_{YY} = \Sigma y_j^2 - (\Sigma y_j)^2 / n = 449{,}73$

$SQR_{eg} = b_1 S_{1y} + b_2 S_{2y} = 331{,}36$

$SQR = S_{YY} - SQR_{eg} = 118{,}37$

Fonte	SQ	GDL	MQ	F
Modelo	331,36	2	165,58	16,80
Residual	118,37	12	9,86	
Total	449,73	14		

Tabela 18 - Tabela ANOVA

$F = 16{,}80 > F_{0{,}05;2;12} = 3{,}89;$ rejeita-se a hipótese nula

E nesse exemplo o coeficiente de determinação vale $SQR_{eg}/S_{YY} = 0{,}737$; ou seja 73,7% da variabilidade total no tempo de entrega é explicada pela relação que essa variável mantém com o número de caixas e a distância do posto de vendas.

PREVISÃO DE VALORES DE Y

Assim como o caso da regressão simples, a relação encontrada pode ser usada para a previsão de um valor médio ou individual de Y. Seja:

Eq 225:
$$X_0 = \begin{bmatrix} X_{10} \\ X_{20} \\ . \\ . \\ . \\ X_{k0} \end{bmatrix}$$

Eq 226:
$$\hat{Y}_0 = b_0 + b_1 X_{10} + \ldots + b_k X_{k0}$$

Pode ser demonstrado o intervalo de confiança de $100(1-\alpha)\%$ para um valor médio e individual de Y são, respectivamente:

Valor médio:

Eq 227: $\hat{Y}_0 \pm t_{\alpha/2,n-k-1}\ S^2 \left(X_0'(X'X)^{-1}\right)^{1/2}$

Valor individual.:

Eq 228: $\hat{Y}_0 \pm t_{\alpha/2,n-k-1}\ S^2 \left(1 + X_0'(X'X)^{-1} X_0\right)^{1/2}$

O fator que multiplica $t_{\alpha/2}$ nas fórmulas acima corresponde ao erro de previsão. A divisão desse fator por $\hat{Y}_0$ produz o coeficiente de variação da previsão.

ANÁLISE DAS SUPOSIÇÕES DO MODELO DE REGRESSÃO

Nas seções anteriores foi feita a suposição $\varepsilon \rightarrow N(0,\sigma^2)$, ou seja, supõe-se normalidade na distribuição dos resíduos e homogeneidade da variância residual. A suposição de normalidade dos resíduos pode ser testada por testes gráficos (papel de probabilidade) ou analíticos (teste do Chi-quadrado, Kolmogorov-Smirnov, etc.).

Para o teste de normalidade, usa-se os resíduos padronizados:

Eq 229: $R_j = \left[Y_j - \hat{Y}_j\right] / S^2$

onde:

Eq 230: $\hat{Y}_j = b_0 + b_1 X_{1j} + ... + b_k X_{kj}$

Para examinar se o erro padrão da estimativa é constante, analisa-se os gráficos $R_j \times \hat{Y}_j$ e $R_j \times X_i$.

Se a suposição de normalidade ou de homogeneidade não forem satisfeitas, muitas vezes é possível contornar o problema aplicando certas transformações matemáticas aos dados. Os resíduos também podem ser analisados para verificar a existência de dados atípicos.

REGRESSÃO POLINOMIAL

O modelo aditivo $Y = X\beta + \varepsilon$ é um modelo geral e pode ser usado para ajustar qualquer relação que seja linear com referência aos parâmetros desconhecidos β. Veja que a exigência de linearidade refere-se a β e não a X. Assim, o modelo pode ser usado para ajustar um polinômio de ordem k em uma variável:

Eq 231: $Y = \beta_0 + \beta_1 x + \beta_2 x^2 + ... + \beta_k x^k + \varepsilon$

ou então para ajustar um polinômio de segundo grau em duas variáveis:

Eq 232: $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_2^2 + \beta_5 x_1 x_2 + \varepsilon$

O uso do modelo $Y = X\beta + \varepsilon$ para ajustar um polinômio é ilustrado a seguir.

Exemplo 10.4

(ver Montgomery (1984)) Pede-se para ajustar o modelo $Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon$ aos dados que aparecem a seguir:

Tabela 19 - Valores observados do exemplo

x	1,0	1,2	1,4	1,6	1,8	2
y	6,15	7,90	9,40	10,50	11,00	14,00

Em notação matricial, usando $X - \overline{X}$, tem-se:

Eq 233: $$Y = \begin{bmatrix} 6,15 \\ 7,90 \\ 9,40 \\ 10,50 \\ 11,00 \\ 14,00 \end{bmatrix} \qquad \mathrm{X} = \begin{bmatrix} 1 & -0,5 & 0,25 \\ 1 & -0,3 & 0,09 \\ 1 & -0,1 & 0,01 \\ 1 & 0,1 & 0,01 \\ 1 & 0,3 & 0,09 \\ 1 & 0,5 & 0,25 \end{bmatrix} \qquad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix}$$

As matrizes $X'X$ e $X'Y$ resultam:

Eq 234: $$X'X = \begin{bmatrix} 6,0 & 0,0 & 0,7 \\ 0,0 & 0,7 & 0,0 \\ 0,7 & 0,0 & 0,1414 \end{bmatrix}; \quad \mathrm{X}'Y = \begin{bmatrix} 58,95 \\ 4,965 \\ 6,938 \end{bmatrix}$$

De modo que as estimativas de β são:

Eq 235: $$b = (X'X)^{-1} X'Y = \begin{bmatrix} 0,3945 & \phi & -1,9527 \\ \phi & 1,4286 & \phi \\ -1,9527 & \phi & 16,737 \end{bmatrix} \begin{bmatrix} 58,95 \\ 4,965 \\ 6,938 \end{bmatrix} = \begin{bmatrix} 9,70 \\ 7,08 \\ 1,00 \end{bmatrix}$$

Assim o modelo de regressão:

Eq 236: $$\hat{Y} = 9,70 + 7,08\left(X - \overline{X}\right) + 1,00\left(X - \overline{X}\right)^2$$

ou

Eq 237: $$\hat{Y} = 1,33 + 4,08X + 1,00X^2$$

Esse método geral pode ser usado para ajustar dados que tenham um formato qualquer. No entanto, se os níveis das variáveis regressoras forem eqüidistantes, então o uso de polinômios ortogonais simplifica bastante o esforço de cálculo. O uso de polinômios ortogonais aparece descrito em Montgomery & PecK (1991) e Nanni & Ribeiro (1991).

Comentários em relação aos modelos polinomiais:

(1) Os polinômios são muito úteis para fornecer uma aproximação para relações não lineares complexas e desconhecidas. Esse tipo de aplicação aparece com freqüência na prática.

(2) É importante manter a ordem do polinômio tão baixa quanto possível. Polinômios de ordem mais alta (k > 2) devem ser evitados, a menos que hajam justificativas técnicas para o seu uso.

(3) Um modelo de ordem mais baixa usando variáveis transformadas é

sempre preferível à modelos de ordem mais alta na métrica original.

(4) Vale lembrar que sempre pode ser obtido um polinômio de ordem $n-1$ que ajusta-se *perfeitamente* aos dados. Tal modelo não ajudaria em nada para a compreensão do fenômeno em estudo e nem tampouco seria um bom estimador.

(5) Extrapolações com polinômios devem ser feitas com muito cuidado. Além do intervalo investigado, os polinômios podem apresentar um comportamento estranho, girando na direção oposta do esperado.

(6) Na medida que cresce a ordem do polinômio, a matriz $X'X$ torna-se mal condicionada e a precisão das estimativas diminui. Esse problema é aliviado quando se centra as variáveis regressoras, isto é, quando se usa $\left(X_{ij} - \overline{X}_i\right)$.

(7) A matriz $X'X$ também tende a tornar-se mal condicionada quando os valores de X estão limitados a um intervalo muito estreito. De forma geral, ampliando o intervalo de investigação, melhoram as estimativas dos coeficientes.

Exercícios

Exercício 10.1

A resistência de uma cera depende da quantidade de Etil-Vinil-Acetato (EVA) e da quantidade de Parafina adicionados à cera. Ajuste um modelo do tipo $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$ aos dados que aparecem a seguir

X1: EVA	4	4	6	6	8	8	4	4	6	6	8	8
X2: Paraf.	8	8	8	8	8	8	12	12	12	12	12	12
Y: Resist.	28,5	26,4	33,0	32,1	35,3	36,7	36,6	34,2	37,9	39,9	42,6	44,2

Exercício 10.2

Calcule o valor dos resíduos $R_j = Y_j - \hat{Y}_j$ para os dados do exercício anterior e a seguir analise esses resíduos plotando os gráficos:

$R_j \times \hat{Y}_j$, $R_j \times X_1$, $R_j \times X_2$.

Exercício 10.3

Ainda em relação aos dados do exercício 10.1, pede-se:

Apresente a matriz de variâncias e covariâncias e a matriz de correlações. Analise a matriz de correlações e indique se há indícios de mal condicionamento;

Calcule a variância residual S^2 e a variância de b_1 e b_2;

Teste de significância de b_1 e b_2;

Teste de significância do modelo;

Calcule o coeficiente de determinação e indique o seu significado técnico;

Exercício 10.4

Considere os dados do exercício 8.2 e use um modelo do tipo $Y = \beta_0 +$

$\beta_1 X + \beta_2 X^2$ para ajustar a resistência à compressão em função da adição de microssílica.

Adição	Resistência (MPa)		
0%	28,1	26,5	24,3
5%	35,3	34,3	37,5
10%	39,8	44,1	42,3
15%	39,1	40,8	43,0

Exercício 10.5

Considere os dados do exercício 8.5 e use o modelo do tipo $Y = \beta_0 + \beta_1 X + \beta_2 X^2$ para ajustar a produtividade mensal em função do intervalo entre manutenções preventivas.

Intervalo	Produtividade				
4	136	137	135	140	136
6	145	146	147	147	148
8	146	144	148	145	145
10	134	131	136	134	133
12	117	119	117	115	116

Exercício 10.6

Os dados a seguir mostram os valores da distribuição normal acumulada para diferentes valores da variável reduzida Z. Ajuste um modelo do tipo $Y = \beta_0 + \Sigma\beta_i X^i$ a esses dados. Após, calcule o valor da variância residual (que no caso deve-se exclusivamente à falta de ajuste) e indique se o ajuste é satisfatório para a maioria das aplicações práticas. Por fim, use o modelo para extrapolações, ou seja, calcule por exemplo F(-4) e F(+4) e indique se o modelo pode ser usado para extrapolações.

Z	-3	-2,5	-2,0	-1,5	-1,0	-0,5	0	0,5	1,0
F(Z)	,0013	,0062	,0228	,0668	,1587	,3085	,5000	,6915	,8413

Z	1,5	2,0	2,5	3,0
F(Z)	,9332	,9772	,9938	,9987

Exercício 10.7

Repita o exercício 10.1 acrescentando um termo $\beta_3 X_1 X_2$ ao modelo. Teste a significância deste termo e conclua se há razões para mantê-lo no modelo.