Métodos Estocáticos da Engenharia II

Capítulo 8 - Análise de Regressão Linear Múltipla

Prof. Magno Silvério Campos

2024/2

Bibliografia

Essas notas de aulas foram baseadas nas seguintes obras:

1. CANCHO, V.G. *Notas de Aulas sobre Noções de Estatística e Probabilidade*. São Paulo: USP, 2010.
2. HINES, W.W.; *et al.* *Probabilidade e Estatística na Engenharia*. 4. ed. Rio de Janeiro: LTC, 2006.
3. MONTGOMERY, D.C.; RUNGER, G.C. *Estatística Aplicada e Probabilidade para Engenheiros*. 6. ed. Rio de Janeiro: LTC, 2016.

Aconselha-se pesquisá-las para se obter um maior aprofundamento e um melhor aproveitamento nos estudos.

Conteúdo Programático

Seção - Modelo de Regressão Linear Múltipla (MRLM)

- Introdução
- Enfoque matricial para o MRLM
- Inferências para o MRLM
- Estudo da adequação do MRLM
- Problemas em um MRLM

Introdução

Em geral, a variável dependente ou resposta Y pode estar relacionada com k variáveis explicativas ou independentes. O modelo

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \varepsilon \tag{1}$$

recebe o nome de modelo de regressão linear múltipla com k variáveis explicativas. Os parâmetros β_j, $j = 0, \ldots, k$ são chamados de coeficientes de regressão. Este modelo descreve um hiperplano no espaço k-dimensional.

Suposições do modelo de regressão linear múltipla (MRLM)

$$\varepsilon \sim NID(0, \sigma^2)$$

Se as suposições do MRLM se verificarem, atendendo à relação na equação (1), a variável aleatória Y segue uma distribuição normal com variância σ^2 e média $\mu_{Y|\boldsymbol{x}}$, sendo

$$E(Y|\boldsymbol{X}=\boldsymbol{x}) = \mu_{Y|\boldsymbol{x}} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k \qquad (2)$$

Observe em (2) que os parâmetros β_j, $j = 1, \ldots, k$, representam a variação esperada na variável resposta Y quando a variável X_j sofre um acréscimo unitário, enquanto todas as demais variáveis explicativas X_i $(i \neq j)$ são mantidas constantes. Por este motivo, os parâmetros β_j, $j = 1, \ldots, k$ são também conhecidos como coeficientes parciais de regressão.

Enfoque Matricial para o MRLM

Suponha que se tenha $n > k+1$ observações de Y e seja X_{ij} a i-ésima observação da variável X_j. As observações são da forma $(x_{i1}, x_{i2}, \ldots, x_{ik}, y_i)$, $i = 1, \ldots, n$. Os dados de uma regressão múltipla podem ser apresentados da seguinte forma:

y	x_1	x_2	$\ldots$	x_k
y_1	x_{11}	x_{12}	$\ldots$	x_{1k}
y_2	x_{21}	x_{22}	$\ldots$	x_{2k}
$\vdots$	$\vdots$	$\vdots$	$\ldots$	$\vdots$
y_n	x_{n1}	x_{n2}	$\ldots$	x_{nk}

Cada observação satisfaz o modelo da equação (1), isto é,

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} + \varepsilon_i, \quad i = 1, \ldots, n. \tag{3}$$

Suponha que existam k variáveis explicativas e n observações. O MRLM é:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} + \varepsilon_i, \quad i = 1, \ldots, n$$

Que é um sistema de n equações que pode ser escrito em notação matricial:

$$\underline{y} = X\underline{\beta} + \underline{\varepsilon} \tag{4}$$

$$\text{onde } \underline{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}_{(n\times 1)}, \; X = \begin{bmatrix} 1 & X_{11} & X_{12} & \ldots & X_{1k} \\ 1 & X_{21} & X_{22} & \ldots & X_{2k} \\ \vdots & \vdots & \vdots & \ldots & \vdots \\ 1 & X_{n1} & X_{n2} & \ldots & X_{nk} \end{bmatrix}_{(n\times p)},$$

$$\underline{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}_{(p\times 1)} \text{ e } \underline{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}_{(n\times 1)}$$

A partir desse enfoque matricial, o estimador de mínimos quadrados de β é dado por

$$\hat{\beta} = (X^t X)^{-1} X^t y \tag{5}$$

Sendo assim, o modelo de regressão múltipla ajustado que era escrito como

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \hat{\beta}_2 X_{i2} + \cdots + \hat{\beta}_k X_{ik}, \quad i = 1, \ldots, n, \tag{6}$$

Na forma matricial, o modelo ajustado passa a ser escrito como

$$\hat{\boldsymbol{y}} = \mathrm{X}\hat{\boldsymbol{\beta}} = \mathrm{X}(\mathrm{X}^t\mathrm{X})^{-1}\mathrm{X}^t\boldsymbol{y} = H\boldsymbol{y} \tag{7}$$

onde $H = \mathrm{X}(\mathrm{X}^t\mathrm{X})^{-1}\mathrm{X}^t$ é chamada de matriz "hat" (chapéu).

A diferença entre a observação y_i e o valor ajustado $\hat{y}_i$ é o resíduo $e_i = y_i - \hat{y}_i$. O vetor de resíduos, de ordem $(n \times 1)$, é então representado por

$$\boldsymbol{e} = \boldsymbol{y} - \hat{\boldsymbol{y}} = (I_n - H)\boldsymbol{y} \tag{8}$$

Exemplo - [Cancho(2010)]

O proprietário de uma casa está interessado no efeito, na conta de luz, de seu aparelho de ar condicionado e de sua secadora de roupas.

Para isso, ele registrou o número de horas que usou o seu aparelho de ar condicionado e o número de vezes que a secadora de roupa foi usada em cada dia, durante 21 dias.

Também monitorou o "consumo" de eletricidade durante esses dias. Os dados são apresentados na tabela que se segue:

Quantidade de eletricidade (Y)	Horas de uso do condicionador de ar (X_1)	Nº de vezes que a secadora foi ligada (X_2)
35	1,5	1
63	4,5	2
66	5,0	2
17	2,0	0
94	8,5	3
79	6,0	3
93	13,5	1
66	8,0	1
94	12,5	1
82	7,5	2
78	6,5	5
65	8,0	1
77	7,5	2
75	8,0	2
62	7,5	1
85	12,0	1
43	6,0	0
57	2,5	3
33	5,0	0
65	7,5	1
33	6,0	0

Na tabela acima, aparecem 21 observações. O ajuste do modelo de regressão múltipla é dado por:

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

Agora utilizaremos o enfoque matricial para ajustar o modelo de regressão anterior a esse conjunto de dados. A matriz $\boldsymbol{X}$ e o vetor $\boldsymbol{y}$ para este modelo são:

$$\boldsymbol{X} = \begin{bmatrix} 1 & 1,5 & 1 \\ 1 & 4,5 & 2 \\ 1 & 5,0 & 2 \\ \vdots & \vdots & \vdots \\ 1 & 7,5 & 1 \\ 1 & 6,0 & 0 \end{bmatrix}, \quad \boldsymbol{y} = \begin{bmatrix} 35 \\ 63 \\ 66 \\ \vdots \\ 65 \\ 33 \end{bmatrix}$$

A matriz $\boldsymbol{X^tX}$ é:

$$\boldsymbol{X^tX} = \begin{bmatrix} 1 & 1 & \dots 1 \\ 1,5 & 4,5 & \dots 6,0 \\ 1 & 2 & \dots 0 \end{bmatrix} \begin{bmatrix} 1 & 1,5 & 1 \\ 1 & 4,5 & 2 \\ \vdots & \vdots & \vdots \\ 1 & 6,0 & 0 \end{bmatrix} = \begin{bmatrix} 21 & 145,5 & 32 \\ 145,5 & 1204,75 & 219 \\ 32 & 219 & 80 \end{bmatrix}$$

e o vetor $\boldsymbol{X^ty}$ é

$$\boldsymbol{X^ty} = \begin{bmatrix} 1 & 1 & \dots 1 \\ 1,5 & 4,5 & \dots 6,0 \\ 1 & 2 & \dots 0 \end{bmatrix} \begin{bmatrix} 35 \\ 63 \\ \vdots \\ 33 \end{bmatrix} = \begin{bmatrix} 1362 \\ 10487 \\ 2371 \end{bmatrix}$$

Os estimadores de mínimos quadrados obtém-se a partir da equação (6), isto é

$$\hat{\boldsymbol{\beta}}^t = (\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t\boldsymbol{y}$$

Ou seja

$$\begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = \begin{bmatrix} 21 & 145,5 & 32 \\ 145,5 & 1204,75 & 219 \\ 32 & 219 & 80 \end{bmatrix}^{-1} \begin{bmatrix} 1362 \\ 10487 \\ 2371 \end{bmatrix}$$

$$= \begin{bmatrix} 0,3757980 & -0,0359507 & -0,0519043 \\ -0,0359507 & 0,0050915 & 0,0004424 \\ -0,0519043 & 0,0004424 & 0,0320506 \end{bmatrix} \begin{bmatrix} 1362 \\ 10487 \\ 2371 \end{bmatrix}$$

$$= \begin{bmatrix} 11,7572 \\ 5,4783 \\ 9,9379 \end{bmatrix}$$

Portanto, o modelo de regressão ajustado é

$$\hat{y} = 11,7572 + 5,4783X_1 + 9,9379X_2$$

Abaixo, estão os 21 valores ajustados $\hat{y}_i$, bem como os respectivos resíduais.

Observação	y_i	x_1	x_2	$\hat{y}_i$	$e_i = y_i - \hat{y}_i$
1	35	1,5	1	29,9126	5,0875
2	63	4,5	2	56,2854	6,7147
3	66	5,0	2	59,0245	6,9755
4	17	2,0	0	22,7138	-5,7138
5	94	8,5	3	88,1365	5,8636
6	79	6,0	3	74,4407	4,5593
7	93	13,5	1	95,6522	-2,6521
8	66	8,0	1	65,5215	0,4785
9	94	12,5	1	90,1739	3,8262
10	82	7,5	2	72,7203	9,2798
11	78	6,5	3	97,0557	-19,0557
12	65	8,0	1	65,5215	-,05215
13	77	7,5	2	72,7203	4,2798
14	75	8,0	2	75,4594	-0,4594
15	62	7,5	1	62,7824	-0,7823
16	85	12,0	1	87,4347	-2,4347
17	43	6,0	0	44,6270	-1,6270
18	57	2,5	3	55,2667	1,7373
19	33	5,0	0	39,1487	-6,1487
20	65	7,5	1	62,7824	-7,9008
21	33	6,0	0	44,6270	-11,6270

Estimação de σ^2

Como no caso do modelo de regressão linear simples, a estimação de σ^2 está definida em termos da soma de quadrados dos residuais (SQR).

$$\begin{aligned} SQR &= \sum_{i=1}^{n} e_i^2 = \boldsymbol{e}^t\boldsymbol{e} = (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^t(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) \\ &= \boldsymbol{y}^t\boldsymbol{y} - 2\hat{\boldsymbol{\beta}}^t\boldsymbol{X}^t\boldsymbol{y} + \hat{\boldsymbol{\beta}}^t\boldsymbol{X}^t\boldsymbol{X}\hat{\boldsymbol{\beta}} \end{aligned}$$

Já que $\boldsymbol{X}^t\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}^t\boldsymbol{y}$, esta última equação se reduz a:

$$SQR = \boldsymbol{y}^t\boldsymbol{y} - \hat{\boldsymbol{\beta}}^t\boldsymbol{X}^t\boldsymbol{y} \tag{9}$$

Logo,

$$\hat{\sigma^2} = \frac{SQR}{n-p} = \frac{\boldsymbol{y}^t\boldsymbol{y} - \hat{\boldsymbol{\beta}}^t\boldsymbol{X}^t\boldsymbol{y}}{n-p} \tag{10}$$

Observação: $p = k + 1$

Exemplo

Estime a variância do erro, σ^2, para o problema dessa seção.

$$\boldsymbol{y}^t\boldsymbol{y} = \sum_{i=1}^{21} y_i^2 = 97914$$

$$\hat{\boldsymbol{\beta}}^t \boldsymbol{X}^t\boldsymbol{y} = (11,7572 \quad 5,4783 \quad 9,9379) \begin{bmatrix} 1362 \\ 10487 \\ 2371 \end{bmatrix} = 97026,5216$$

Portanto, a soma de quadrados do residual é obtida com equação (9):

$$\begin{aligned} SQR &= \boldsymbol{y}^t\boldsymbol{y} - \hat{\boldsymbol{\beta}}^t \boldsymbol{X}^t\boldsymbol{y} \\ &= 97914 - 97026,5216 = 887,4784 \end{aligned}$$

Logo,

$$\hat{\sigma^2} = \frac{SQR}{n-p} = \frac{887,4787}{21-3} = 49,30436.$$

Inferência no modelo de regressão linear múltipla

Para fazermos inferência, isto é, construir intervalos de confiança e realizar testes de hipóteses no MRLM, é necessário que as suposições do MRLM sejam válidas:

$$\varepsilon \sim NID(0, \sigma^2)$$

Intervalos de confiança para β_j, $j = 0, 1, \ldots, k$

A variável aleatória,

$$T = \frac{\hat{\beta}_j - \beta_j}{\sqrt{C_{jj}\hat{\sigma}^2}}, \quad j = 0, 1, \ldots, k$$

tem distribuição t-Student com $(n - p)$ graus de liberdade.
A partir disso, pode-se mostrar que um intervalo de $100(1 - \alpha)\%$ de confiança para β_j, $j = 0, 1, \ldots, k$, é dado por:

$$IC(\beta_j; 1 - \alpha) = \left(\hat{\beta}_j - t_{(\frac{\alpha}{2}, n-p)} \sqrt{C_{jj}\hat{\sigma}^2}; \hat{\beta}_j + t_{(\frac{\alpha}{2}, n-p)} \sqrt{C_{jj}\hat{\sigma}^2} \right) \quad (11)$$

onde C_{jj} é o j-ésimo elemento diagonal da matriz $(\boldsymbol{X}^t\boldsymbol{X})^{-1}$, com $j = 0, 1, \ldots, k$.

Exemplo

Suponha que no exemplo desta seção, temos interesse em estimar, com 95% de confiança a variação sofrida na quantidade de energia transformada, quando o ar condicionado sofre um acréscimo de uma hora de uso, sendo mantido constante o uso da secadora.

$$IC(\beta_1, 0, 95) = \left(\hat{\beta}_1 - t_{0,025,18}\sqrt{C_{11}\hat{\sigma}^2}; \hat{\beta}_1 + t_{0,025,18}\sqrt{C_{11}\hat{\sigma}^2}\right)$$

Dos dados, temos:
$\hat{\beta}_1 = 5,4783$,

$$(\boldsymbol{X}^t\boldsymbol{X})^{-1} = \begin{bmatrix} 0,3757980 & -0,0359507 & -0,0519043 \\ -0,0359507 & 0,0050915 & 0,0004424 \\ -0,0519043 & 0,0004424 & 0,0320506 \end{bmatrix},$$

$C_{11} = 0,0050915$, $\hat{\sigma}^2 = 49,30436$ e $t_{0,025,18} = 2,101$.
Logo,

$$\begin{aligned} IC(\beta_1, 0, 95) &= \left(5,4783 \pm 2,101\sqrt{(0,0050915)(49,30436)}\right) \\ &= (4,4256; 6,5309) \end{aligned}$$

Intervalos de confiança para previsão de novas observações

Um modelo de regressão pode ser utilizado para prever observações futuras da variável resposta Y, correspondentes a valores particulares das variáveis independentes, por exemplo, $x_{01}, x_{02}, \ldots, x_{0k}$. Se $\boldsymbol{x_0}^t = [1, x_{01}, x_{02}, \ldots, x_{0k}]$, então uma estimação pontual da observação futura Y_0 no ponto $x_{01}, x_{02}, \ldots, x_{0k}$ é

$$\hat{Y_0} = \boldsymbol{x_0}^t \hat{\boldsymbol{\beta}}.$$

Um intervalo de $100(1-\alpha)\%$ de confiança para esta observação futura é

$$IC(Y_0; 1-\alpha) = \left[\hat{Y_0} \pm t_{\frac{\alpha}{2}, n-p} \sqrt{\hat{\sigma}^2 (1 + \boldsymbol{x_0}^t (\boldsymbol{X}^t \boldsymbol{X})^{-1} \boldsymbol{x_0})}\right] \qquad (12)$$

Exemplo

Suponha que temos interesse em construir um intervalo de 95% de confiança, para o consumo de eletricidade num dia, quando o ar condicionado for ligado durante 8 horas e a secadora de roupa for ligada uma vez nesse dia.

$$IC(Y_0; 1-\alpha) = \left[\hat{Y_0} \pm t_{\frac{\alpha}{2}, n-p}\sqrt{\hat{\sigma}^2(1 + \boldsymbol{x_0}^t(\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{x_0})}\right] \tag{13}$$

Note que $\boldsymbol{x_0}^t = [1\ 8\ 1]$. A estimação pontual de Y é

$$\hat{Y_0} = \boldsymbol{x_0}^t\hat{\boldsymbol{\beta}} = 65,52.$$

Temos que calcular

$$\hat{\sigma}^2(1 + \boldsymbol{x_0}^t(\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{x_0}) = \hat{\sigma}^2 + \hat{\sigma}^2\boldsymbol{x_0}^t(\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{x_0}$$

Cálculos auxiliares:

$$\hat{\sigma}^2 \boldsymbol{x_0}^t (\boldsymbol{X}^t \boldsymbol{X})^{-1} \boldsymbol{x_0} =$$

$$= 49,30436 [1 \;\; 8 \;\; 1] \begin{bmatrix} 0,3757980 & -0,0359507 & -0,0519043 \\ -0,0359507 & 0,0050915 & 0,0004424 \\ -0,0519043 & 0,0004424 & 0,0320506 \end{bmatrix} \begin{bmatrix} 1 \\ 8 \\ 1 \end{bmatrix} =$$

$$= 3,0451$$

Logo,

$$\hat{\sigma}^2 (1 + \boldsymbol{x_0}^t (\boldsymbol{X}^t \boldsymbol{X})^{-1} \boldsymbol{x_0}) = \hat{\sigma}^2 + \hat{\sigma}^2 \boldsymbol{x_0}^t (\boldsymbol{X}^t \boldsymbol{X})^{-1} \boldsymbol{x_0} =$$

$$= 49,30436 + 3,0451 = 52,34946$$

Portanto, substituindo na equação (13), tem-se que:

$$\begin{aligned} IC(Y_0, 0, 95) &= \left(65,52 - 2,101\sqrt{52,34946}; 65,52 + 2,101\sqrt{52,34946}\right) \\ &= (50,31865; 80,72135) \end{aligned}$$

Teste de significância da regressão

Com o objetivo de determinar se existe um relacionamento linear entre a variável resposta Y e o conjunto de variáveis explicativas, $X_1, X_2, \ldots, X_k$, pode ser utilizado o teste de significância da regressão. Nesse teste, as hipóteses apropriadas são:

$$\begin{aligned} H_0 &: \beta_1 = \beta_2 = \cdots = \beta_k = 0 \\ H_1 &: \beta_j \neq 0, \text{ para pelo menos um } j. \end{aligned} \tag{14}$$

A rejeição de H_0 significa que pelo menos uma das variáveis independentes ou regressoras $X_1, X_2, \ldots, X_n$ tem contribuição significativa no modelo.

Estatística de Teste

$$F_{obs} = \frac{SQreg/k}{SQR/n-p} = \frac{QMreg}{QMR} \tag{15}$$

que tem distribuição F com k e $n-p$ graus de liberdade no numerador e no denominador, respectivamente.
Rejeita-se H_o se $F_{obs} > F_{\alpha,k,n-p}$.
O procedimento é resumido em uma tabela de análise de variância, tal como a tabela abaixo:

Tabela: Análise de variância para o teste de $H_0 : \beta_1 = \beta_2 = \ldots, = \beta_k = 0$

Fonte de variação	Soma de Quadrados	Graus de Liberdade	Quadrado Médio	F
Regressão	$SQreg$	k	$QMreg$	$QMreg/QMR$
Residual	SQR	$n-p$	QMR	
Total	SQT	$n-1$		

Da equação (9), pode-se calcular SQR, isto é,

$$SQR = \boldsymbol{y}^t\boldsymbol{y} - \hat{\boldsymbol{\beta}}^t\boldsymbol{X}^t\boldsymbol{y}$$

Agora, pode-se determinar uma fórmula simples para o cálculo da SQR :

Já que $SQT = \boldsymbol{y}^t\boldsymbol{y} - \dfrac{\left(\sum_{i=1}^{n} Y_i\right)^2}{n}$, então a equação anterior fica assim:

$$SQR = \boldsymbol{y}^t\boldsymbol{y} - \frac{\left(\sum_{i=1}^{n} Y_i\right)^2}{n} - \left[\hat{\boldsymbol{\beta}}^t\boldsymbol{X}^t\boldsymbol{y} - \frac{\left(\sum_{i=1}^{n} Y_i\right)^2}{n}\right]$$

$$SQR = SQT - SQreg$$

Portanto, $SQreg = \hat{\boldsymbol{\beta}}^t\boldsymbol{X}^t\boldsymbol{y} - \dfrac{\left(\sum_{i=1}^{n} Y_i\right)^2}{n}$.

Exemplo

Para avaliar se de fato existe relação linear entre a variável quantidade de eletricidade transformada e as variáveis número de horas de uso do ar condicionado e o número de vezes que secadora foi usada, decidiu-se testar se os coeficientes de regressão β_1 e β_2 do MRLM pederiam ser ambos iguais a zero. Isto é, $H_0 : \beta_1 = \beta_2 = 0$; $H_1 : \beta_1 \neq 0$, ou $\beta_2 \neq 0$.

A soma de quadrados total é

$$\begin{aligned} SQT &= \boldsymbol{y}^t\boldsymbol{y} - \frac{\left(\sum_{i=1}^{n} Y_i\right)^2}{n} \\ &= 97914 - \frac{(1362)^2}{21} = 9578,57. \end{aligned}$$

A soma de quadrados da regressão é dada por

$$\begin{aligned} SQreg &= \hat{\boldsymbol{\beta}}^t \boldsymbol{X}^t \boldsymbol{y} - \frac{\left(\sum_{i=1}^{n} Y_i\right)^2}{n} \\ &= 97026,5216 - \frac{(1362)^2}{21} = 8691,05. \end{aligned}$$

E por diferença

$$SQR = SQT - SQreg = 9578,57 - 8691,05 = 887,52$$

Na tabela a seguir, encontra-se a respectiva análise de variância.

Tabela: Análise de variância para o teste de $H_0 : \beta_1 = \beta_2 = 0$

Fonte de variação	Soma de Quadrados	Graus de Liberdade	Quadrado Médio	F
Regressão	$8691,05$	2	$4345,5$	$88,14$
Residual	$887,52$	18	$49,3$	
Total	$9578,57$	20		

Para testar $H_0 : \beta_1 = \beta_2 = 0$, calcula-se a estatística de teste

$$F_{obs} = \frac{QMreg}{QMR} = \frac{4345,5}{49,3} = 88,14$$

Já que $F_{obs} > F_{0,05,2,18} = 3,55$, rejeita-se H_0. Concluí-se ao nível de significância de 5%, que a variável quantidade de eletricidade transformada se relaciona linearmente com as variáveis número de horas de uso do ar condicionado e número de vezes em que a secadora foi ligada.

Teste para avaliar se um único $H_0 : \beta_j = 0$

Suponha que temos interesse em determinar a importância da variável explicativa x_j no modelo de regressão adotado. Neste caso, as hipóteses a serem testadas são:

$$H_0 : \beta_j = 0$$
$$H_1 : \beta_j \neq 0.$$

Estatística de Teste

$$T = \frac{\hat{\beta}_j}{\sqrt{\hat{\sigma}^2 C_{jj}}} \tag{16}$$

a qual tem distribuição t-Student com $(n - p)$ graus de liberdade se a hipótese nula é verdadeira. Onde, C_{jj} é o j-ésimo elemento da diagonal principal da matriz $(\boldsymbol{X}^t \boldsymbol{X})^{-1}$.

Critério de rejeição

Rejeita-se $H_0 : \beta_j = 0$ se $|T_{obs}| > t_{\alpha/2,n-p}$.

Se $H_0 : \beta_j = 0$ não for rejeitada, este resultado indica que a variável X_j poderá ser excluída do modelo.

Exemplo

Considere os dados do exemplo desta seção e suponha que se deseja testar a hipótese de que o coeficiente de regressão para X_2 é zero. As hipóteses são:

$$H_0 : \beta_2 = 0$$
$$H_1 : \beta_2 \neq 0.$$

O elemento da diagonal principal da matriz $(\boldsymbol{X}^t\boldsymbol{X})^{-1}$ que corresponde a $\hat{\beta}_2$ é $C_{22} = 0,0320506$, de modo que a estatística de teste T da equação (16) é:

$$T_{obs} = \frac{\hat{\beta}_2}{\sqrt{\hat{\sigma}^2 C_{22}}} = \frac{9,9379}{\sqrt{49,30436 \times 0,0320506}} = 7,91$$

Já que $|T_{obs}| > t_{0,025,18} = 2,101$, rejeitamos $H_0 : \beta_2 = 0$. Portanto podemos concluir, ao nível de significância de 5%, que a inclusão da variável número de vezes que a secadora de roupa foi usada (X_2) contribuiu para

Estudo da Adequação de um MRLM

O ajuste de um modelo de regressão requer várias suposições:

- A estimação dos parâmetros do modelo requer a suposição de que os erros sejam variáveis aleatórias não correlacionadas com média zero e variância constante;
- A construção de intervalos de confiança e testes de hipóteses requer que os erros sejam normalmente distribuídos;
- Além disso, é assumindo que a ordem do modelo é correta; isto é, se ajustamos um modelo de regressão linear simples, considera-se que o fenômeno realmente se comporta dessa forma.

O pesquisador deve sempre questionar a validade dessas suposições e realizar análises para verificar a adequação do modelo adotado. Nesta subseção serão discutidos métodos úteis para o estudo da adequação do modelo de regressão.

(1) - Análise Residual

A análise de resíduos é útil para verificar a suposição de que os erros são não correlacionados e têm uma distribuição que é aproximadamente normal com média zero e variância constante, assim como para determinar se é necessária a adição de termos adicionais ao modelo.

Os resíduos de um modelo de regressão são definidos como

$$e_i = y_i - \hat{y}_i, \quad i = 1, \ldots, n \qquad \rightarrow \text{resíduo regular}.$$

onde y_i é uma observação real de Y e $\hat{y}_i$ é o valor correspondente estimado através do modelo de regressão.

Um procedimento muito útil consiste em padronizar os resíduos assim:

1. $$d_i = \frac{e_i}{\sqrt{\hat{\sigma^2}}}, \quad i = 1, \ldots, n \qquad \rightarrow \text{resíduo padronizado.}$$

2. $$z_i = \frac{e_i}{\sqrt{\hat{\sigma^2}(1 - h_{ii})}}, \quad i = 1, \ldots, n \qquad \rightarrow \text{resíduo padronizado.}$$

3. $$z_i^* = \frac{e_i}{\sqrt{\hat{\sigma^2_{(i)}}(1 - h_{ii})}}, \quad i = 1, \ldots, n \qquad \rightarrow \text{resíduo estudentizado.}$$

Onde:

- h_{ii} é o i-ésimo elemento da diagonal da matriz $H = \mathrm{X}(\mathrm{X}^t\mathrm{X})^{-1}\mathrm{X}^t$.
- $\sigma^2_{(i)}$ é a variância estimada sem utilizarmos a i-ésima observação.

Como foi proposto no capítulo de Regressão Linear Simples, vários gráficos de resíduos são frequentemente úteis. Em MRLM, a interpretação é a mesma do caso de MRLS.

É útil também plotar os resíduos contra as variáveis que não estejam presentes no modelo, mas que sejam possíveis candidatas à inclusão no modelo.

Padrões de comportamento nesses gráficos indicam que o modelo pode ser melhorado através da adição das variáveis candidatas.

(2) - Coeficiente de determinação múltipla

Uma medida largamente usada para um modelo de regressão é a razão de soma dos quadrados.
A quantidade:

$$R^2 = \frac{SQreg}{SQT} = 1 - \frac{SQR}{SQT} \tag{17}$$

recebe o nome de **coeficiente de determinação múltipla** que é usado para ajudar a julgar a adequação do modelo de regressão. Temos que $0 \leq R^2 \leq 1$.

O coeficiente de determinação pode ser interpretado como a proporção da variabilidade presente nas observações da variável resposta Y, que é explicada pelas variáveis independentes $X_1, X_2, \ldots, X_k$ no modelo de regressão.

A raiz quadrada positiva de R^2 é o coeficiente de correlação múltipla entre Y e o conjunto de variáveis regressoras $X_1, X_2, \ldots, X_k$.

Exemplo

Para os dados do exemplo desta seção, determinar R^2.
Da equação (17) tem-se:

$$R^2 = \frac{SQreg}{SQT} = \frac{8,691,1}{9578,57} = 0,907$$

Esse resultado significa que o modelo ajustado explicou 90,7% da variação na variável resposta Y ("consumo" de energia). Isto é, 90,7% da variabilidade de Y é explicada quando são usadas as duas variáveis regressoras, *número de horas de uso do condicionador de ar* (x_1) e *número de vezes de uso da secadora de roupas* (x_2).

Nota:

Desenvolveu-se o modelo que relaciona Y a X_1, apenas. O valor de R^2 para esse modelo é $R^2 = 0,586$. Assim, o acréscimo da variável X_2 ao modelo aumentou R^2 de 0,586 para 0,971.

Observação

A estatística R^2 é de algum modo problemática como uma medida da qualidade do ajuste para um modelo de regressão múltipla, uma vez que ela sempre aumenta quando uma variável é adicionada a um modelo.

Uma vez que R^2 sempre aumenta quando um regressor é adicionado, pode ser difícil julgar se o aumento está nos dizendo qualquer coisa útil acerca do novo regressor. É particularmente difícil interpretar um pequeno aumento.

Por esse fato, muitos usuários de regressão preferem usar o coeficiente de determinação múltipla ajustado, R^2_{aj} ajustado, definido como

$$R^2_{aj} = 1 - \frac{\frac{SQR}{n-p}}{\frac{SQT}{n-1}}. \tag{18}$$

R^2_{aj} penalizará a adição de termos ao modelo que não sejam significantes na modelagem da resposta. A interpretação de R^2_{aj} é idêntica à de R^2.

Exemplo

Para os dados do exemplo desta seção, determinar R^2_{aj}.
Da equação (18) tem-se:

$$R^2_{aj} = 1 - \frac{\frac{SQR}{n-p}}{\frac{SQT}{n-1}} = 1 - \frac{\frac{887{,}05}{21-3}}{\frac{9578{,}57}{21-1}} = 0{,}897$$

Esse resultado significa que o modelo ajustado explicou 89,7% da variação na variável resposta Y ("consumo de energia"). Isto é, 89,7% da variabilidade de Y é explicada quando são usadas as duas variáveis regressoras, *número de horas de uso do condicionador de ar* (x_1) e *número de vezes de uso da secadora de roupas* (x_2).

Observação

Quando R^2 e R^2_{aj} são muito diferentes, existem variáveis X no modelo que não estão contibuindo em nada para o MRLM.

Problemas em um MRLM

1 - Observações influentes

Talvez, algumas observações podem se distanciar muito da massa de dados e exercer forte influência na estimação dos parâmetros de um MRLM. Existem vários métodos de detecção de observações influentes. Um excelente diagnóstico é uma medida desenvolvida por *Dennis R. Cook*, conhecida como Distância de Cook e é assim definida:

$$D_i = \frac{e_i^2}{p \cdot \hat{\sigma^2}} \cdot \frac{h_{ii}}{(1-h_{ii})^2} \qquad i = 1, 2, \ldots, n. \tag{19}$$

onde h_{ii} é o i-ésimo elemento da diagonal da matriz $H = \mathrm{X}(\mathrm{X}^t\mathrm{X})^{-1}\mathrm{X}^t$.

Se $D_i > 1$, a i-ésima observação é influente.

Exemplo

A tabela abaixo apresenta os valores calculados para as Distâncias de Cook para cada observação do exemplo desta seção.

Observações i	h_{ii}	Distância de Cook D_i
1	0,209	0,058
2	0,084	0,030
3	0,073	0,028
4	0,252	0,100
5	0,132	0,041
6	0,121	0,022
7	0,273	0,025
8	0,062	0,000
9	0,212	0,034
10	0,057	0,037
11	0,435	3,336 *
12	0,062	0,000
13	0,057	0,008
14	0,061	0,000
15	0,058	0,000
16	0,185	0,011
17	0,128	0,003
18	0,212	0,007
19	0,144	0,050
20	0,058	0,002
21	0,128	0,153

2 - Multicolinearidade

Em MRLM, o problema de multicolinearidade surge quando as variáveis regressoras apresentam uma forte relação linear entre si. Algumas indicações de multicolinearidade são:

(1) Ocorrência de valores próximos de +1 ou -1 para os coeficientes de correlação linear r_{ij} entre pares de variáveis regressoras (X_i, X_j). O coeficiente de correlação r_{ij} é definido por:

$$r_{ij} = \frac{S_{ij}}{\sqrt{S_{ii} \cdot S_{jj}}}, \qquad i, j = 1, 2, \ldots, k \tag{20}$$

onde:

$$S_{ij} = \sum_{l=1}^{n} x_{li} x_{lj} - n\bar{x}_i \bar{x}_j$$

$$S_{ii} = \sum_{l=1}^{n} x_{li}^2 - n\bar{x}_i{}^2 \quad \text{e} \quad S_{jj} = \sum_{l=1}^{n} x_{lj}^2 - n\bar{x}_j{}^2$$

(2) A hipótese $H_0: \quad \beta_1 = \beta_2 = \ldots = \beta_k = 0$ é rejeitada por meio do teste F, mas nenhuma hipótese do tipo $H_0: \quad \beta_j = 0, \quad j = 1, 2, \ldots, k$, é rejeitada por meio da realização do teste t-Student sobre os coeficientes individuais.

(3) Obtenção de estimativas para os coeficientes de regressão com sinais algébricos opostos àqueles que seriam esperados a partir de conhecimentos teóricos disponíveis.

(4) Obtenção da estatística FIV (Fator Inflacionário da Variância) maior do que 5 para cada variável explicativa. O FIV é definido como:

$$FIV_j = \text{j-ésimo elemento diagonal de } R^{-1}, \qquad j = 1, 2, \ldots, k, \qquad (21)$$

$$\boldsymbol{R} = \begin{bmatrix} 1 & r_{12} & r_{13} & \ldots & r_{1k} \\ r_{21} & 1 & r_{23} & \ldots & r_{2k} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ r_{k1} & r_{k2} & r_{k3} & \ldots & 1 \end{bmatrix}$$

Exemplo

Para o nosso exemplo, determine se há problema de multicolinearidade. A matriz $\mathbf{R}$ é dada por:

$$\boldsymbol{R} = \begin{bmatrix} 1 & -0,03463 \\ -0,03463 & 1 \end{bmatrix}$$

Logo, R^{-1} é calculada:

$$\boldsymbol{R^{-1}} = \begin{bmatrix} 1,00120 & 0,03467 \\ 0,03467 & 1,00120 \end{bmatrix}$$

Assim,

$$FIV_1 = 1,00120 < 5$$

$$FIV_2 = 1,00120 < 5$$

Portanto, concluí-se que não existe problema de multicolinearidade no modelo construído.

Várias medidas corretivas têm sido propostas para tratar o problema de multicolinearidade, entre elas destacam-se:

- Eliminar do modelo as variáveis regressoras que estejam altamente correlacionadas a outras variáveis regressoras;
  Este procedimento apresenta a desvantagem de descartar a informação contida nas variáveis que serão eliminadas.
- Aumentar os dados com novas observações especialmente planejadas para romper essa dependência.

Seleção de variáveis e contrução de modelos

Introdução

Um problema importante em muitas aplicações da análise de regressão envolve selecionar o conjunto de variáveis regressoras a ser usado no modelo.

Algumas vezes, experiência prévia ou considerações teóricas em foco podem ajudar o analista a especificar o conjunto de variáveis regressoras a se usar em uma situação particular.

Mas estamos certos de que nem todos os regressores candidatos são necessários para modelar adequadamente a resposta Y. Em tal situação, estamos interessados na seleção de variáveis; ou seja, filtrar as variáveis candidatas para obter um modelo de regressão que contenha o melhor subconjunto de variáveis regressoras.

Todas as regressões possíveis

Essa aboradgem requer que o analista ajuste todas as equações de regressão envolvendo uma variável candidata, todas as equações de regressão envolvendo duas variáveis candidatas e assim por diante.

Então essas equações são avaliadas de acordo com alguns critérios adequados para selecionar o melhor modelo de regressão.

Se houver k regressores candidatos, haverá 2^k equaçòes de regressão.

Procure por uma opção tal como regressão de Melhores Subconjuntos (*Best Subsets*).

Critérios de seleção

Vários critérios podem ser usados para avaliar e comparar os diferentes modelos obtidos de regressão. Sejam alguns deles:

R^2_{aj}

Geralmente, o modelo que maximiza R^2_{aj} é considerado como um bom candidato para a melhor equação de regressão.

Estatística C_p de Mallows

$$C_p = \frac{SQR \cdot (p)}{\hat{\sigma^2}} - n + 2p$$

Escolhemos como a melhor equação de regressão um modelo com $C_p \cong p$.

Exemplo - [Montgomery e Runger(2016)]

A tabela a seguir apresenta dados sobre o teste de sabor de 38 marcas de vinho. A variável resposta é Y =qualidade do vinho e desejamos encontrar a melhor equação de regressão que relaciona qualidade aos outros cinco parâmetros.

Table 12-17 Wine Quality Data

	x_1 Clarity	x_2 Aroma	x_3 Body	x_4 Flavor	x_5 Oakiness	y Quality
1	1.0	3.3	2.8	3.1	4.1	9.8
2	1.0	4.4	4.9	3.5	3.9	12.6
3	1.0	3.9	5.3	4.8	4.7	11.9
4	1.0	3.9	2.6	3.1	3.6	11.1
5	1.0	5.6	5.1	5.5	5.1	13.3
6	1.0	4.6	4.7	5.0	4.1	12.8
7	1.0	4.8	4.8	4.8	3.3	12.8
8	1.0	5.3	4.5	4.3	5.2	12.0
9	1.0	4.3	4.3	3.9	2.9	13.6
10	1.0	4.3	3.9	4.7	3.9	13.9
11	1.0	5.1	4.3	4.5	3.6	14.4
12	0.5	3.3	5.4	4.3	3.6	12.3
13	0.8	5.9	5.7	7.0	4.1	16.1
14	0.7	7.7	6.6	6.7	3.7	16.1
15	1.0	7.1	4.4	5.8	4.1	15.5
16	0.9	5.5	5.6	5.6	4.4	15.5
17	1.0	6.3	5.4	4.8	4.6	13.8
18	1.0	5.0	5.5	5.5	4.1	13.8
19	1.0	4.6	4.1	4.3	3.1	11.3
20	0.9	3.4	5.0	3.4	3.4	7.9
21	0.9	6.4	5.4	6.6	4.8	15.1
22	1.0	5.5	5.3	5.3	3.8	13.5
23	0.7	4.7	4.1	5.0	3.7	10.8
24	0.7	4.1	4.0	4.1	4.0	9.5
25	1.0	6.0	5.4	5.7	4.7	12.7
26	1.0	4.3	4.6	4.7	4.9	11.6
27	1.0	3.9	4.0	5.1	5.1	11.7
28	1.0	5.1	4.9	5.0	5.1	11.9
29	1.0	3.9	4.4	5.0	4.4	10.8
30	1.0	4.5	3.7	2.9	3.9	8.5
31	1.0	5.2	4.3	5.0	6.0	10.7
32	0.8	4.2	3.8	3.0	4.7	9.1
33	1.0	3.3	3.5	4.3	4.5	12.1
34	1.0	6.8	5.0	6.0	5.2	14.9
35	0.8	5.0	5.7	5.5	4.8	13.5
36	0.8	3.5	4.7	4.2	3.3	12.2
37	0.8	4.3	5.5	3.5	5.8	10.3
38	0.8	5.2	4.8	5.7	3.5	13.2

Fonte: [Montgomery e Runger(2016)]

A figura a seguir é a matriz de gráficos de dispersão para os dados da qualidade de vinho, como construída pelo Minitab.

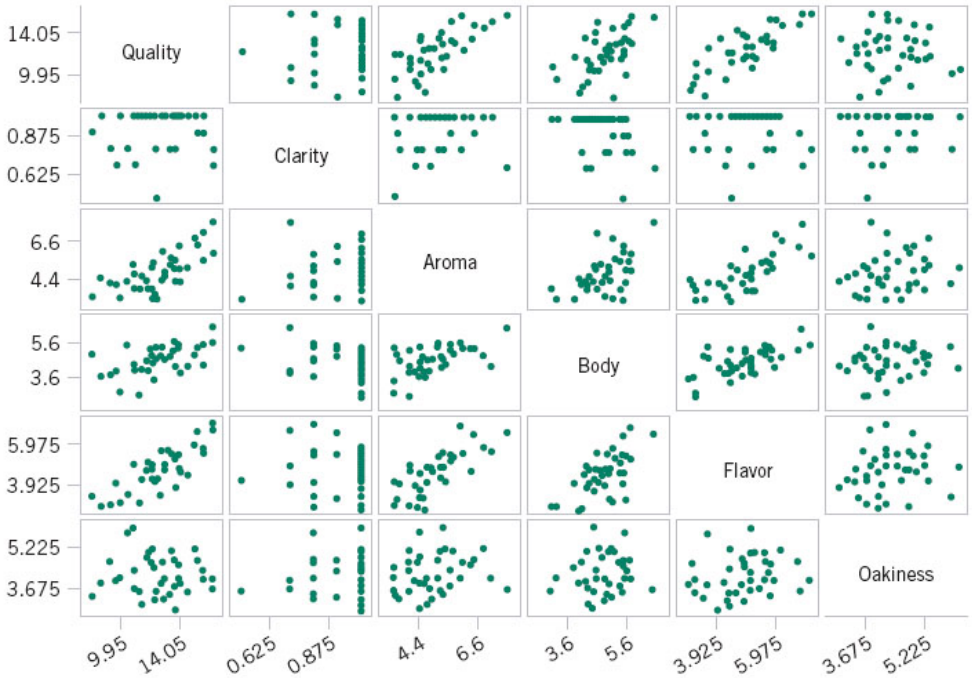

Figure 12-12 A Matrix of Scatter Plots from Minitab for the Wine Quality Data.

Fonte: [Montgomery e Runger(2016)]

Notamos que há algumas indicações de possíveis relações lineares entre qualidade e os regressores, porém não há uma impressão visual óbvia de quais regressores seriam apropriados.

Abaixo, estão listadas as 3 melhores (opção nossa) equações de regressão para cada tamanho de subconjuntos.

Table 12-18 Minitab All Possible Regressions Output for the Wine Quality Data

Best Subsets Regression: Quality versus Clarity, Aroma, . . .

Response is Quality

Vars	R-Sq	R-Sq (adj)	C–p	S	Clarity	Aroma	Flavor	Oakiness
1	62.4	61.4	9.0	1.2712			X	
1	50.0	48.6	23.2	1.4658		X		
1	30.1	28.2	46.0	1.7335			X	
2	66.1	64.2	6.8	1.2242			X	X
2	65.9	63.9	7.1	1.2288		X	X	
2	63.3	61.2	10.0	1.2733	X		X	
3	70.4	67.8	3.9	1.1613		X	X	X
3	68.0	65.2	6.6	1.2068	X		X	X
3	66.5	63.5	8.4	1.2357			X	X X
4	71.5	68.0	4.7	1.1568	X	X	X	X
4	70.5	66.9	5.8	1.1769		X	X	X X
4	69.3	65.6	7.1	1.1996	X		X	X X
5	72.1	67.7	6.0	1.1625	X	X	X	X X

Fonte: [Montgomery e Runger(2016)]

Cancho, V., 2010. Notas de aulas sobre noções de estatística e probabilidade - São Paulo: USP.

Montgomery, D., Runger, G., 2016. Estatística Aplicada e Probabilidade para Engenheiros. Rio de Janeiro: LTC.