Matemática Computacional

Teórica 8

Departamento de Matemática
Instituto Superior de Engenharia do Porto

2º Semestre 21-22

Conteúdo

Regressão Linear Múltipla

RLM

Na análise de regressão encontramos situações com mais do que uma variável explicativa. Esse modelo de regressão recebe o nome de modelo de regressão múltipla (RLM).
A variável dependente ou resposta Y pode estar relacionada com k variáveis explicativas ou independentes.

Regressão Linear Múltipla

Modelo matemático

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

- X_1 , X_2, ..., X_k - são k variáveis explicativas independentes não aleatórias
- Y - variável aleatória resposta
- ε - variável aleatória erro
- β_0, β_1, β_2,..., β_k são os coeficientes de regressão
- β_1, β_2,..., β_k representam a variação da resposta Y por unidade de variação de X_j quando assumimos como constantes as restantes variáveis explicativas.

Modelo matemático

Para qualquer n-uplo de observações

$(x_{i1}, x_{i2}, ..., x_{ik}, y_i),\ i = 1, 2, ..., n (n > k)$

cada resposta y_i obtém-se da seguinte forma

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik} + \varepsilon_i,\ i = 1, 2, ..., n \quad (1)$$

Forma matricial RLM

O modelo (1) apresentado é um sistema de n equações com a representação matricial

$$Y = X\beta + \varepsilon \tag{2}$$

$$\mathbf{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \vdots & x_{1k} \\ 1 & x_{21} & x_{22} & \vdots & x_{2k} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & \vdots & x_{nk} \end{bmatrix}, \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}, \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

Estimação dos coeficientes de regresssão

Pretende-se encontrar o vetor de estimadores dos mínimos quadrados $\hat{\beta}$ que minimize a soma dos quadrados dos resíduos. Da equação (2) obtemos $\boldsymbol{\varepsilon} = \mathbf{Y} - \mathbf{X}\boldsymbol{\beta}$ e, consequentemente

$$SQ_E = \sum_{i=1}^{n} \varepsilon_i^2 = (Y - X\beta)^T (Y - X\beta) = Y^T Y - 2\beta^T X^T Y + \beta^T X^T X \beta \tag{3}$$

Estimação dos coeficientes de regresssão

O estimador dos mínimos quadrados $\hat{\beta}$ é a solução das seguintes $k+1$ equações normais:

$$\frac{\partial SQ_E}{\partial \hat{\beta}} = 0 \Leftrightarrow -2X^T Y + 2X^T X\hat{\beta} = 0 \Leftrightarrow X^T X\hat{\beta} = X^T Y \quad (4)$$

$$\hat{\beta} = \left(X^T X\right)^{-1} X^T Y \quad (5)$$

Estimação dos coeficientes de regresssão

As matrizes X^TX e X^TY são

$$X^TX = \begin{bmatrix} n & \sum_{i=1}^n x_{i1} & \sum_{i=1}^n x_{i2} & \cdots & \sum_{i=1}^n x_{ik} \\ \sum_{i=1}^n x_{i1} & \sum_{i=1}^n x_{i1}^2 & \sum_{i=1}^n x_{i1}x_{i2} & \cdots & \sum_{i=1}^n x_{i1}x_{ik} \\ \sum_{i=1}^n x_{i2} & \sum_{i=1}^n x_{i2}x_{i1} & \sum_{i=1}^n x_{i2}^2 & \cdots & \sum_{i=1}^n x_{i2}x_{ik} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \sum_{i=1}^n x_{ik} & \sum_{i=1}^n x_{ik}x_{i1} & \sum_{i=1}^n x_{ik}x_{i2} & \cdots & \sum_{i=1}^n x_{ik}^2 \end{bmatrix}$$

Estimação dos coeficientes de regresssão

$$X^T Y = \begin{bmatrix} 1 & 1 & 1 & \dots & 1 \\ x_{11} & x_{21} & x_{31} & \dots & x_{n1} \\ x_{12} & x_{22} & x_{32} & \dots & x_{n2} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{1k} & x_{2k} & x_{3k} & \dots & x_{nk} \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_{i1} y_i \\ \sum_{i=1}^n x_{i2} y_i \\ \vdots \\ \sum_{i=1}^n x_{ik} y_i \end{bmatrix}$$

Estimação dos coeficientes de regresssão

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_k \end{bmatrix} =$$

$$\begin{bmatrix} n & \sum_{i=1}^n x_{i1} & \sum_{i=1}^n x_{i2} & \cdots & \sum_{i=1}^n x_{ik} \\ \sum_{i=1}^n x_{i1} & \sum_{i=1}^n x_{i1}^2 & \sum_{i=1}^n x_{i1}x_{i2} & \cdots & \sum_{i=1}^n x_{i1}x_{ik} \\ \sum_{i=1}^n x_{i2} & \sum_{i=1}^n x_{i2}x_{i1} & \sum_{i=1}^n x_{i2}^2 & \cdots & \sum_{i=1}^n x_{i2}x_{ik} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \sum_{i=1}^n x_{ik} & \sum_{i=1}^n x_{ik}x_{i1} & \sum_{i=1}^n x_{ik}x_{i2} & \cdots & \sum_{i=1}^n x_{ik}^2 \end{bmatrix}^{-1} \begin{bmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_{i1}y_i \\ \sum_{i=1}^n x_{i2}y_i \\ \vdots \\ \sum_{i=1}^n x_{ik}y_i \end{bmatrix}$$

O modelo de regressão ajustado correspondente a 1

$$\hat{y}_i = \hat{\beta}_0 + \sum_{j=1}^{k} \hat{\beta}_j x_{ij}, i = 1, ..., n \tag{6}$$

e matricialmente

$$\hat{Y} = X\hat{\boldsymbol{\beta}} \tag{7}$$

Significância do modelo de regressão

O teste de significância para a regressão é um teste para determinar se existe uma relação linear ente a variável resposta Y e as variáveis independentes (explicativas, regressoras) $x_1, x_2, ..., x_n$. Os testes de hipóteses a realizar pressupoem que os erros ε_i sejam normalmente distribuídos de média 0 e variância σ^2.

$$H_0 : \beta_1 = \beta_2 = ... = \beta_k = 0$$

$$H_1 : \beta_j \neq 0 \text{ para algum } j, j = 1, .., k$$

Significância do modelo de regressão

Utilizamos a análise de variância para avaliarmos a significância do modelo.

$$SQ_T = SQ_R + SQ_E$$

onde

$SQ_T = Y^T Y - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n} = Y^T Y - n\bar{y}^2$ - mede a variação total das observações em torno da média

$SQ_R = \hat{\boldsymbol{\beta}}^T X^T Y - n\bar{y}^2$ - mede a variação da variável dependente explicada pelo modelo

$SQ_E = Y^T Y - \hat{\boldsymbol{\beta}}^T X^T Y$ - mede a variação não explicada pelo modelo.

Significância do modelo de regressão

Procedimento

$H_0 : \beta_1 = \beta_2 = ... = \beta_k = 0$

$H_1 : \beta_j \neq 0$ para algum $j, j = 1, .., k$

Estatística de teste : $F_0 = \dfrac{MQ_R}{MQ_E}$ onde $MQ_R = \dfrac{SQ_R}{gl_R}$ e $MQ_E = \dfrac{SQ_E}{gl_E}$

Critério de rejeição: $f_0 > f_\alpha(k, n-k+1)$

Significância do modelo de regressão

Fonte de variação	Soma de quadrados	Graus de liberdade	Média quadrática	Estatística de de teste f
Regressão	SQ_R	k	$MQ_R = \frac{SQ_R}{k}$	$\frac{MQ_R}{MQ_E}$
Erro	SQ_E	$n-(k+1)$	$MQ_E = \frac{SQ_E}{n-(k+1)}$	
Total	SQ_T	$n-1$		

Decisão:

i) Se $f_0 > f_\alpha(k, n-(k+1))$ rejeita-se H_0, o que permite concluir que há pelo menos uma variável independente que contribui significativamente para explicar a variação da variável dependente Y. Diz-se que o modelo de regressão linear apresentado é significativo.

ii) Se $f_0 \leq f_\alpha(k, n-(k+1))$ não se rejeita H_0, o que permite concluir que há um mau ajuste do modelo linear apresentado. Diz-se que o modelo não é significativo e não deve ser utilizado.

Coeficiente de determinação

Coeficiente de determinação

Tal como no modelo RLS o coeficiente de determinação é dado por:

$$R^2 = \frac{SQ_R}{SQ_T} = 1 - \frac{SQ_E}{SQ_T}$$

Este coeficiente é uma medida da proporção da variação da variável resposta Y que é explicada pela equação de regressão. Também é utilizado o coeficiente de determinação ajustado que é dado por:

$$R^2_{ajust.} = 1 - \frac{\frac{SQ_E}{n-(k+1)}}{\frac{SQ_T}{n-1}} = 1 - \left(\frac{n-1}{n-(k+1)}\right)(1-R^2)$$

Intervalos de confiança

Os estimadores $\hat{\boldsymbol{\beta}} = \left(X^T X\right)^{-1} X^T Y$ têm uma distribuição normal multivariada. Portanto $\hat{\boldsymbol{\beta}} \sim N\left(\beta, \sigma^2 C\right)$, com $C = \left(X^T X\right)^{-1}$ que é uma matriz simétrica. Logo,

$$\hat{\beta}_j \sim N\left(\beta_j, \sigma^2 C_{jj}\right)$$

C_{jj} - é o elemento j da diagonal principal da matriz C.
A variância σ^2 é estimada por $\hat{\sigma}^2 = S^2 = \frac{SQ_E}{n-(k+1)} = MQ_E$

I.C. para os coeficientes de regressão

O **intervalo de confiança** a $(1-\alpha) \times 100\%$ para os coeficientes de regressão, β_j's, com $j = 0, 1, ..., k$ é dado por

$$\left]\hat{\beta}_j - t_{1-\alpha/2}[n-(k+1)]\sqrt{\hat{\sigma}^2 C_{jj}} \ , \hat{\beta}_j + t_{1-\alpha/2}[n-(k+1)]\sqrt{\hat{\sigma}^2 C_{jj}}\right[$$

Intervalos de confiança

No caso de se pretender estimar o valor esperado da resposta Y dado $x_0^T = [1 \;\; x_{01} \;\; x_{02} \;\; \ldots x_{0k}]$. O valor esperado de $\hat{Y}$, $E\left[\hat{Y}_0\right]$ considerando x_0, é estimado por $\hat{\mu}_{Y_0} = x_0^T \hat{\boldsymbol{\beta}}$ e a variância $V\left[\hat{Y}_0\right]$ é estimada por $\hat{\sigma}^2 x_0^T \left(X^T X\right)^{-1} x_0 = \hat{\sigma}^2 x_0^T C x_0$

I.C. para os valores esperados $E(Y_0)$

O intervalo de confiança a $(1-\alpha) \times 100\%$ para $E[Y_0]$ é dado por

$$\left]\hat{\mu}_{Y_0} - t_{1-\alpha/2[n-(k+1)]}\sqrt{\hat{\sigma}^2 x_0^T C x_0} \;\; , \;\; \hat{\mu}_{Y_0} + t_{1-\alpha/2[n-(k+1)]}\sqrt{\hat{\sigma}^2 x_0^T C x_0}\right[$$

Intervalos de confiança

No caso de se pretender estimar a previsão da resposta Y dado $x_0^T = [1 \;\; x_{01} \;\; x_{02} \;\; \ldots x_{0k}]$, vamos considerar uma estimativa da previsão como $\hat{y}_0 = x_0^T \hat{\beta}$. É possível mostrar que

$$T = \frac{\hat{Y}_0 - Y_0}{\sqrt{\hat{\sigma}^2\left(1 + x_0^T C x_0\right)}} \sim t_{n-(k+1)}$$

I.C. para a previsão da resposta Y_0

O intervalo de confiança a $(1 - \alpha) \times 100\%$ para a previsão da resposta Y_0 é dado por

$$\left]\hat{y}_0 - t_{1-\alpha/2[n-(k+1)]}\sqrt{\hat{\sigma}^2\left(1 + x_0^T C x_0\right)} \;, \hat{y}_0 + t_{1-\alpha/2[n-(k+1)]}\sqrt{\hat{\sigma}^2\left(1 + x_0^T C x_0\right)}\right[$$

Testes de hipóteses

Um teste de hipóteses para os coeficientes de regressão β_j´s obedece ao seguinte procedimento:

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

Estatística de teste : $T_0 = \dfrac{\hat{\beta}_j}{\sqrt{\hat{\sigma}^2 C_{jj}}}$

Critério de rejeição: $|t_0| > t_{1-\alpha/2, n-(k+1)}$

Decisão: A rejeição de H_0 permite concluir que o regressor x_j tem poder explicativo. A não rejeição de H_0 permite concluir que o regressor x_j pode ser "eliminado".

Exemplo

Considere a seguinte tabela de 12 observações.
Y - variável dependente
X_1 e $X2$ - variáveis independentes

observação	Y	X1	X2
1	2256	80	8
2	2340	93	9
3	2426	100	10
4	2293	82	12
5	2330	90	11
6	2368	99	8
7	2250	81	8
8	2409	96	10
9	2364	94	12
10	2379	93	11
11	2440	97	13
12	2364	95	11

Exemplo

O modelo a ser ajustado é:
$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$ com forma matricial $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$
Começamos por determinar o vetor de estimadores dos mínimos quadrados $\hat{\boldsymbol{\beta}} = \left(X^T X\right)^{-1} X^T Y$. Onde

$$X = \begin{bmatrix} 1 & 80 & 8 \\ 1 & 93 & 9 \\ 1 & 100 & 10 \\ 1 & 82 & 12 \\ \vdots & \vdots & \vdots \\ 1 & 93 & 11 \\ 1 & 97 & 13 \\ 1 & 95 & 11 \end{bmatrix}, X^T = \begin{bmatrix} 1 & 1 & 1 & 1 & \dots & 1 & 1 & 1 \\ 80 & 93 & 100 & 82 & \dots & 93 & 97 & 95 \\ 8 & 9 & 10 & 12 & \dots & 11 & 13 & 11 \end{bmatrix}$$

Exemplo

As restantes matrizes necessárias

$$X^T X = \begin{bmatrix} 12 & 1100 & 123 \\ 1100 & 101370 & 11308 \\ 123 & 11308 & 1293 \end{bmatrix}$$

$$\left(X^T X\right)^{-1} = \begin{bmatrix} 16,4449 & -0,1614 & -0,1527 \\ -0,1614 & 0,0020 & -0,0020 \\ -0,1527 & -0,0020 & 0,0331 \end{bmatrix}$$

$$X^T Y = \begin{bmatrix} 28219 \\ 2591095 \\ 289809 \end{bmatrix}$$

Exemplo

Obtemos o vetor de estimadores dos coeficientes de regressão

$$\hat{\boldsymbol{\beta}} = \left(X^T X\right)^{-1} X^T Y = \begin{bmatrix} 1562,7301 \\ 7,5084 \\ 9,8131 \end{bmatrix}$$

O modelo de regressão fica definido por

$$\hat{Y} = 1562,7301 + 7,5084 X_1 + 9,8131 X_2$$

Exemplo- significância do modelo obtido

Aplicamos o teste de hipóteses

$H_0 : \beta_1 = \beta_2 = 0, k = 2$

$H_1 : \beta_j \neq 0$ para algum $j, j = 1, 2$

Estatística de teste: $F_0 = \frac{MQ_R}{MQ_E}$ onde $MQ_R = \frac{SQ_R}{gl_R}$ e $MQ_E = \frac{SQ_E}{gl_E}$

Critério de rejeição: $f_0 > f_\alpha(k, n-(k+1)) = f_\alpha(2, 12-(2+1))$

$$SQ_T = Y^T Y - \frac{\left(\sum_{i=1}^n y_i\right)^2}{n} = Y^T Y - n\bar{y}^2 = 41468,9167$$

$$SQ_R = \hat{\boldsymbol{\beta}}^T X^T Y - n\bar{y}^2 = 38223,5606$$

$$SQ_E = Y^T Y - \hat{\boldsymbol{\beta}}^T X^T Y = 3245,3561$$

Exemplo- significância do modelo obtido

Sumariamos na seguinte tabela Anova:

Fonte de variação	Soma de quadrados	Graus de liberdade	Média quadrática	Estatística de de teste f
Regressão	$38223,5606$	2	$MQ_R = 19111,7803$	$\frac{MQ_R}{MQ_E} \approx 53,0007$
Erro	$3245,3561$	9	$MQ_E = 360,5951$	
Total	$41468,9167$	11		

Decisão: Observando a tabela, verificamos que $f_0 \approx 53,0007$. Considerando $\alpha = 0,05$, obtemos $f_{0,05(2,9)} \approx 4,2565$. Como $f_0 \approx 53,0007 > 4,2565$, rejeita-se H_0, o que permite concluir que este modelo de regressão é significativo.

Exemplo- Coeficiente de determinação

Verificamos que o modelo de regressão
$\hat{Y} = 1562,7301 + 7,5084X_1 + 9,8131X_2$ é significativo.
O valor do coeficiente de determinação múltiplo é dado por:

$$R^2 = \frac{SQ_R}{SQ_T} = \frac{38223,5606}{41468,9167} = 0,9217$$

O valor do coeficiente de determinação **ajustado** é dado por:

$$R^2_{ajust.} = 1 - \left(\frac{n-1}{n-(k+1)}\right)(1-R^2) = 1 - \frac{11}{9}(1-0,9217) = 0,9043$$

Exemplo- Intervalos de confiança

I.C. para o coeficiente de regressão β_2

O **intervalo de confiança** a 95% para o coeficiente de regressão, β_2 é dado por

$$\left]\hat{\beta}_2 - t_{0,975[12-(2+1)]}\sqrt{\hat{\sigma}^2 C_{22}}\ ,\hat{\beta}_2 + t_{0,975[12-(2+1)]}\sqrt{\hat{\sigma}^2 C_{22}}\right[$$

Com

$\hat{\beta}_2 = 9,8131$

$\hat{\sigma}^2 = MQ_E = 360,5951$

$C_{22} = 0,0331$, elemento da diagonal principal da matriz $\left(X^T X\right)^{-1}$

$t_{0,975[12-(2+1)]} = t_{0,975[9]} = 2,2622$

O $I.C._{95\%}(\beta_2) =]1,9990, 17,6272[$

Exemplo- Intervalos de confiança

Consideremos $x_1 = 80$ e $x_2 = 8$, isto é , $x_0^T = \begin{bmatrix} 1 & 80 & 8 \end{bmatrix}$

I.C. para o valor esperado $E(Y_0)$

O intervalo de confiança a 95% para $E[Y_0]$ é dado por $\left]\hat{\mu}_{Y_0} - t_{0,975[9]}\sqrt{\hat{\sigma}^2 x_0^T C x_0}\ ,\ \hat{\mu}_{Y_0} + t_{0,975[9]}\sqrt{\hat{\sigma}^2 x_0^T C x_0}\right[$

Onde

$$\hat{\mu}_{Y_0} = x_0^T \hat{\boldsymbol{\beta}} = \begin{bmatrix} 1 & 80 & 8 \end{bmatrix} \begin{bmatrix} 1562,7301 \\ 7,5084 \\ 9,8131 \end{bmatrix} = 2241,9060$$

Exemplo- Intervalos de confiança

$$\hat{\sigma}^2 x_0^T C x_0 =$$

$$360,5951 \begin{bmatrix} 1 & 80 & 8 \end{bmatrix} \begin{bmatrix} 16,4449 & -0,1614 & -0,1527 \\ -0,1614 & 0,0020 & -0,0020 \\ -0,1527 & -0,0020 & 0,0331 \end{bmatrix} \begin{bmatrix} 1 \\ 80 \\ 8 \end{bmatrix}$$

$$= 149,5319$$

Assim, o $I.C._{95\%}(\mu_{Y_0}) =]2214,2436 \ ; \ 2269,5684[$

Exemplo- Intervalos de confiança

Consideremos novamente $x_1 = 80$ e $x_2 = 8$. Pretendemos estimar a previsão da resposta Y dado $x_0^T = \begin{bmatrix} 1 & 80 & 8 \end{bmatrix}$

I.C. para a previsão da resposta Y_0

O intervalo de confiança a 95% para a previsão da resposta Y_0 é dado por

$$\left]\hat{y}_0 - t_{0,975[9]}\sqrt{\hat{\sigma}^2\left(1 + x_0^T C x_0\right)}\ , \hat{y}_0 + t_{0,975[9]}\sqrt{\hat{\sigma}^2\left(1 + x_0^T C x_0\right)}\right[$$

Onde

$$\hat{y}_0 = x_0^T \hat{\boldsymbol{\beta}} = 2241,9060$$

Exemplo- Intervalos de confiança

$$x_0^T C x_0 =$$

$$\begin{bmatrix} 1 & 80 & 8 \end{bmatrix} \begin{bmatrix} 16,4449 & -0,1614 & -0,1527 \\ -0,1614 & 0,0020 & -0,0020 \\ -0,1527 & -0,0020 & 0,0331 \end{bmatrix} \begin{bmatrix} 1 \\ 80 \\ 8 \end{bmatrix} = 0,4147$$

$$\hat{\sigma}^2 \left(1 + x_0^T C x_0\right) = 510,1270$$

Assim

$$I.C._{95\%}(y_0) =]2190,8129 \ ; \ 2292,9990[$$

Exemplo - Testes de hipóteses

Pretende-se verificar se é possível admitir que o coeficiente de regressão $\beta_2 \neq 0$ com um nível de significância de 5%.

$$H_0 : \beta_2 = 0 \quad \text{v.s.} \quad H_1 : \beta_2 \neq 0$$

Estatística de teste: $T_0 = \frac{\hat{\beta}_2}{\sqrt{\hat{\sigma}^2 C_{22}}}$ onde $\hat{\sigma}^2 = MQ_E = 360,5951$ e $t_0 = \frac{9,8131}{\sqrt{360,5951 \times 0,0331}} = 2,8409$

Decisão:

$|t_0| > t_{0,975,[9]} \approx 2.2622$, rejeita-se H_0. O teste é conclusivo, o que quer dizer que a um nível de significância de 5% há evidência estatística suficiente para se afirmar que o coeficiente de regressão $\beta_2 \neq 0$.