Modelos Lineares
Regressão Linear Multipla

Susana Faria

Regressão Linear Múltipla (RLM)

- Se num modelo de regressão linear simples (RLS) introduzirmos mais variáveis explicativas passamos a ter um modelo de regressão linear múltipla (RLM). Neste caso estaremos a relacionar uma variável dependente Y com mais do que uma variável independente.
- A análise de um modelo de regressão linear múltipla é análoga à do modelo de regressão linear simples.

Considere a relação linear entre a variável resposta Y e as p variáveis explicativas $x_1, x_2, \ldots, x_p$, representada por:

$$Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p\, x_p + \epsilon$$

onde β_0, β_1, ..., β_p são designados por parâmetros ou coeficientes de regressão desconhecidos do modelo. Ao termo ϵ designamos por erro aleatório.

A este modelo designamos por Modelo de Regressão Linear Múltipla.

Regressão Linear Múltipla (RLM)

Dada uma amostra $(x_{11}, x_{12}, \ldots, x_{1p}, y_1), \ldots, (x_{n1}, x_{n2}, \ldots, x_{np}, y_n)$ de n observações independentes onde x_{ij} e y_i são, respectivamente, os valores da variável x_j e Y para o individuo i, tem-se:

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

em que:
$y_i -$ resposta aleatória do indivíduo i (variável dependente aleatória)
$x_{ij} -$ $i-$ésima observação da variável independente x_j
$\beta_0, \beta_1, \ldots, \beta_p -$ parâmetros desconhecidos do modelo
$\epsilon_i -$ erro aleatório associado à observação da resposta do indivíduo i com distribuição Normal de média 0 e variância σ^2.

Nota:

- Repare-se que agora deixamos de ter uma recta para passarmos a ter uma superfície.
- Os valores x_{ij} são considerados determinísticos (pré-determinados à partida). Os valores y_i representam a variàvel dependente e estes sim são considerados variáveis aleatórias.

Regressão Linear Múltipla (RLM)

Interpretação dos coeficientes:

β_0- representa o valor esperado da variável dependente Y quando as variáveis explicativas são simultaneamente iguais a zero.

β_j- representa a variação do valor esperado de Y por cada incremento unitário em x_j quando se mantém constantes as restantes variáveis explicativas.

Nota: Caso de interação

Regressão Linear Múltipla (RLM)

Um tratamento matricial simplifica consideravelmente os cálculos neste tipo de modelos.

O modelo de regressão pode ser representado por:

$$Y = X\beta + \epsilon$$

onde

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, X = \begin{bmatrix} 1 & x_{11}, \dots, x_{1p} \\ 1 & x_{21}, \dots, x_{2p} \\ \vdots & \\ 1 & x_{n1}, \dots, x_{np} \end{bmatrix}, \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}, \epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

em que Y é um vector coluna $(n \times 1)$ de observações da variável resposta, X é uma matriz $(n \times (p+1))$ cujas linhas são constituídas pelos valores das variáveis independentes, β é um vector coluna $((p+1) \times 1)$ de parâmetros da regressão e ϵ é um vector coluna $(n \times 1)$ de erros aleatórios.

Regressão Linear Múltipla (RLM)

Pressupostos usuais do modelo RLM :

- $E[\epsilon_i] = 0$
- $Var[\epsilon_i] = \sigma^2 \ \ \forall i$ (variância constante desconhecida).
- ϵ_i's são variáveis aleatórias independentes.
- $\epsilon_i \sim N(0, \sigma^2), \ i = 1, \ldots, n$ então $\epsilon \sim N(\mathbf{0}, \sigma^2 I_n)$ onde I_n é a matriz identidade de ordem n.
- $cov(y_i, y_j) = 0, i \neq j, (i, j = 1, \ldots, n)$
- as variáveis explicativas não devem estar correlacionadas.

De forma semelhante ao modelo de regressão linear:

- A variável resposta é função linear das variáveis explicativas.

Estimação dos Parâmetros de um Modelo RLM

- Um método de estimação dos coeficientes de regressão β é o método dos mínimos quadrados que consiste em minimizar a soma de quadrados dos erros aleatórios:

$$SSE = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip})]^2 = (Y - X\beta)^T (Y - X\beta)$$

- Obtém-se o estimador dos mínimos quadrados:

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

Nota: $(X^T X)$ não é invertível quando:

- $n < p + 1$, ou seja, há poucas observações.
- uma (ou várias) variável explicativa é uma combinação linear das outras variáveis explicativas.

Exercicio

Considere o modelo $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \epsilon_i$ e os seguintes dados:

$y = [-43.6, 3.3, 12.4, 7.6, 11.4, 5.9, -4.5, 22.7, -14.4, -28.3]$

$x_1 = [27, 33, 27, 24, 31, 40, 15, 26, 22, 23]$

$x_2 = [34, 30, 33, 11, 16, 30, 17, 12, 21, 27]$

Obter as estimativas de minimos quadrados de β. (Não usar o comando lm)

Estimação dos Parâmetros de um Modelo RLM

- A **equação da superfície de regressão pode ser escrita:**

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_p x_p$$

onde $(x_1, \ldots, x_p)$ **representa qualquer ponto de** IR^p**.**

- **Os valores** $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_p x_{ip}$ **designam-se por valores estimados ou valores preditosde** y_i **, em inglês, "fitted values"ou "predicted values".**
- **Na forma matricial** $\hat{Y} = X\hat{\beta} = X(X^T X)^{-1} X^T Y = H\ Y$ **onde** H **designa-se a matriz hat.**
- **As quantidades** $e_i = y_i - \hat{y}_i$ **são designados por resíduos.**
- **O vector** $(n \times 1)$ **dos resíduos é** $e = Y - \hat{Y} = (I - H)Y$**.**

Propriedades dos Estimadores de um Modelo RLM

- Tal como acontecia no modelo de regressão linear simples, prova-se que as estimativas para os parâmetros da regressão são centradas, ou seja:

$$E[\hat{\beta}] = \beta$$

- As variâncias dos estimadores são dadas pelos elementos da diagonal da matriz:

$$\sigma^2 (X^T X)^{-1}$$

Os estimadores dos Mínimos Quadrados dos parâmetros :

- são combinações lineares de Y_i.
- são **centrados** ou **não enviesados**.
- têm variância mínima.
- são, de entre os centrados, os de menor variância. (**BLUES**)

Estimador centrado de σ^2:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n-p-1} = \frac{SSE}{n-p-1}$$

onde SSE é a soma dos quadrados dos resíduos.

Nota: $\frac{(n-p-1)^2 \hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-p-1}$

Inferências no Modelo de Regressão Linear Múltipla

Inferências sobre β_j

- Pretende-se testar a hipótese:

$$H_0 : \beta_j = b_j \quad \text{vs} \quad H_1 : \beta_j \neq b_j$$

A estatística-teste é:

$$T = \frac{\hat{\beta}_j - b_j}{\sqrt{\hat{\sigma}^2 C_{jj}}} \quad \sim \quad t_{n-p-1}$$

omde C_{jj} é o $j-$ésimo elemento da diagonal principal da matriz $C = (X^T X)^{-1}$.

A região de rejeição é:
$RC = \{t : |t| > t_{\frac{\alpha}{2};n-p-1}\}$ em que α é o nível de significância.

- Pretende-se calcular o intervalo de confiança para β_j:

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

Inferências no Modelo de Regressão Linear Múltipla

Estimação do valor esperado de Y para um determinado x_0:
$E[Y_0] = E[Y|x = \mathbf{x}_0] = \mathbf{x}^{\mathbf{T}}{}_0\beta$

- Estimador pontual:

$$\hat{E}[Y_0] = \mathbf{x}^{\mathbf{T}}{}_0\hat{\beta}$$

- Intervalo de confiança a $100(1-\alpha)\%$ para $E[Y_0]$:

$$\left(\hat{E}[Y_0] - t_{\frac{\alpha}{2};n-p-1}\sqrt{\hat{\sigma}^2\mathbf{x}^{\mathbf{T}}{}_0\ C\ \mathbf{x}_0}; \hat{E}[Y_0] + t_{\frac{\alpha}{2};n-p-1}\sqrt{\hat{\sigma}^2\mathbf{x}^{\mathbf{T}}{}_0\ C\ \mathbf{x}_0}\right)$$

Nota: As inferências podem não ser válidas fora do intervalo de valores de x considerado.

Inferências no Modelo de Regressão Linear Múltipla

Previsão do valor de Y para o ponto x_0: $Y_0 = \mathbf{x^T}_0\beta$

- Estimador pontual:

$$\hat{Y}_0 = \mathbf{x^T}_0\hat{\beta}$$

- Intervalo de confiança a $100(1-\alpha)\%$ para Y_0:

$$\left(\hat{Y}_0 - t_{\frac{\alpha}{2};n-p-1}\sqrt{\hat{\sigma}^2[1 + \mathbf{x^T}_0\ C\ \mathbf{x}_0]};\ \hat{Y}_0 + t_{\frac{\alpha}{2};n-p-1}\sqrt{\hat{\sigma}^2[1 + \mathbf{x^T}_0\ C\ \mathbf{x}_0]}\right)$$

Nota: As inferências podem não ser válidas fora do intervalo de valores de x considerado.

ANOVA

Para testar:

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0 \quad \text{vs} \quad H_1 : \text{Pelo menos um} \beta_j \neq 0 \;\; j = 1, \ldots, p$$

A tabela da ANOVA correspondente é:

Fonte de Variação	SS	gl	MS	F_0	$p - value$
Regressão	SSR	p	MSR	$\frac{MSR}{MSE}$	
Erros	SSE	n-p-1	MSE		
Total	SST	n-1			

Rejeita-se a hipótese H_0 ao nível de significância α se o valor da estatística de teste, F_0 for maior do que o valor de F com (p,n-p-1) graus de liberdade.

Notas:

-

Avaliação da qualidade e significado da regressão

Para avaliar a qualidade e significado da regressão vamos considerar vários métodos.

- Métodos Gráficos
- Teste ao significado da regressão
- Coeficiente de Determinação

Métodos Gráficos

- Habitualmente constrói-se diagramas de dispersão para visualizar a relação entre Y e cada um dos regressores individualmente.
- É também habitual construir um gráfico de dispersão que apresente os valores observados Y_i versus os valores preditos $\hat{Y}_i$.

Teste ao significado da regressão

Será que Y depende mesmo de x? Podemos responder a esta questãoo através da ANOVA

Coeficiente de Determinação

Definição

O coeficiente de determinação é uma medida relativa de ajustamento do modelo de regressão linear, dada por:

$$R^2 = 1 - \frac{SSE}{SST} = \frac{SSR}{SST}$$

É interpretada como a percentagem da variabilidade de Y que é explicada pelo modelo de regressão linear.

O coeficiente de determinação é tal que $0 \leq R^2 \leq 1$.

Atenção: Este coefiente pode induzir em erro. Ao adicionarmos variáveis (regressores) ao modelo estamos sempre a aumentar o valor de R^2 e nem sempre essas variáveis são estatisticamente significativas.

Exercício: Verifique que: $F = \frac{n-p-1}{p} \frac{R^2}{1-R^2}$

Coeficiente de Determinação

- Tal como na regressão simples, define-se um coeficiente de determinação ajustado:

$$R_a^2 = 1 - \frac{\frac{SSE}{(n-p-1)}}{\frac{SST}{(n-1)}} = 1 - \frac{(n-1)}{(n-p-1)}(1-R^2)$$

Relação entre o Coeficiente de Correlação e o Coeficiente de Determinação

- No caso do modelo RLM, o coeficiente de determinação (R^2) é o quadrado do coeficiente de correlação entre $\hat{y}$ e y, ($r_{\hat{y}y}$).

Notas Importantes:

- Para comparar dois modelos que têm o mesmo número de variáveis explicativas, pode-se comparar os respectivos valores de R^2 e escolher o modelo com maior coeficiente de determinação.
- É possível comparar dois modelos que não têm o mesmo número de variáveis explicativas, escolhendo o modelo com maior valor de R_a^2.
- Critério de C_p de Mallows.

Testar se alguns Parâmetros são Nulos

- Avaliar se apenas um parâmetro é Nulo
- Avaliar se um número r de parâmetros é nulo:

$$H_0 : Y_i = \beta_0 + \beta_1 \, x_{i1} + \cdots + \beta_{p-r} \, x_{i(p-r)} + \epsilon_i \quad \text{vs}$$
$$H_1 : Y_i = \beta_0 + \beta_1 \, x_{i1} + \cdots + \beta_p \, x_{ip} + \epsilon_i$$

A estatística-teste é:

$$F_0 = \frac{n-p-1}{r} \frac{SSE(H_0) - SSE(H_1)}{SSE(H_1)}$$

em que:
$SSE(H_0)$: a soma dos quadrados dos resíduos obtidos usando o modelo H_0
$SSE(H_1)$: a soma dos quadrados dos resíduos obtidos usando o modelo H_1

Rejeita-se a hipótese H_0 ao nível de significância α se o valor da estatística de teste, F_0 for maior do que o valor de F com (r,n-p-1) graus de liberdade.

Exemplo de Aplicação

Considere $n = 46$ e
X_1 : Percentagem da população urbana
X_2 : 100/(número de crian ças nascidas)
X_3 : Consumo de vinho por pessoa
X_4 : Consumo de alcool por pessoa
Y : Taxa de falecimentos devido a doença no figado
Tem-se $SSE(X_1, X_2, X_3, X_4) = 4609.704$ $SSE(X_1, X_2, X_3) = 4624.796$
$SSE(X_1, X_2, X_4) = 7024.307$ $SSE(X_1, X_3, X_4) = 5045.36$
$SSE(X_2, X_3, X_4) = 4627.987$ $SSE(X_1, X_2) = 8909.233$
$SSE(X_1, X_3) = 5673.415$ $SSE(X_1, X_4) = 7050.957$
$SSE(X_2, X_3) = 4630.927$ $SSE(X_1, X_3) = 8602.508$
$SSE(X_3, X_4) = 6526.226$ $SSE(X_1) = 10866.33$
$SSE(X_2) = 9586.892$ $SSE(X_3) = 7091.082$ $SSE(X_4) = 13230.48$
$SST = 24753.5$
a) Qual o modelo que escolheria?
b) Teste as seguintes hipóteses:

- $H_0 : \beta_4 = 0$
- $H_0 : \beta_1 = \beta_4 = 0$
- $H_0 : \beta_1 = \beta_2 = \beta_4 = 0$