SME0822 Análise Multivariada e Aprendizado Não-Supervisionado

Aula 6c: **Regressão Multivariada**

Prof. Cibele Russo

cibele@icmc.usp.br

http://www.icmc.usp.br/~cibele

Baseado em Johnson, R. A., & Wichern, D. W. (2007). Applied Multivariate Statistical Analysis. Prentice Hall.

Modelo de regressão linear múltipla (univariado)

Motivação: Deseja-se construir um modelo para explicar

- Y: valor de mercado de uma casa

  utilizando variáveis explicativas
- Z_1: área
- Z_2: localização
- Z_3: valor da casa no ano anterior
- Z_4: qualidade da construção

Um possível modelo linear (nos parâmetros) seria:

$$\underbrace{Y}_{\text{v. resposta}} = \underbrace{\beta_0 + \beta_1 Z_1 + \beta_2 Z_2 + \beta_3 Z_3 + \beta_4 Z_4}_{\text{componente sistemática}} + \underbrace{\epsilon}_{\text{erro aleatório}}.$$

Modelo de regressão linear múltipla (univariado)

Se coletarmos n observações dessas variáveis, podemos escrever

$$Y_i = \beta_0 + \beta_1 Z_{1i} + \beta_2 Z_{2i} + \beta_3 Z_{3i} + \beta_4 Z_{4i} + \epsilon_i, i = 1, \ldots, n.$$

Nomenclatura:

- Y_i: variável resposta (dependente),
- β_j: parâmetros desconhecidos,
- Z_{ji}: variáveis explicativas (covariáveis, variáveis independentes),
- ϵ_i: erro aleatório.

Suposições:

- $\mathrm{E}(\epsilon_i) = 0$ para $i = 1, \ldots, n$,
- $\mathrm{Var}(\epsilon_i) = \sigma^2$ para $i = 1, \ldots, n$,
- $\mathrm{Cov}(\epsilon_i, \epsilon_j) = 0$ para $i, j = 1, \ldots, n$ e $i \neq j$.

Modelo de regressão linear múltipla (univariado)

Se coletarmos n observações dessas variáveis, podemos escrever

$$Y_i = \beta_0 + \beta_1 Z_{1i} + \beta_2 Z_{2i} + \beta_3 Z_{3i} + \beta_4 Z_{4i} + \epsilon_i, i = 1, \ldots, n.$$

Nomenclatura:

- Y_i: variável resposta (dependente),
- β_j: parâmetros desconhecidos,
- Z_{ji}: variáveis explicativas (covariáveis, variáveis independentes),
- ϵ_i: erro aleatório.

Suposições:

- $\text{E}(\epsilon_i) = 0$ para $i = 1, \ldots, n$,
- $\text{Var}(\epsilon_i) = \sigma^2$ para $i = 1, \ldots, n$,
- $\text{Cov}(\epsilon_i, \epsilon_j) = 0$ para $i, j = 1, \ldots, n$ e $i \neq j$.

Modelo de regressão linear múltipla (univariado)

Poderíamos estender esse modelo para p covariáveis,

$$Y_i = \beta_0 + \beta_1 Z_{1i} + \beta_2 Z_{2i} + \ldots + \beta_p Z_{pi} + \epsilon_i, i = 1, \ldots, n.$$

Note que a variável resposta Y_i é unidimensional.

Modelo de regressão linear múltipla (univariado)

Poderíamos "empilhar" os dados de n indivíduos em linhas. Teríamos então matricialmente

$$\underset{\sim}{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \; Z = \begin{bmatrix} 1 & Z_{11} & \dots & Z_{1p} \\ 1 & Z_{21} & \dots & Z_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & Z_{n1} & \dots & Z_{np} \end{bmatrix}, \underset{\sim}{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}, \; \underset{\sim}{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

ou seja,

$$\underset{\sim}{Y}_{n\times 1} = Z_{n\times(p+1)} \underset{\sim}{\beta}_{(p+1)\times 1} + \underset{\sim}{\epsilon}_{n\times 1}.$$

Modelo de regressão linear múltipla (univariado)

No modelo

$$\underset{\sim}{Y} = Z\underset{\sim}{\beta} + \underset{\sim}{\epsilon}.$$

com as suposições

- $\mathsf{E}(\underset{\sim}{\epsilon}) = \underset{\sim}{0}$,
- $\mathsf{Var}(\underset{\sim}{\epsilon}) = \sigma^2 I$,

o estimador de mínimos quadrados é dado por

$$\widehat{\underset{\sim}{\beta}} = (Z^\top Z)^{-1} Z^\top \underset{\sim}{Y}.$$

Modelo de regressão linear multivariado

Considere agora que, para cada indivíduo, sejam observadas m variáveis respostas, e que cada uma delas tenha uma relação linear com as p covariáveis.

Assim, teriamos m modelos de regressão:

$$
\begin{aligned}
Y_1 &= \beta_{01} + \beta_{11}Z_1 + \beta_{21}Z_2 + \beta_{31}Z_3 + \ldots + \beta_{p1}Z_p + \epsilon_1 \\
Y_2 &= \beta_{02} + \beta_{12}Z_1 + \beta_{22}Z_2 + \beta_{32}Z_3 + \ldots + \beta_{p2}Z_p + \epsilon_2 \\
&\vdots \\
Y_m &= \beta_{0m} + \beta_{1m}Z_1 + \beta_{2m}Z_2 + \beta_{3m}Z_3 + \ldots + \beta_{pm}Z_p + \epsilon_m
\end{aligned}
$$

Para cada um dos n indivíduos, vamos observar as m variáveis resposta e as p covariáveis.

Modelo de regressão linear multivariado

Para cada um dos n indivíduos, vamos observar as m variáveis resposta e as p covariáveis. Assim, podemos definir um modelo de regressão multivariado

$$Y_{n\times m} = Z_{n\times(p+1)}\beta_{(p+1)\times m} + \epsilon_{n\times m}$$

em que

Modelo de regressão linear multivariado

$$Y_{n\times m} = Z_{n\times(p+1)}\beta_{(p+1)\times m} + \epsilon_{n\times m}$$

em que

$$Y = \begin{bmatrix} Y_{11} & Y_{12} & \dots & Y_{1m} \\ Y_{21} & Y_{22} & \dots & Y_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ Y_{n1} & Y_{n2} & \dots & Y_{nm} \end{bmatrix}, \; Z = \begin{bmatrix} 1 & Z_{11} & \dots & Z_{1p} \\ 1 & Z_{21} & \dots & Z_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & Z_{n1} & \dots & Z_{np} \end{bmatrix},$$

$$\beta = \begin{bmatrix} \beta_{01} & \beta_{02} & \dots & \beta_{0m} \\ \beta_{11} & \beta_{12} & \dots & \beta_{1m} \\ \vdots & \vdots & \ddots & \vdots \\ \beta_{p1} & \beta_{p2} & \dots & \beta_{pm} \end{bmatrix}, \; \epsilon = \begin{bmatrix} \epsilon_{11} & \epsilon_{12} & \dots & \epsilon_{1m} \\ \epsilon_{21} & \epsilon_{22} & \dots & \epsilon_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \epsilon_{n1} & \epsilon_{n2} & \dots & \epsilon_{nm} \end{bmatrix}$$

Modelo de regressão linear multivariado

Considere

$$\epsilon = \begin{bmatrix} \epsilon_{11} & \epsilon_{12} & \cdots & \epsilon_{1m} \\ \epsilon_{21} & \epsilon_{22} & \cdots & \epsilon_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \epsilon_{n1} & \epsilon_{n2} & \cdots & \epsilon_{nm} \end{bmatrix} = \left[\underset{\sim}{\epsilon}_{(1)}, \ldots, \underset{\sim}{\epsilon}_{(n)} \right]^\top$$

Modelo de regressão linear multivariado

$$Y_{n\times m} = Z_{n\times (p+1)}\beta_{(p+1)\times m} + \epsilon_{n\times m}$$

Suposições:

- $\text{E}(\underset{\sim}{\epsilon}_{(i)}) = \underset{\sim}{0}$ para $i = 1, \ldots, n$,
- $\text{Var}(\underset{\sim}{\epsilon}_{(i)}) = \Sigma$ para $i = 1, \ldots, n$,
- $\text{Cov}(\underset{\sim}{\epsilon}_i, \underset{\sim}{\epsilon}_k) = \sigma^2_{ik} I$ para $i, k = 1, \ldots, m$ e $i \neq k$.

Ou seja, erros em indivíduos distintos são não-correlacionados mas as observações de variáveis diferentes podem ser correlacionadas para um mesmo indivíduo.

Modelo de regressão linear multivariado

Podemos estimar β fazendo

$$\widehat{\beta} = (Z^\top Z)^{-1} Z^\top Y.$$

Exercício: verifique as dimensões dos elementos acima.

Modelo de regressão linear multivariado

As **predições** podem ser obtidas fazendo

$$\widehat{Y} = Z\widehat{\beta} = Z(Z^\top Z)^{-1}Z^\top Y,$$

que é linear em Y.

Os **resíduos** são dados por

$$Y - \widehat{Y} = Y - Z(Z^\top Z)^{-1}Z^\top Y = (I - Z(Z^\top Z)^{-1}Z^\top)Y.$$

Modelo de regressão linear multivariado

Propriedades:

Se $\epsilon_{(i)} \sim N(\underset{\sim}{0}, \Sigma)$, $i = 1, \ldots, r(\Sigma) = p + 1$ e $n \geq p + 1 + m$, então

1. $\widehat{\beta} = (Z^\top Z)^{-1} Z^\top Y$ é o EMV de β.
2. $\widehat{\beta}$ tem distribuição normal com $E(\widehat{\beta}) = \beta$.
3. $\widehat{\Sigma} = \dfrac{1}{n}(Y - Z\widehat{\beta})^\top (Y - Z\widehat{\beta})$.

Modelo de regressão linear multivariado

Somas de Quadrados:

- $Y^\top Y$: somas de quadrados e produtos cruzados **total**
- $\widehat{Y}^\top \widehat{Y}$: somas de quadrados e produtos cruzados **predito**
- $\widehat{\epsilon}^\top \widehat{\epsilon}$: somas de quadrados e produtos cruzados **do resíduo**

Modelo de regressão linear multivariado

Soma de quadrados e produtos cruzados do resíduo:

$$\widehat{\epsilon}^\top \widehat{\epsilon} = Y^\top Y - \widehat{Y}^\top \widehat{Y} =$$

$$= Y^\top Y - \widehat{\beta}^\top X^\top X \widehat{\beta} =$$

$$= Y^\top Y - Y^\top X (X^\top X)^{-1} X^\top X (X^\top X)^{-1} X^\top Y$$

$$= Y^\top (I - H) Y$$

Modelo de regressão linear multivariado

Soma de quadrados e produtos cruzados do resíduo:

$$\widehat{\epsilon}^\top\widehat{\epsilon} = Y^\top Y - \widehat{Y}^\top\widehat{Y} =$$

$$= Y^\top Y - \widehat{\beta}^\top X^\top X\widehat{\beta} =$$

$$= Y^\top Y - Y^\top X(X^\top X)^{-1}X^\top X(X^\top X)^{-1}X^\top Y$$

$$= Y^\top(I - H)Y$$

Modelo de regressão linear multivariado

Soma de quadrados e produtos cruzados do resíduo:

$$\widehat{\epsilon}^\top \widehat{\epsilon} = Y^\top Y - \widehat{Y}^\top \widehat{Y} =$$

$$= Y^\top Y - \widehat{\beta}^\top X^\top X \widehat{\beta} =$$

$$= Y^\top Y - Y^\top X (X^\top X)^{-1} X^\top X (X^\top X)^{-1} X^\top Y$$

$$= Y^\top (I - H) Y$$

Modelo de regressão linear multivariado

Soma de quadrados e produtos cruzados do resíduo:

$$\widehat{\epsilon}^\top \widehat{\epsilon} = Y^\top Y - \widehat{Y}^\top \widehat{Y} =$$
$$= Y^\top Y - \widehat{\beta}^\top X^\top X \widehat{\beta} =$$
$$= Y^\top Y - Y^\top X (X^\top X)^{-1} X^\top X (X^\top X)^{-1} X^\top Y$$
$$= Y^\top (I - H) Y$$

Modelo de regressão linear multivariado: Exemplo

Exercício 7.26 de Johnson & Wichern (2007): Deseja-se explicar a resistência de alguns tipos de fibra de celulose. Em um experimento, foram obtidas n=62 medidas de fibras de celulose e papel. Esses dados estão disponíveis na library robustbase do R sob o nome de pulpfiber. As variáves são:

- Y_1: comprimento na quebra
- Y_2: módulo de elasticidade
- Y_3: estresse na falha
- Y_4: resistência à quebra
- Z_1: comprimento da fibra
- Z_2: fração de fibra grossa
- Z_3: fração de fibra fina
- Z_4: extensão à tração nula

Modelo de regressão linear multivariado: Exemplo

1. Ajuste modelos de regressão linear múltipla com cada variável resposta Y_i e realize uma análise de resíduos para cada um desses modelos.
2. Ajuste um modelo de regressão linear multivariada para o vetor de respostas $(Y_1, Y_2, Y_3, Y_4)^\top$.