



SME0822 Análise Multivariada e Aprendizado Não-Supervisionado

Aula 6c: **Regressão Multivariada**

Prof. Cibeles 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:

$$Y = \beta_0 + \beta_1 Z_1 + \beta_2 Z_2 + \beta_3 Z_3 + \beta_4 Z_4 + \epsilon.$$

v. resposta

componente sistemática

erro aleatório

Modelo de regressão linear múltipla (univariado)

Se coletarmos n observações dessas variáveis, podemos escrever

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

Nomenclatura:

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

Suposições:

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

Modelo de regressão linear múltipla (univariado)

Se coletarmos n observações dessas variáveis, podemos escrever

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

Nomenclatura:

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

Suposições:

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

Modelo de regressão linear múltipla (univariado)

Poderíamos estender esse modelo para p covariáveis,

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

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

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

ou seja,

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

Modelo de regressão linear múltipla (univariado)

No modelo

$$\underline{Y} = \underline{Z}\underline{\beta} + \underline{\epsilon}.$$

com as suposições

- $E(\underline{\epsilon}) = \underline{0}$,
- $\text{Var}(\underline{\epsilon}) = \sigma^2 I$,

o estimador de mínimos quadrados é dado por

$$\hat{\underline{\beta}} = (\underline{Z}^T \underline{Z})^{-1} \underline{Z}^T \underline{Y}.$$

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, teríamos m modelos de regressão:

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

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

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

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

Modelo de regressão linear multivariado

Considere

$$\epsilon = \begin{bmatrix} \epsilon_{11} & \epsilon_{12} & \dots & \epsilon_{1m} \\ \epsilon_{21} & \epsilon_{22} & \dots & \epsilon_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \epsilon_{n1} & \epsilon_{n2} & \dots & \epsilon_{nm} \end{bmatrix} = [\underline{\epsilon}_{(1)}, \dots, \underline{\epsilon}_{(n)}]^T$$

Modelo de regressão linear multivariado

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

Suposições:

- $E(\underline{\epsilon}_{(i)}) = \underline{0}$ para $i = 1, \dots, n$,
- $\text{Var}(\underline{\epsilon}_{(i)}) = \Sigma$ para $i = 1, \dots, n$,
- $\text{Cov}(\underline{\epsilon}_i, \underline{\epsilon}_k) = \sigma_{ik}^2 I$ para $i, k = 1, \dots, m$ 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

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

Exercício: verifique as dimensões dos elementos acima.

Modelo de regressão linear multivariado

As **predições** podem ser obtidas fazendo

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

que é linear em Y .

Os **resíduos** são dados por

$$Y - \hat{Y} = Y - Z(Z^T Z)^{-1}Z^T Y = (I - Z(Z^T Z)^{-1}Z^T)Y.$$

Modelo de regressão linear multivariado

Propriedades:

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

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

Modelo de regressão linear multivariado

Somas de Quadrados:

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

Modelo de regressão linear multivariado

Soma de quadrados e produtos cruzados do resíduo:

$$\begin{aligned}\hat{\epsilon}^T \hat{\epsilon} &= Y^T Y - \hat{Y}^T \hat{Y} = \\ &= Y^T Y - \hat{\beta}^T X^T X \hat{\beta} = \\ &= Y^T Y - Y^T X (X^T X)^{-1} X^T X (X^T X)^{-1} X^T Y \\ &= Y^T (I - H) Y\end{aligned}$$

Modelo de regressão linear multivariado

Soma de quadrados e produtos cruzados do resíduo:

$$\begin{aligned}\hat{\epsilon}^T \hat{\epsilon} &= Y^T Y - \hat{Y}^T \hat{Y} = \\ &= Y^T Y - \hat{\beta}^T X^T X \hat{\beta} = \\ &= Y^T Y - Y^T X (X^T X)^{-1} X^T X (X^T X)^{-1} X^T Y \\ &= Y^T (I - H) Y\end{aligned}$$

Modelo de regressão linear multivariado

Soma de quadrados e produtos cruzados do resíduo:

$$\begin{aligned}\hat{\epsilon}^T \hat{\epsilon} &= Y^T Y - \hat{Y}^T \hat{Y} = \\ &= Y^T Y - \hat{\beta}^T X^T X \hat{\beta} = \\ &= Y^T Y - Y^T X (X^T X)^{-1} X^T X (X^T X)^{-1} X^T Y \\ &= Y^T (I - H) Y\end{aligned}$$

Modelo de regressão linear multivariado

Soma de quadrados e produtos cruzados do resíduo:

$$\begin{aligned}\hat{\epsilon}^T \hat{\epsilon} &= Y^T Y - \hat{Y}^T \hat{Y} = \\ &= Y^T Y - \hat{\beta}^T X^T X \hat{\beta} = \\ &= Y^T Y - Y^T X (X^T X)^{-1} X^T X (X^T X)^{-1} X^T Y \\ &= Y^T (I - H) Y\end{aligned}$$

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 variáveis 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 $(Y_1, Y_2, Y_3, Y_4)^T$.