DATA ANALYTICS

Unit 2:Multivariate Regression

Mamatha.H.R
Department of Computer Science and Engineering

DATA ANALYTICS

Unit 2:Multivariate Regression

Mamatha H R

Department of Computer Science and Engineering

Multivariate analysis (MVA): Involves simultaneous analysis of more than one outcome variable.

– Some multivariate analytic methods includes no independent variables

– Others include several independent Variables

Examples:

- Bivariate probit regression
- Multivariate probit regression
- Multivariate analysis of variance (MANOVA)
- Latent class analysis
- Path analysis

MULTIVARIATE Regression

Multivariate regression is a technique that estimates a single regression model with more than one outcome variable.

When there is more than one predictor variable in a multivariate regression model, the model is a multivariate multiple regression.

Examples of multivariate regression

- Example 1. A researcher has collected data on three psychological variables, four academic variables (standardized test scores), and the type of educational program the student is in for 600 high school students. She is interested in how the set of psychological variables is related to the academic variables and the type of program the student is in.
- Example 2. A doctor has collected data on cholesterol, blood pressure, and weight. She also collected data on the eating habits of the subjects (e.g., how many grams of meat, fish, dairy products, and chocolate consumed per week). She wants to investigate the relationship between the three measures of health and eating habits.

MvLR Model: Scalar Form

Multivariate Regression: Predict multiple dependent variables
using multiple independent variables

The multivariate (multiple) linear regression model has the form

$$y_{ik} = b_{0k} + \sum_{j=1}^{p} b_{jk} x_{ij} + e_{ik}$$

for $i \in \{1, \ldots, n\}$ and $k \in \{1, \ldots, m\}$ where

- $y_{ik} \in \mathbb{R}$ is the k-th real-valued response for the i-th observation
- $b_{0k} \in \mathbb{R}$ is the regression intercept for k-th response
- $b_{jk} \in \mathbb{R}$ is the j-th predictor's regression slope for k-th response
- $x_{ij} \in \mathbb{R}$ is the j-th predictor for the i-th observation
- $(e_{i1}, \ldots, e_{im}) \overset{\text{iid}}{\sim} \mathrm{N}(\mathbf{0}_m, \boldsymbol{\Sigma})$ is a multivariate Gaussian error vector

MvLR Model: Assumptions

The fundamental assumptions of the MLR model are:

1. Relationship between X_j and Y_k is linear (given other predictors)
2. x_{ij} and y_{ik} are observed random variables (known constants)
3. $(e_{i1}, \ldots, e_{im}) \overset{\text{iid}}{\sim} \mathrm{N}(\mathbf{0}_m, \mathbf{\Sigma})$ is an unobserved random vector
4. $\mathbf{b}_k = (b_{0k}, b_{1k}, \ldots, b_{pk})'$ for $k \in \{1, \ldots, m\}$ are unknown constants
5. $(y_{ik} | x_{i1}, \ldots, x_{ip}) \sim \mathrm{N}(b_{0k} + \sum_{j=1}^{p} b_{jk} x_{ij}, \sigma_{kk})$ for each $k \in \{1, \ldots, m\}$
   note: homogeneity of variance for each response

Note: b_{jk} is expected increase in Y_k for 1-unit increase in X_j with all other predictor variables held constant

MLR Model: Matrix Form

The multivariate multiple linear regression model has the form

$$\mathbf{Y} = \mathbf{X}\mathbf{B} + \mathbf{E}$$

where

- $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_m] \in \mathbb{R}^{n \times m}$ is the $n \times m$ response matrix
  - $\mathbf{y}_k = (y_{1k}, \ldots, y_{nk})' \in \mathbb{R}^n$ is k-th response vector $(n \times 1)$
- $\mathbf{X} = [\mathbf{1}_n, \mathbf{x}_1, \ldots, \mathbf{x}_p] \in \mathbb{R}^{n \times (p+1)}$ is the $n \times (p+1)$ design matrix
  - $\mathbf{1}_n$ is an $n \times 1$ vector of ones
  - $\mathbf{x}_j = (x_{1j}, \ldots, x_{nj})' \in \mathbb{R}^n$ is j-th predictor vector $(n \times 1)$
- $\mathbf{B} = [\mathbf{b}_1, \ldots, \mathbf{b}_m] \in \mathbb{R}^{(p+1) \times m}$ is $(p+1) \times m$ matrix of coefficients
  - $\mathbf{b}_k = (b_{0k}, b_{1k}, \ldots, b_{pk})' \in \mathbb{R}^{p+1}$ is k-th coefficient vector $(p+1 \times 1)$
- $\mathbf{E} = [\mathbf{e}_1, \ldots, \mathbf{e}_m] \in \mathbb{R}^{n \times m}$ is the $n \times m$ error matrix
  - $\mathbf{e}_k = (e_{1k}, \ldots, e_{nk})' \in \mathbb{R}^n$ is k-th error vector $(n \times 1)$

MLR Model: Matrix Form (another look)

Matrix form writes MLR model for all *nm* points simultaneously

$$\mathbf{Y} = \mathbf{XB} + \mathbf{E}$$

$$\begin{pmatrix} y_{11} & \cdots & y_{1m} \\ y_{21} & \cdots & y_{2m} \\ y_{31} & \cdots & y_{3m} \\ \vdots & \ddots & \vdots \\ y_{n1} & \cdots & y_{nm} \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ 1 & x_{31} & x_{32} & \cdots & x_{3p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix} \begin{pmatrix} b_{01} & \cdots & b_{0m} \\ b_{11} & \cdots & b_{1m} \\ b_{21} & \cdots & b_{2m} \\ \vdots & \ddots & \vdots \\ b_{p1} & \cdots & b_{pm} \end{pmatrix} + \begin{pmatrix} e_{11} & \cdots & e_{1m} \\ e_{21} & \cdots & e_{2m} \\ e_{31} & \cdots & e_{3m} \\ \vdots & \ddots & \vdots \\ e_{n1} & \cdots & e_{nm} \end{pmatrix}$$

Fitted Values and Residuals

SCALAR FORM:

Fitted values are given by

$$\hat{y}_{ik} = \hat{b}_{0k} + \sum_{j=1}^{p} \hat{b}_{jk} x_{ij}$$

and residuals are given by

$$\hat{e}_{ik} = y_{ik} - \hat{y}_{ik}$$

MATRIX FORM:

Fitted values are given by

$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\mathbf{B}}$$

and residuals are given by

$$\hat{\mathbf{E}} = \mathbf{Y} - \hat{\mathbf{Y}}$$

Hat Matrix

Note that we can write the fitted values as

$$\begin{aligned}\hat{\mathbf{Y}} &= \mathbf{X}\hat{\mathbf{B}} \\ &= \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \\ &= \mathbf{H}\mathbf{Y}\end{aligned}$$

where $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ is the hat matrix.

$\mathbf{H}$ is a symmetric and idempotent matrix: $\mathbf{H}\mathbf{H} = \mathbf{H}$

$\mathbf{H}$ projects $\mathbf{y}_k$ onto the column space of $\mathbf{X}$ for $k \in \{1, \ldots, m\}$.

References

Text Book:

“Business Analytics, The Science of Data-Driven Decision Making”, U. Dinesh Kumar, Wiley 2017 (Ch 10.1-10.19.1)

Additional reference (for the interested student)

http://users.stat.umn.edu/~helwig/notes/mvlr-Notes.pdf

References

http://users.stat.umn.edu/~helwig/notes/mvlr-Notes.pdf

THANK YOU

Dr.Mamatha H R
Professor,Department of Computer Science
mamathahr@pes.edu

+91 80 2672 1983 Extn 834