INTRODUCTION TO MULTIVARIATE ANALYSIS

BY

Prof. Gafar Matanmi Oyeyemi

Department of Statistics

University of Ilorin

INTRODUCTION TO MULTIVARIATE ANALYSIS

Multivariate Analysis is the study or exploration of not only the relationship which may exist among set of variables, but the inherent structure of such variables is also very important.

INTRODUCTION

Most of the Multivariate Analysis deals with any of these:

. Estimation of Parameters (with confidence set)

. Test of hypothesis for means, variance-covariances, correlation coefficients

. Dimension reduction of complex data (Big Data) without losing much information

. Sorting, clustering, classification, pattern recognition and the related techniques (Machine Learning)

MULTIVARIATE DATA AND ITS NOTATION

Suppose we observed p-variables on a sample of n items. Let X_{ij} be measurement obtained on the i^{th} item on the variable j^{th} where

$i = 1, 2, \ , \ , \ n$ and $j = 1, 2, 3, \ , \ , p$. The result of these measurements can be represented by the following data matrix.

$$X_{ij} = \begin{bmatrix} x_{11} & x_{12} & , & , & x_{1p} \\ x_{21} & x_{22} & , & , & x_{2p} \\ , & , & , & , & , \\ , & , & , & , & , \\ x_{n1} & x_{n2} & , & , & x_{np} \end{bmatrix}$$

Most often the data matrix is always denoted by X

MULTIVARIATE NORMAL DISTRIBUTION

Assume that $x_1, x_2, x_3, \ , \ , x_n$ are iid vectors of random sample of size n (n > p) observed from p - variables, the joint probability density function is given as;

$$f\left(x_1, , , x_p; \mu, \Sigma\right) = (2\pi)^{-\frac{p}{2}} \left|\Sigma\right|^{-\frac{1}{2}} exp - \frac{1}{2}\left(X - \mu\right)^I \Sigma^{-1}\left(X - \mu\right)$$

Where μ is vector of population means and Σ is a positive definite population variance-covariance matrix.

BIVARIATE NORMAL

The simplest form of Multivariate Normal distribution is a Bivariate Normal distribution.

$$X = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, \; \mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} \text{ and } \Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{bmatrix}$$

Given the pdf,

$$f\Big((x \,|\, \mu, \Sigma) \Big) = (2\pi)^{-\frac{p}{2}} |\Sigma|^{-\frac{1}{2}} exp - \frac{1}{2}(X - \mu)^{I} \Sigma^{-1} (X - \mu)$$

Where $|\Sigma| = \sigma_{11}\sigma_{22}(1 - \rho^2)$ and

$$\Sigma^{-1} = \frac{1}{\sigma_{11}\sigma_{22}(1 - \rho^2)} \begin{bmatrix} \sigma_{22} & -\sigma_{12} \\ -\sigma_{21} & \sigma_{11} \end{bmatrix},$$

$\rho = \dfrac{\sigma_{12}}{\sigma_1 \sigma_2}$ and p = 2.

BIVARIATE NORMAL

$$f(x_1, x_2) = (2\pi)^{-\frac{2}{2}}(\sigma_{11}\sigma_{22}(1 - \rho^2)^{-\frac{1}{2}} exp - \frac{1}{2}\begin{pmatrix} x_1 & - & \mu_1 \\ x_2 & - & \mu_2 \end{pmatrix}^I \frac{1}{\sigma_{11}\sigma_{22}(1 - \rho^2)}\begin{bmatrix} \sigma_{22} & -\sigma_{12} \\ -\sigma_{21} & \sigma_{11} \end{bmatrix}\begin{pmatrix} x_1 & - & \mu_1 \\ x_2 & - & \mu_2 \end{pmatrix}$$

$$f(x_1, x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1 - \rho^2}} exp - \frac{1}{2(1 - \rho^2)}\left\{\left(\frac{x_{1} - \mu_1}{\sigma_1}\right)^2 + \left(\frac{x_{2} - \mu_2}{\sigma_2}\right)^2 - 2\rho\left(\frac{x_{1} - \mu_1}{\sigma_1}\right)\left(\frac{x_{2} - \mu_2}{\sigma_2}\right)\right\}$$

If the two variables x_1 and x_2 are independent, $\sigma_{12} = \sigma_{21} = 0$, therefore $\rho = 0$

$$f(x_1, x_2) = \frac{1}{2\pi\sigma_1\sigma_2} exp - \frac{1}{2}\left\{\left(\frac{x_{1} - \mu_1}{\sigma_1}\right)^2 + \left(\frac{x_{2} - \mu_2}{\sigma_2}\right)^2\right\}$$

MEAN AND VARIANCE-COVARIANCE MATRIX

The two parameters of Multivariate Normal Distribution are the μ and Σ. The μ is a vector containing the population mean of each variable while Σ is a p x p variance-covariance matrix containing the variances of all the variables and their covariances.

$$\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \\ , \\ , \\ \mu_p \end{bmatrix} \qquad \Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{21} & , & , & \sigma_{p1} \\ \sigma_{12} & \sigma_{22} & , & , & \sigma_{p2} \\ , & , & , & , & , \\ , & , & , & , & , \\ \sigma_{1p} & \sigma_{2p} & , & , & \sigma_{pp} \end{bmatrix}$$

TYPES OF MULTIVARIATE NORMAL

The variance-covariance matrix (Σ) determines the type of multivariate density function. If the variance-covariance matrix is a diagonal matrix, that is $\Sigma = \sigma^2 I$ where I is an identity matrix of order p.

$$\sigma^2 I = \begin{bmatrix} \sigma^2 & 0 & , & , & 0 \\ 0 & \sigma^2 & , & , & 0 \\ & & , & , & , \\ 0, & 0, & , & \sigma^2 & , \\ 0 & 0 & , & , & \sigma^2 \end{bmatrix}$$

Prof. Gafar Matanmi Oyeyemi, Department of Statistics, University of Ilorin.
matanmi@unilorin.edu.ng

TYPES OF MULTIVARIATE NORMAL

Then the joint probability density function becomes:

$$f(x/\mu,\ \sigma^2 I) = (2\pi)^{-\frac{p}{2}} \left|\sigma^2 I\right|^{-\frac{1}{2}} exp - \frac{1}{2}(x\ -\ \mu)^I(\sigma^2 I)^{-1}(x\ -\ \mu)$$

The above pdf is called **Independent Multivariate Normal.** The variables in the multivariate normal are pair-wise independent.

TYPES OF MULTIVARIATE NORMAL

Also, if the variables are standardized, such that, $z_i = \dfrac{x_i - \mu_i}{\sigma_i}$,

where the $E[z_i] = 0$ and $V[z_i] = 1$

$$\text{Then the } \mu = \begin{bmatrix} 0 \\ 0 \\ ' \\ ' \\ 0 \end{bmatrix} = \underline{0} \text{ and } \Sigma = I = \begin{bmatrix} 1 & 0 & , & , & 0 \\ 0 & 1 & , & , & 0 \\ 0\ , & 0\ , & , & , & , \\ & & , & 1 & , \\ 0 & 0 & , & , & 1 \end{bmatrix}$$

Prof. Gafar Matanmi Oyeyemi, Department of Statistics, University of Ilorin.
matanmi@unilorin.edu.ng

TYPES OF MULTIVARIATE NORMAL

The joint density function becomes:

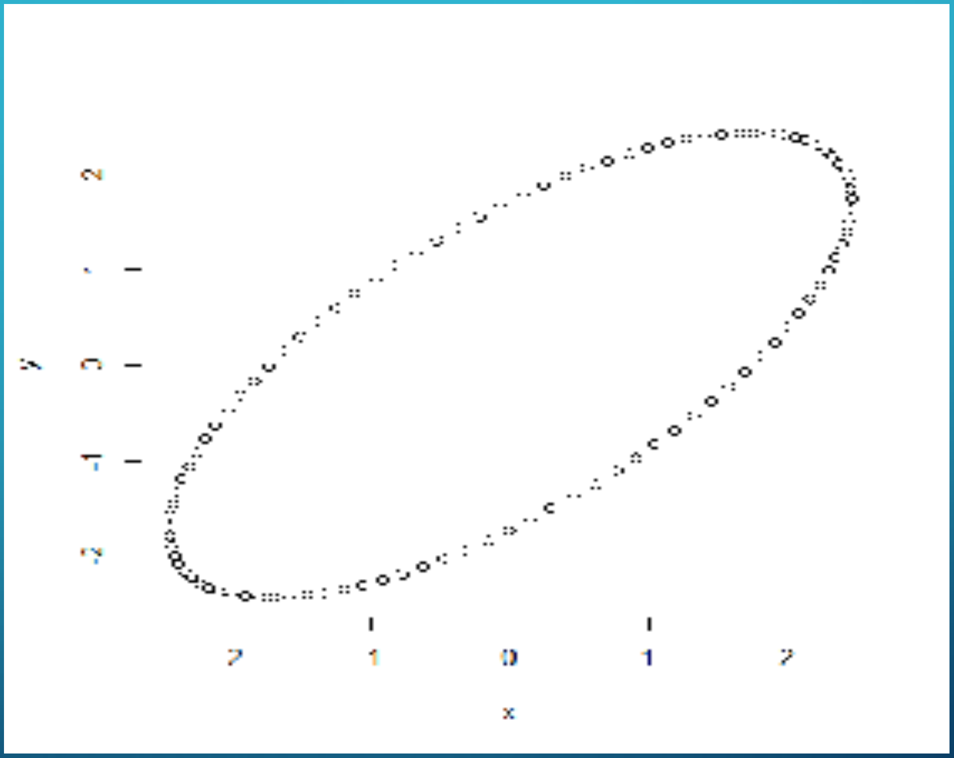
$$f(z/0,\ I) = (2\pi)^{-\frac{p}{2}}\left|I\right|^{-\frac{1}{2}} exp - \frac{1}{2} Z^I (I)^{-1} Z$$

$$= (2\pi)^{-\frac{p}{2}} exp - \frac{1}{2} Z^I Z$$

The above pdf is known as **Standardized Independent Multivariate Normal**. The variables are not only standardized but are also pair-wise independent.

BIVARIATE NORMAL ELLIPSOID

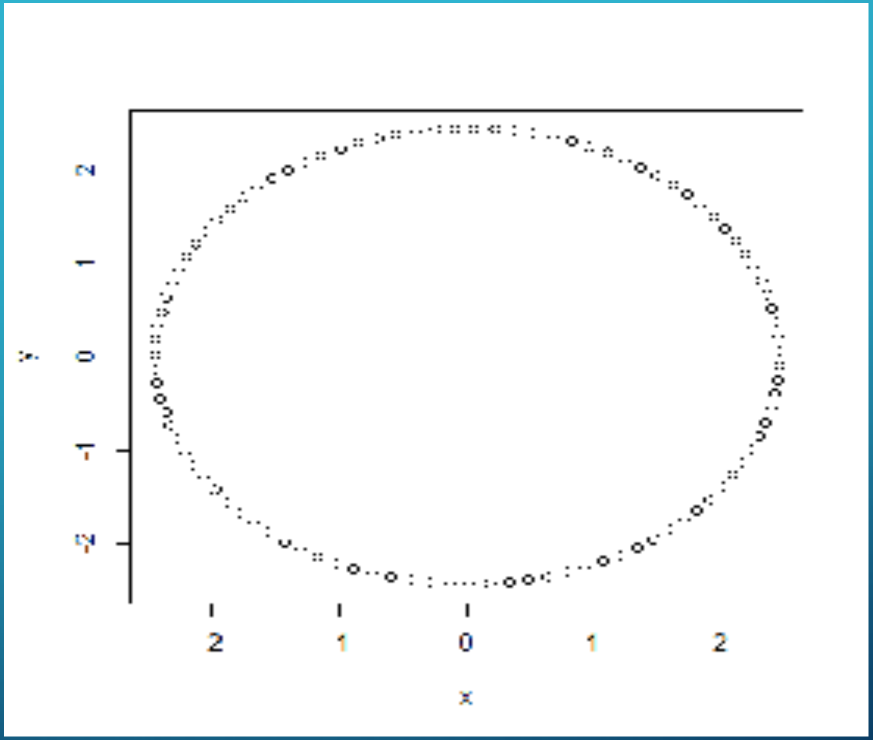
Bivariate Normal; $\Sigma = \begin{bmatrix} 2 & 2 \\ 2 & 4 \end{bmatrix}, \rho = 0.707$



Prof. Gafar Matanmi Oyeyemi, Department of Statistics, University of Ilorin.
matanmi@unilorin.edu.ng

BIVARIATE NORMAL ELLIPSOID

Independent Bivariate Normal; $\Sigma = \begin{bmatrix} 2 & 0 \\ 0 & 4 \end{bmatrix}, \rho = 0$



Prof. Gafar Matanmi Oyeyemi, Department of Statistics, University of Ilorin.
matanmi@unilorin.edu.ng

MARGINAL DISTRIBUTIONS OF MULTIVARIATE NORMAL

The marginal distribution of any random variable in the P-variate normal is the simple (univariate) normal distribution.

$$f(x;\ \mu,\ \sigma^2) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{1}{2\sigma^2}(x\ -\ \mu)^2}$$

Generally, if random variables $x_1,\ x_2,\ x_3,\ \ ,\ \ ,\ x_p$ are jointly normal (Multivariate normal), the joint marginal distribution of any subset of s variables (s < p) is the s-variate normal distribution.

MARGINAL DISTRIBUTIONS OF MULTIVARIATE NORMAL

Let X_3 be a trivariate normal; $X \sim N(\mu, \Sigma)$,

$$X \sim N\left(\begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{bmatrix}, \begin{bmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} \\ \sigma_{21} & \sigma_{22} & \sigma_{23} \\ \sigma_{31} & \sigma_{32} & \sigma_{33} \end{bmatrix} \right)$$

The marginal distribution of X_1 is $x_1 \sim N(\mu_1, \sigma_{11})$, while the joint marginal distribution of X_1 and X_3 is given by;

$$\begin{bmatrix} X_1 \\ X_3 \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_1 \\ \mu_3 \end{bmatrix}, \begin{bmatrix} \sigma_{11} & \sigma_{13} \\ \sigma_{31} & \sigma_{33} \end{bmatrix} \right)$$

CONDITIONAL DISTRIBUTION AND ITS EXPECTATION

Suppose the p-vector of random variables X from a multivariate normal, is partitioned into two vectors; $X = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}$, the conditional density function of X_1 given that the elements of X_2 are fixed is denoted by:

$$h\left(X_1 \middle| X_2 = x_2\right) = \frac{f\left(x_1,\ x_2\right)}{g\left(x_2\right)}$$

$f\left(x_1,\ x_2\right)$ is the joint density of x_1 *and* x_2 while $g\left(x_2\right)$ is the marginal density function of x_2.

CONDITIONAL DISTRIBUTION AND ITS EXPECTATION

Skipping the proof:

$$h\left(X_1 \middle| X_2 = x_2\right) = \frac{1}{(2\pi)^{\frac{p}{2}}\left|\Sigma_{1.2}\right|^{\frac{1}{2}}} exp - \frac{1}{2}\left[\left(x_1 - \mu_1\right) - \Sigma_{12}\Sigma_{22}^{-1}\left(x_2 - \mu_2\right)\right]^1 \Sigma_{1.2}^{-1}\left[\left(x_1 - \mu_1\right) - \Sigma_{12}\Sigma_{22}^{-1}\left(x_2 - \mu_2\right)\right]$$

$$= \frac{1}{(2\pi)^{\frac{p}{2}}\left|\Sigma_{1.2}\right|^{\frac{1}{2}}} exp - \frac{1}{2}\left[x_1 - \left(\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}\left(x_2 - \mu_2\right)\right)\right]^1 \Sigma_{1.2}^{-1}\left[x_1 - \left(\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}\left(x_2 - \mu_2\right)\right)\right]$$

From the above conditional density function, $h\left(X_1 \middle| X_2 = x_2\right)$

$E\left[h\left(X_1 \middle| X_2 = x_2\right)\right] = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}\left(x_2 - \mu_2\right)$ while the

$$\mathrm{V}\left[h\left(X_1 \middle| X_2 = x_2\right)\right] = \Sigma_{1.2}$$

Σ_{12} is the variance-covarian matrix of vectors of variables X_1 and X_2

Σ_{22} is the variance-covariance matrix of vector X_2

CONDITIONAL DISTRIBUTION AND ITS EXPECTATION

The $E\left[h\left(X_1 \middle| X_2 = x_2\right)\right] = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}\left(x_2 - \mu_2\right)$, given that X_1 and X_2 are (q x 1) and (r x 1) vectors respectively. The expectation will be a multivariate linear regression model obtained as follows:

$$= \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}\left(x_2 - \mu_2\right)$$

$$= \mu_1 - \Sigma_{12}\Sigma_{22}^{-1}\mu_2 + \Sigma_{12}\Sigma_{22}^{-1}x_2$$

$$= \beta_{0(qx1)} + \beta_{(qxr)}x_{2(rx1)}$$

$$= \begin{bmatrix} \beta_{01} \\ \beta_{02} \\ , \\ , \\ \beta_{0q} \end{bmatrix} + \begin{bmatrix} \beta_{11} & \beta_{12} & , & , & \beta_{1r} \\ \beta_{21} & \beta_{22} & , & , & \beta_{2r} \\ , & , & , & , & , \\ , & , & , & , & , \\ \beta_{q1} & \beta_{q2} & , & , & \beta_{qr} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ , \\ , \\ x_r \end{bmatrix}$$

MULTIVARIATE LINEAR REGRESSION MODEL

$$= \begin{bmatrix} \beta_{01} & \beta_{11} & \beta_{12} & , & \beta_{1r} \\ \beta_{02} & \beta_{21} & \beta_{22} & , & \beta_{2r} \\ \beta_{03} & \beta_{31} & \beta_{32} & , & \beta_{3r} \\ , & , & , & , & , \\ \beta_{0q} & \beta_{q2} & \beta_{q2} & , & \beta_{qr} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ , \\ , \\ x_r \end{bmatrix} = Bx$$

In the Multivariate Regression Model: $Y = X_1$ & $X = X_2$

$$Y_1 = \beta_{01} + \beta_{11}x_1 + \beta_{12}x_2 + , \quad , \quad , + \beta_{1r}x_r + e_1$$

$$Y_2 = \beta_{02} + \beta_{21}x_1 + \beta_{22}x_2 + , \quad , \quad , + \beta_{2r}x_r + e_2$$

$$Y_3 = \beta_{03} + \beta_{31}x_1 + \beta_{32}x_2 + , \quad , \quad , + \beta_{3r}x_r + e_3$$

$$, \quad , \quad , \quad , \quad , \quad , \quad , \quad , \quad , \quad , \quad ,$$

$$Y_q = \beta_{0q} + \beta_{q1}x_1 + \beta_{q2}x_2 + , \quad , \quad , + \beta_{rr}x_r + e_q$$

MULTIVARIATE LINEAR REGRESSION MODEL

In the Matrix Notation, the model is presented as;

$$Y = BX + \epsilon$$

The matrix B contains the regression coefficients of the multivariate regression models and it can be shown that

$$B = \left(X_2^1 X_2\right)^{-1} X_2^1 X_1 = \left(X^1 X\right)^{-1} X^1 Y$$

is the Maximum Likelihood Estimate (MLE) of the regression coefficients

MULTIVARIATE LINEAR REGRESSION MODEL

To test for the significance of the fitted model, is equivalent to testing that the matrix B is equal to a null matrix, that is;

$$H_0: B = 0 \qquad \text{vs} \quad H_1: B \neq 0; \text{ equivalent to}$$

$$H_0: \Sigma_{12}\Sigma_{22}^{-1} = 0 \text{ vs } H_1: \Sigma_{12}\Sigma_{22}^{-1} \neq 0; \; B = \Sigma_{12}\Sigma_{22}^{-1}$$

$$H_0: \Sigma_{12} = 0 \qquad \text{vs} \quad H_1: \Sigma_{12} \neq 0; \text{ since } \Sigma_{22}^{-1} \neq 0$$

MULTIVARIATE LINEAR REGRESSION MODEL

The test statistics:

$$\Lambda = \frac{\left| S_{11} - S_{12} S_{22}^{-1} S_{21} \right|}{\left| S_{11} \right|}$$

S_{11}, S_{12} and S_{21} are the sample estimates of Σ_{11}, Σ_{12} and Σ_{22} respectively. The test statistic, Wilks' Lambda, Λ, can be converted to a Chi-square distribution as follows;

$$X = -\left[\left(N - q - 1 \right) - \frac{1}{2} \left(q - r + 1 \right) \right] ln\Lambda \sim \chi^2_{qr}$$

The null hypothesis is rejected if the Xcal > $\chi^2_{(1-\alpha);qr}$ at level of significance alpha.

Prof. Gafar Matanmi Oyeyemi, Department of Statistics, University of Ilorin.
matanmi@unilorin.edu.ng

MULTIPLE LINEAR REGRESSION MODEL

Given that q = 1, that is X_1 is a single variable and X_2 is a vector of (r x 1) variables, the expectation will result into multiple linear regression model as follows:

$$E\left[h\left(X_1 \middle| X_2 = x_2\right)\right] = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}\left(x_2 - \mu_2\right)$$

$$= \mu_1 - \Sigma_{12}\Sigma_{22}^{-1}\mu_2 + \Sigma_{12}\Sigma_{22}^{-1}x_2$$

$$= \mu_1 - \sum_{i=1}^{r} \beta_i \mu_i + \sum_{i=1} \beta_i x_i$$

$$= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + , \quad , \quad + \beta_r x_r$$

SIMPLE LINEAR REGRESSION MODEL

Finally, if q = 1 and r =1, the resulting expectation will give simple linear regression model obtained as follow:

$$E\left[h\left(X_1 \middle| X_2 = x_2\right)\right] = \mu_1 + \sigma_{12}\sigma_{22}^{-1}\left(x_2 - \mu_2\right)$$

$$= \mu_1 - \frac{\sigma_{12}}{\sigma_{22}}\mu_2 + \frac{\sigma_{12}}{\sigma_{22}}x_2$$

$$= \mu_1 - \beta_1\mu_2 + \beta_1 x_2$$

$$= \beta_0 + \beta_1 x_2$$

SOME USEFUL RESULTS FOR MULTIVARIATE NORMAL

- ➢Each variable has a univariate normal distribution
- ➢Any subset of the variables also has a multivariate normal distribution.
- ➢Any linear combination of the variables has a univariate normal distribution.
- ➢The conditional distribution for a subset of the variables conditional on known values for another subset of variables is a multivariate distribution

MULTIVARIATE LINEAR REGRESSION MODEL

Example (Using stock data in STATA)

There are five variables; x_1 = volume; x_2 = close price, x_3 = open price, x_4 = high price and x_5 = low price: The estimates of the population parameters: μ and Σ are $\bar{X}$ and S respectively.

$$\bar{X} = \begin{bmatrix} 12320.68 \\ 1194.179 \\ 1194.884 \\ 1204.044 \\ 1183.334 \end{bmatrix}$$

$$S = \begin{bmatrix} 6687028.7 & -69719 & -71935.4 & -65353 & -77493.5 \\ -69719 & 7533.32 & 7434.47 & 7463.59 & 7541.77 \\ -71935.4 & 7434.47 & 7591.3 & 7492.74 & 7561.28 \\ -65353 & 7463.59 & 7492.74 & 7488.25 & 7521.26 \\ -77493.5 & 7541.77 & 7561.28 & 7521.26 & 7652.01 \end{bmatrix}$$

MULTIVARIATE LINEAR REGRESSION MODEL

Multivariate regression model (q = 2 and r =3): $X_1 = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}$ and $X_2 = \begin{bmatrix} X_3 \\ X_4 \\ X_5 \end{bmatrix}$

$$E\left[h\left(X_1 \middle| X_2 = x_2\right)\right] = \bar{x}_1 - S_{12}S_{22}^{-1}\bar{x}_2 + S_{12}S_{22}^{-1}x_2$$

Where $\bar{x}_1 = \begin{bmatrix} \bar{x}_1 \\ \bar{x}_2 \end{bmatrix} = \begin{bmatrix} 12320.68 \\ 1194.179 \end{bmatrix}$, $\bar{x}_2 = \begin{bmatrix} \bar{x}_3 \\ \bar{x}_4 \\ \bar{x}_5 \end{bmatrix} = \begin{bmatrix} 1194.884 \\ 1204.044 \\ 1183.334 \end{bmatrix}$

$$S_{12} = \begin{bmatrix} s_{13} & s_{14} & s_{15} \\ s_{23} & s_{24} & s_{25} \end{bmatrix} = \begin{bmatrix} -71935.4 & -65353 & -77493.5 \\ 7434.47 & 7463.59 & 7541.77 \end{bmatrix}$$

$$S_{22} = \begin{bmatrix} s_{33} & s_{34} & s_{35} \\ s_{43} & s_{44} & s_{45} \\ s_{53} & s_{54} & s_{55} \end{bmatrix} = \begin{bmatrix} 7591.3 & 7492.74 & 7561.28 \\ 7492.74 & 7488.25 & 752126 \\ 7561.28 & 7521.26 & 7652.01 \end{bmatrix}$$

MULTIVARIATE LINEAR REGRESSION MODEL

$$E\left[h\left(X_1 \middle| X_2 = x_2\right)\right] =$$

$$\begin{bmatrix} 12320.68 \\ 1194.179 \end{bmatrix} + \begin{bmatrix} -71935.4 & -65353 & -77493.5 \\ 7434.47 & 7463.59 & 7541.77 \end{bmatrix} \begin{bmatrix} 7591.3 & 7492.74 & 7561.28 \\ 7492.74 & 7488.25 & 752126 \\ 7561.28 & 7521.26 & 7652.01 \end{bmatrix}^{-1} \left(\begin{bmatrix} x_3 \\ x_4 \\ x_5 \end{bmatrix} - \left[\begin{bmatrix} 1194.884 \\ 1204.044 \\ 1183.334 \end{bmatrix} \right] \right)$$

$$E\left[h\left(X_1 \middle| X_3 = x_3, X_4 = x_4,\ X_5 = x_5\right)\right] =$$

$$\begin{bmatrix} X_1 \\ X_2 \end{bmatrix} = \begin{matrix} 19657.7116 - 27.4892x_3 + 130.7129x_4 - 111.4434x_5 \\ 5.9147 - 0.6170x_3 + 0.9226x_4 + 0.6885x_5 \end{matrix}$$

MULTIPLE LINEAR REGRESSION MODEL

Multiple regression of x_1 (volume) on x_3 (open), x_4 (high) and x_5 (low)

$$E\left[h\left(X_1 \middle| X_3 = x_3, X_4 = x_4, X_5 = x_5\right)\right] = 12320.68 - \begin{bmatrix} -71935.4 & -65353 & -77493.5 \end{bmatrix} \begin{bmatrix} 7591.3 & 7492.74 & 7561.28 \\ 7492.74 & 7488.25 & 752126 \\ 7561.28 & 7521.26 & 7652.01 \end{bmatrix}^{-1} \left(\begin{bmatrix} x_3 \\ x_4 \\ x_5 \end{bmatrix} - \left[\begin{bmatrix} 1194.884 \\ 1204.044 \\ 1183.334 \end{bmatrix} \right] \right)$$

$$E\left[\left(X_1 \middle| X_3 = x_3, X_4 = x_4,\ X_5 = x_5\right)\right] = 19657.711 - 27.4892x_3 + 130.7130x_4 - 111.4434x_5$$

Prof. Gafar Matanmi Oyeyemi, Department of Statistics, University of Ilorin.
matanmi@unilorin.edu.ng

CONDITIONAL DISTRIBUTION AND ITS EXPECTATION

For simple linear regression; regression of x_1 (volume) on x_3 (open)

$$E\left[h\left(X_1 \middle| X_3 = x_3\right)\right] = \mu_1 + \sigma_{13}\sigma_{33}^{-1}\left(x_3 - \mu_3\right)$$

$$= 12320.68 + (-71935.4)\left(7591.3\right)^{-1}(x_3 - 1194.884)$$

$$= 23643.44 - 9.4760x_3$$

MULTIVARIATE ANALYSIS OF VARIANCE (MANOVA)

Multivariate Analysis of Variance is extension of Analysis of Variance (ANOVA), it generalizes ANOVA to allow for multiple/multivariate responses.

The model can be written as:

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij} = \mu_i + \varepsilon_{ij}, \text{ where } y_{ij} \sim N_p(\mu_i, \quad \Sigma)$$

The model can be written in vector/matrix form as follows:

$$\begin{bmatrix} y_{ij1} \\ y_{ij2} \\ , \\ , \\ , \\ y_{ijr} \end{bmatrix} = \begin{bmatrix} \mu_{i1} \\ \mu_{i2} \\ , \\ , \\ , \\ \mu_{ir} \end{bmatrix} + \begin{bmatrix} \varepsilon_{ij1} \\ \varepsilon_{ij2} \\ , \\ , \\ , \\ \varepsilon_{ijr} \end{bmatrix}$$

The null and alternative hypotheses are:

$$H_0: \underline{\mu}_1 = \underline{\mu}_2 = , \quad , \quad , \quad = \underline{\mu}_k$$ against

$H_1: \underline{\mu}_i \neq \underline{\mu}_j$ for at least one pair $i \neq j$

MULTIVARIATE ANALYSIS OF VARIANCE (MANOVA)

Data layout

	Sample 1	Sample 2	Sample 3	. . .	Sample k
	$N_p(\mu_1, \Sigma)$	$N_p(\mu_2, \Sigma)$	$N_p(\mu_3, \Sigma)$		$N_p(\mu_k, \Sigma)$
	,	,	, , , ,		,
	,	,	, , , ,		,
	,	,	, , , ,		,
Total					
Mean					

MULTIVARIATE ANALYSIS OF VARIANCE (MANOVA)

The MANOVA Table to test the hypothesis

Source	df	Sum of Square Matrices
Group	k-1	
Error	N-k	
Total	**N-1**	

Prof. Gafar Matanmi Oyeyemi, Department of Statistics, University of Ilorin.
matanmi@unilorin.edu.ng

MULTIVARIATE ANALYSIS OF VARIANCE (MANOVA)

The B and W matrices are both p x p, but not necessarily full rank. The rank of B is min(p, V_B), where V_B is the degree of freedom associated with hypothesis, that is, k – 1.

Using Wilk's Λ Test Statistic:

$$\Lambda = \frac{|W|}{|W + B|}$$

The null hypothesis is rejected if $\Lambda > \Lambda_{\alpha,\ p,\ V_B, V_W}$, where V_B and V_W are degrees of freedom for the hypothesis (k – 1) and Error (N – k) respectively.

MULTIVARIATE ANALYSIS OF VARIANCE (MANOVA)

This test statistic can be converted to an F-statistic.

$$F = \frac{1-\Lambda}{\Lambda}\frac{V_W - p + 1}{p} \sim F_{p,\ V_w - p + 1}$$

The null hypothesis is rejected if F $> \mathrm{F}_{\alpha;\ p,\ V_w-p+1}$

If the null hypothesis is rejected, the follow up tests could be made. Fixing $r \in \{1,\ 2,\ ,\ ,p\}$, one could test:

$$H_0:\ \mu_{1r} = \mu_{2r} = \ .\ \ .\ \ .\ = \mu_{kr},$$

this is a univariate ANOVA test to see if the k populations differ on variable r.

MULTIVARIATE ANALYSIS OF VARIANCE (MANOVA)

Some properties of Wilks' Λ

- The $V_w = N - k \geq p$ for the determinants to be positive
- The degree of for error and hypothesis (group) are the same for equivalent univariate ANOVA
- Λ is in the interval [0, 1]
- Increasing the number of variables p, decrease the critical value for Λ needed to reject the null hypothesis.
- When V_B = 1, 2 or p = 1, 2, Wilks' Λ is equivalent to an F statistic.

MULTIVARIATE ANALYSIS OF VARIANCE (MANOVA)

Other alternative statistics to Wilks' Λ are:

- Hotelling's Trace statistic, $tr\left(W^{-1}B\right)$
- Pillai's Trace statistic; $tr[\left(W+B)^{-1}B\right)$
- Roy's largest root: $\dfrac{\lambda_1}{1+\lambda_1}$, λ_1 is the largest eigenvalue of

  $W^{-1}B$.

MULTIVARIATE ANALYSIS OF VARIANCE (MANOVA)

Two Groups (Populations)

$H_0\colon \underline{\mu}_1 = \underline{\mu}_2$ against $H_1\colon \underline{\mu}_1 \neq \underline{\mu}_2$

The test statistic is given as follows:

$$T = \frac{N_1 N_2}{N_1 + N_2}\left(\bar{y}_{1.} - \bar{y}_{2.}\right)^1 S_p^{-1}\left(\bar{y}_{1.} - \bar{y}_{2.}\right) \sim T_{p,\ N_1+N_2-2}$$

Where the pooled variance-covariance matrix,

$S_p = \dfrac{(N_1 - 1)S_1 + (N_1 - 1)S_2}{N_1 + N_2 - 2}$, S_1 and S_2 are sample variance-covariance matrices for groups (populations) 1 and 2 respectively.

Converting it to F- statistic: $F = \dfrac{N_1 + N_2 - p - 1}{p(N_1 + N_2 - 2)} * T$ and the null hypothesis will be rejected if $F > F_{\alpha;\, p,\ N_1+N_2-p-1}$

MULTIVARIATE ANALYSIS OF VARIANCE (MANOVA)

Example

$N_1=5$ $N_2 = 4$ $N_3 = 3$ $N_4 = 6$

Prof. Gafar Matanmi Oyeyemi, Department of Statistics, University of Ilorin.
matanmi@unilorin.edu.ng

MULTIVARIATE ANALYSIS OF VARIANCE (MANOVA)

The Hypothesis:

$$H_0: \underline{\mu}_1 = \underline{\mu}_2 = \underline{\mu}_3 = \underline{\mu}_4 \text{ against } H_1: \underline{\mu}_i \neq \underline{\mu}_j \text{ for at least one pair}$$

$i \neq j$

$$\bar{y}_{1.} = [10.20 \quad 3.50 \quad 1.68];$$

$$\bar{y}_{2.} = [8.63 \quad 3.00 \quad 1.73];$$

$$\bar{y}_{3.} = [9.43 \quad 2.90 \quad 1.63]$$

$$\bar{y}_{4.} = [8.90 \quad 3.47 \quad 1.65] \text{ and}$$

$$\bar{y}_{..} = [9.29 \quad 3.28 \quad 1.67]$$

MULTIVARIATE ANALYSIS OF VARIANCE (MANOVA)

$$B = \sum_{i=1}^{4} N_i(\bar{y}_{i.} - \bar{y}_{...})(\bar{y}_{i.} - \bar{y}_{..})^1 = 5\begin{bmatrix} 10.20 - & 9.29 \\ 3.50 & -3.28 \\ 1.68 - & 1.67 \end{bmatrix}\begin{bmatrix} 10.20 - & 9.29 \\ 3.50 & -3.28 \\ 1.68 - & 1.67 \end{bmatrix}^1 + 4\begin{bmatrix} 8.63 - & 9.29 \\ 3.00 & -3.28 \\ 1.73 - & 1.67 \end{bmatrix}\begin{bmatrix} 8.63 - & 9.29 \\ 3.00 & -3.28 \\ 1.73 - & 1.67 \end{bmatrix}^1 + 3\begin{bmatrix} 9.43 - & 9.29 \\ 2.90 & -3.28 \\ 1.63 - & 1.67 \end{bmatrix}\begin{bmatrix} 9.43 - & 9.29 \\ 2.90 & -3.28 \\ 1.63 - & 1.67 \end{bmatrix}^1 + 6\begin{bmatrix} 8.90 - & 9.29 \\ 3.47 & -3.28 \\ 1.65 - & 1.67 \end{bmatrix}\begin{bmatrix} 8.90 - & 9.29 \\ 3.47 & -3.28 \\ 1.65 - & 1.67 \end{bmatrix}^1$$

$$B = \begin{bmatrix} 6.4299 & 1.3659 & -0.0956 \\ 1.3659 & 1.0907 & -0.0185 \\ -0.0956 & -0.0185 & 0.0175 \end{bmatrix}$$

MULTIVARIATE ANALYSIS OF VARIANCE (MANOVA)

$$T=\sum_{i=1}^{k}\sum_{j}^{N_i}\left(y_{ij}-\bar{y}_{..}\right)\left(y_{ij}-\bar{y}_{..}\right)^1=\begin{bmatrix}10.5- & 9.29\\ 3.0 & -3.28\\ 1.5- & 1.67\end{bmatrix}\begin{bmatrix}10.5- & 9.29\\ 3.0 & -3.28\\ 1.5- & 1.67\end{bmatrix}^1+\begin{bmatrix}7.0- & 9.29\\ 3.5 & -3.28\\ 1.7- & 1.67\end{bmatrix}\begin{bmatrix}7.0- & 9.29\\ 3.5 & -3.28\\ 1.7- & 1.67\end{bmatrix}^1+\;.\;\;.\;\;.+\begin{bmatrix}8.5- & 9.29\\ 4.5 & -3.28\\ 1.4- & 1.67\end{bmatrix}\begin{bmatrix}8.5- & 9.29\\ 4.5 & -3.28\\ 1.4- & 1.67\end{bmatrix}^1$$

$$T=\begin{bmatrix}44.0978 & 0.8656 & 0.5344\\ 0.8656 & 4.4911 & -0.2911\\ 0.5344 & -0.2911 & 0.3361\end{bmatrix}$$

MULTIVARIATE ANALYSIS OF VARIANCE (MANOVA)

$$W = T - B = \begin{bmatrix} 37.6679 & -0.5003 & 0.6300 \\ -0.5003 & 3.4004 & -0.2726 \\ 0.6300 & -0.2726 & 0.3186 \end{bmatrix}$$

$$\Lambda = \frac{|W|}{|W+B|} = \frac{\begin{vmatrix} 37.6679 & -0.5003 & 0.6300 \\ -0.5003 & 3.4004 & -0.2726 \\ 0.6300 & -0.2726 & 0.3186 \end{vmatrix}}{\begin{vmatrix} 44.0978 & 0.8656 & 0.5344 \\ 0.8656 & 4.4911 & -0.2911 \\ 0.5344 & -0.2911 & 0.3361 \end{vmatrix}} = 0.6023 \text{ Converting to F gives}$$

$$F = \frac{1-\Lambda}{\Lambda}\frac{V_W - p + 1}{p} = \frac{1-0.6023}{0.6023}\frac{14 - 3 + 1}{3} = 2.6412$$

$$\mathrm{F}_{1-\alpha,\ p,\ V_w - p + 1} = \mathrm{F}_{0.95;\ 3,\ 12} = 3.490$$

MULTIVARIATE ANALYSIS OF VARIANCE (MANOVA)

Two Populations (Comparing vector of means for two populations)

The hypothesis:

$$H_0: \underline{\mu}_1 = \underline{\mu}_2 \text{ against } H_1: \underline{\mu}_1 \neq \underline{\mu}_2$$

Using the above data but restricting ourselves to just populations 1 and 2.

The test statistics:

$$T = \frac{N_1 N_2}{N_1 + N_2}\left(\bar{y}_{1.} - \bar{y}_{2.}\right)^1 S_p^{-1}\left(\bar{y}_{1.} - \bar{y}_{2.}\right)$$

$$S_1 = \begin{bmatrix} 3.4500 & 0.1250 & -0.0075 \\ & 0.1250 & 0.1250 \\ & & 0.0170 \end{bmatrix} \text{ and } S_2 = \begin{bmatrix} 2.3958 & 0.0000 & 0.0125 \\ & 0.000 & 0.0000 \\ & & 0.0158 \end{bmatrix}$$

MULTIVARIATE ANALYSIS OF VARIANCE (MANOVA)

Therefore $S_p = \dfrac{(5-1)S_1 + (4-1)S_2}{5+4-2} = \begin{bmatrix} 2.9982 & 0.0714 & 0.0011 \\ & 0.0714 & 0.0071 \\ & & 0.0165 \end{bmatrix}$

$$T = \frac{5*4}{5+4}\begin{pmatrix} 10.200 - & 8.625 \\ 3.500 & -3.00 \\ 1.680 - & 1.725 \end{pmatrix}^{1} \begin{bmatrix} 2.9982 & 0.0714 & 0.0011 \\ & 0.0714 & 0.0071 \\ & & 0.0165 \end{bmatrix}^{-1} \begin{pmatrix} 10.200 - & 8.625 \\ 3.500 & -3.00 \\ 1.680 - & 1.725 \end{pmatrix} = 9.864$$

$$F = \frac{5+4-3-1}{3(5+4-2)} * 9.864 = 2.349; \qquad F_{0.95;\,3,\,5} = 5.409$$

We fail to reject the null hypothesis since F_{cal} (2.349) < F_{tab} (5.409) at 0.05 level of significance.

LINEAR DISCRIMINANT ANALYSIS FOR TWO CLASSES

We assume that the 2 populations have the same variance-covariance matrix Σ and with distinct mean vectors $\underline{\mu}_1$ and $\underline{\mu}_2$. Assume we have p-dimensional training data set X_{p1} of N_1, of which belong to ω_1 and X_{p2} of N_2 from ω_2.

$$X_{p1} = \begin{bmatrix} x_{11} \\ x_{12} \\ , \\ , \\ , \\ x_{1N_1} \end{bmatrix} \qquad X_{p2} = \begin{bmatrix} x_{21} \\ x_{22} \\ , \\ , \\ , \\ x_{2N_2} \end{bmatrix}$$

LINEAR DISCRIMINANT ANALYSIS FOR TWO CLASSES

Population 1	Population 2
$N_p(\mu_1, \Sigma)$	$N_p(\mu_2, \Sigma)$

,	,
,	,
,	,

Prof. Gafar Matanmi Oyeyemi, Department of Statistics, University of Ilorin.
matanmi@unilorin.edu.ng

LINEAR DISCRIMINANT ANALYSIS FOR TWO CLASSES

We can obtain a unit vector which forms a linear combination of the p-variables that best discriminates between the two populations.

The Fisher Linear Discriminant function is the linear combination of the p variables (a^1x) that maximizes the distance between the two classes projected mean vectors normalized by the within-class covariance matrix of the projected samples:

$$\max_{J(a)} = \frac{(\mu_1 - \mu_2)}{\Sigma_1 + \Sigma_2}$$

LINEAR DISCRIMINANT ANALYSIS FOR TWO CLASSES

With assumption of same variance-covariance matrix, the pooled sample variance covariance is used.

$$\max_{J(a)} = \frac{(\overline{x}_1 - \overline{x}_2)}{S_p} = a^1 = s_p^{-1}(\overline{x}_1 - \overline{x}_2), \text{ where } S_p = \frac{(N_1 - 1)S_1 + (N_2 - 1)S_2}{N_1 + N_2 - 2}$$

The linear combination z = a^1x is therefore used to classify the observed vectors into class ω_1 or ω_2 given the cut-off or classification rule.

The cut-off = $\frac{1}{2}(\overline{z}_1 + \overline{z}_2)$, where $\overline{z}_i = a^1\overline{x}_{pi}$, i = 1 or 2

Therefore, an observe vector x_j is assign to ω_1 if a^1x_j is greater than the cut-off and assign to ω_2 if otherwise.

LINEAR DISCRIMINANT ANALYSIS FOR TWO CLASSES

Example

Sample 1	Sample 2
[	
[33, 60]	[35, 57]
[36, 61]	[36, 59]
[35, 64]	[38, 59]
[38, 63]	[39, 61]
[40, 65]	[42, 63]
	[43, 65]
	[41, 59]

LINEAR DISCRIMINANT ANALYSIS FOR TWO CLASSES

$$\bar{x}_1 = \begin{bmatrix} 36.40 \\ 62.60 \end{bmatrix} \qquad \bar{x}_2 = \begin{bmatrix} 39.00 \\ 60.40 \end{bmatrix} \qquad S_1 = \begin{bmatrix} 7.30 & 4.20 \\ & 4.30 \end{bmatrix} \; S_1 = \begin{bmatrix} 8.33 & 6.67 \\ & 7.62 \end{bmatrix}$$

$$S_p = \frac{(5-1)S_1 + (7-1)S_2}{5+7-2} = \begin{bmatrix} 7.92 & 5.68 \\ & 6.29 \end{bmatrix}$$

$$a^1 = s_p^{-1}(\bar{x}_1 - \bar{x}_2) = \begin{bmatrix} 7.92 & 5.68 \\ & 6.29 \end{bmatrix}^{-1} \begin{bmatrix} 36.40 - 39.00 \\ 62.60 - 60.40 \end{bmatrix} = \begin{bmatrix} -1.6334 \\ 1.8198 \end{bmatrix}$$

$$z = a^1 x = -1.6334 x_{1j} + 1.8198 x_{2j}$$

LINEAR DISCRIMINANT ANALYSIS FOR TWO CLASSES

The cut-off $= \frac{1}{2}(\bar{z}_1 + \bar{z}_2) = \frac{1}{2}a^1(\bar{x}_1 + \bar{x}_2)$

$$= 0.5[-1.6334 + 1.8198]\begin{bmatrix} 36.40 + 39.00 \\ 62.60 + 60.40 \end{bmatrix} = 50.364$$

Therefore, an observed vector, x_j, is classified into ω_1 if $z = a^1 x_j$ is greater than 50.364 and into ω_2 if otherwise.

LINEAR DISCRIMINANT ANALYSIS FOR TWO CLASSES

Classify the twelve observed vectors used in obtaining the discriminant function into ω_1 or ω_2 using linear combination,

$$z = a^1 x = -1.6334 x_{1j} + 1.8198 x_{2j}$$

True Pop	Classified		Total
	1	2	
1	5 (100%)	0 (0%)	5
2	0 (0)	7 (100%)	7
Total	5	7	12

PRACTICAL WITH STATA

. Practical session

. We will have some illustrations of the techniques discussed using STATA Package

THANKS FOR YOUR ATTENTION