

## CHAPTER 1

### Introduction to multivariate distributions-5

#### 1.1

#### 1.2

#### 1.3

#### 1.4

#### 1.5 Multivariate linear models

The linear model in univariate analysis arises when the response and the covariates (can be several in number) are related through a linear function of some parameters. This models can be easily extended to use multivariate responses. In general, when the covariate variables are categorical, we deal with Multivariate Analysis of Variance (MANOVA). When they are quantitative, the multivariate regression is needed. The errors attached to different response variables need not be independent, and we can assume them to be correlated. In this section, we formulate simple examples of both kind, and derive the likelihood ratio methods.

### 1.5.1 MANOVA

We assume that there is one categorical explanatory variable, which will be called the treatment. We want to investigate problems involving the comparison of several mean vectors, namely the mean vectors corresponding to the response variable under different treatments. The comparisons of means emanate from mainly designed experiments. The Multivariate Analysis of Variance arises from the scenario where there is more than one variable measured per plot in the design of an experiment. The simplest among such problems is the multivariate one-way classification.

Suppose there are  $k$  treatments assigned randomly to plots. There are  $n_j$  plots receiving the  $j$ th treatment and  $\mathbf{x}_{ij}$  is the  $p$ -dimension yield vector for the  $i$ th plot receiving the  $j$ th treatment. The model is,

$$\mathbf{x}_{ij} = \boldsymbol{\mu} + \boldsymbol{\tau}_j + \boldsymbol{\epsilon}_{ij}$$

where  $\boldsymbol{\epsilon}_{ij}$  are independent  $N(0, \Sigma)$  errors,  $\boldsymbol{\mu}$  is the general effect, and  $\boldsymbol{\tau}_j$  is the effect for  $j$ th treatment. Let  $n = \sum_{i=1}^k n_j$ , the total number of plots. The goal is to test the null hypothesis  $H_0 : \tau_1 = \dots = \tau_k$ . To add identifiability to the problem, we also assume that  $\boldsymbol{\mu} = 0$ . Define, the group mean

$$\bar{\mathbf{x}}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} \mathbf{x}_{ij},$$

the grand mean

$$\bar{\mathbf{x}} = \frac{1}{n} \sum_{j=1}^k \sum_{i=1}^{n_j} \mathbf{x}_{ij},$$

the within group variance

$$\mathbf{W} = \sum_{j=1}^k \sum_{i=1}^{n_j} (\mathbf{x}_{ij} - \bar{\mathbf{x}}_j)(\mathbf{x}_{ij} - \bar{\mathbf{x}}_j)^t,$$

the between group variance

$$\mathbf{B} = \sum_{j=1}^k n_j (\bar{\mathbf{x}}_j - \bar{\mathbf{x}})(\bar{\mathbf{x}}_j - \bar{\mathbf{x}})^t,$$

and the total variance

$$\mathbf{T} = \sum_{j=1}^k \sum_{i=1}^{n_j} (\mathbf{x}_{ij} - \bar{\mathbf{x}})(\mathbf{x}_{ij} - \bar{\mathbf{x}})^t.$$

Then,  $\mathbf{T} = \mathbf{W} + \mathbf{B}$ ,  $\mathbf{W} \sim W_p(\Sigma, n - k)$ , and under  $H_0$ ,  $\mathbf{B} \sim W_p(\Sigma, k - 1)$ . Moreover,  $\mathbf{W}$  and  $\mathbf{B}$  are independent. The likelihood ratio statistics is  $\Lambda^{n/2}$  where

$$\Lambda = \frac{|\mathbf{W}|}{|\mathbf{T}|}$$

has a Wilk's lambda distribution with parameters  $p, n - k, k - 1$ . (If  $A \sim W_p(I, m)$ ,  $B \sim W_p(I, n)$  and they are mutually independent, then  $|A|/|A + B| = |I + A^{-1}B|^{-1}$  has a Wilk's lambda distribution with parameters  $p, m, n$ ). Further,  $\Lambda = \prod_{j=1}^p (1 + \lambda_j)^{-1}$  where  $\lambda_1, \dots, \lambda_p$  are eigenvalues of the matrix  $\mathbf{W}^{-1}\mathbf{B}$ . To carry out the test, there is also a transformation that leads to the  $F$  distribution and hence can be computed easily. [If  $U \sim \Lambda(p, m, 2)$ , then  $\frac{(m-p+1)(1-\sqrt{U})}{p\sqrt{U}} \sim F_{2p, 2(m-p+1)}$ .]

*Testing general contrasts :*

In the same flavor as in univariate one-way analysis, consider testing the null hypothesis  $H_0 : a_1\tau_1 + \dots + a_k\tau_k = 0$  where  $a_1 + \dots + a_k = 0$ . Then, the test statistics turns out to be

$$\Lambda = \frac{|\mathbf{W}|}{|\mathbf{W} + \mathbf{C}|} \sim \Lambda(p, n - k, 1)$$

where  $\mathbf{C} = (\sum_{j=1}^k a_j \bar{\mathbf{x}}_j)(\sum_{j=1}^k a_j \bar{\mathbf{x}}_j)^t / (\sum_{j=1}^k a_j^2 / n_j)$ . Similarly, we can also carry out two-way analysis, when there are more than one categorical covariates; analysis of covariance when there are quantitative as well as categorical variables and so on.

### 1.5.2 Multivariate regression

In this section, we consider the quantitative variables as covariates. Consider the problem of modeling the relationship between  $p$  responses  $Y_1, \dots, Y_p$  and a set of predictor variables  $X_1, \dots, X_r$ . Each response is assumed to follow

its own regression model

$$\begin{aligned} Y_1 &= \beta_{01} + \beta_{11}X_1 + \dots + \beta_{r1}X_r + u_1 \\ Y_2 &= \beta_{02} + \beta_{12}X_1 + \dots + \beta_{r2}X_r + u_2 \\ &\vdots \\ Y_p &= \beta_{0p} + \beta_{1p}X_1 + \dots + \beta_{rp}X_r + u_p \end{aligned}$$

The errors are assumed to have a normal distribution with  $\mathbb{E}(\mathbf{u}) = 0$ , and  $\text{var}(\mathbf{u}) = \Sigma$ . Thus, the error terms corresponding to different responses may be correlated. The data consists of  $n$  trials of the experiment where  $\mathbf{Y}_j$  and  $\mathbf{X}_j$  are the responses and the predictors for each trial. Writing the corresponding data matrices

$$\mathbf{Y} = \begin{bmatrix} \mathbf{Y}_1^t \\ \vdots \\ \mathbf{Y}_n^t \end{bmatrix} = \begin{bmatrix} Y_{11} \dots Y_{1p} \\ \vdots \\ Y_{n1} \dots Y_{np} \end{bmatrix}$$

and

$$\mathbf{X} = \begin{bmatrix} 1 & \mathbf{X}_1^t \\ \vdots & \\ 1 & \mathbf{X}_n^t \end{bmatrix} = \begin{bmatrix} 1 & X_{11} \dots X_{1r} \\ \vdots & \vdots \\ 1 & X_{n1} \dots X_{nr} \end{bmatrix}$$

we can write it as a regression model. As an extension of the univariate regression model, we have a model consisting of an observed matrix  $\mathbf{Y}_{n \times p}$  with  $p$  variables on  $n$  individual responses, a known design matrix (or independent variable)  $\mathbf{X}_{n \times q=r+1}$ ,  $\mathbf{B}_{q \times p}$  a matrix of unknown parameters, and  $\mathbf{U}$  the unobserved errors. The linear regression model is

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

such that the rows of  $\mathbf{U}$  are uncorrelated with mean 0 and variance  $\Sigma$ . We assume that  $\mathbf{U}$  is normally distributed. Then, the log-likelihood for  $(\mathbf{B}, \Sigma)$  is given by

$$l(\mathbf{B}, \Sigma) = -\frac{n}{2} \log |2\pi\Sigma| - \frac{1}{2} \text{tr}(\mathbf{Y} - \mathbf{XB})\Sigma^{-1}(\mathbf{Y} - \mathbf{XB})^t$$

*Maximum likelihood Estimation :*

We assume that  $\mathbf{X}$  is of full row rank  $q$  s.t. the matrix  $\mathbf{X}^t\mathbf{X}$  is invertible. Write  $P = I - \mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t$ , the projector onto the ortho-complement of  $\mathbf{X}$  in  $\mathbb{R}^n$ .

Then, the MLE's are

$$\hat{\mathbf{B}} = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{Y}, \quad \hat{\Sigma} = \frac{1}{n}\mathbf{Y}^tP\mathbf{Y}$$

*Outline :* Observe that

$$tr(\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}})\Sigma^{-1}(\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}})^t = tr[\Sigma^{-1}(\mathbf{Y}^tP\mathbf{Y} + (\hat{\mathbf{B}} - \mathbf{B})^t(\mathbf{X}^t\mathbf{X})(\hat{\mathbf{B}} - \mathbf{B}))]$$

Clearly,  $\hat{\mathbf{B}}$  maximizes the second term. Further,

$$l(\hat{\mathbf{B}}, \Sigma) = -\frac{n}{2} \log |2\pi\Sigma| - \frac{n}{2} tr \Sigma^{-1} \hat{\Sigma}$$

maximized at  $\Sigma = \hat{\Sigma}$ .

As usual,  $\hat{\mathbf{B}}$  is unbiased for  $\mathbf{B}$ . Define the error estimates  $\hat{\mathbf{U}} = \mathbf{Y} - \mathbf{X}\hat{\mathbf{B}} = P\mathbf{Y}$ .

Observe that, both  $\hat{\mathbf{B}}$  and  $\hat{\mathbf{U}}$  are linear functions of  $\mathbf{Y}$  and hence Gaussian.

Moreover,  $E(\hat{\mathbf{U}}) = 0$  and  $\hat{\mathbf{B}}$ ,  $\hat{\mathbf{U}}$  are uncorrelated (hence independent). More-

over, the covariance between the different rows of the estimated coefficient matrix is given by

$$cov(\mathbf{B}_{(i)}, \mathbf{B}_{(j)}) = \sigma_{ij}((\mathbf{X}^t\mathbf{X})^{-1})$$

Consequently,  $\hat{\mathbf{B}}$  and  $\hat{\Sigma}$  are independent (since  $\hat{\Sigma} = \frac{1}{n}\hat{\mathbf{U}}^t\hat{\mathbf{U}}$ ). Also,  $n\hat{\Sigma} \sim W_p(\Sigma, n - q)$ .

The likelihood ratio tests for the regression parameters can be carried out as in the simple multiple regression fashion, where we test whether a particular set of parameters are zero or not. The question whether a set of variables are effective or not, can be translated to a hypothesis

$$H_0 : \mathbf{B} = 0$$

or more specifically,

$$H_0 : \mathbf{B}_{(2)} = 0$$

where  $\mathbf{B} = \begin{pmatrix} \mathbf{B}_{(1)} \\ \mathbf{B}_{(2)} \end{pmatrix}$ . The likelihood ratio statistics all turn out to be Wilk's

Lambda type statistics and can be tested as before. Check yourself !!!