IOP Conf. Series: Materials Science and Engineering **263** (2017) 042134 doi:10.1088/1757-899X/263/4/042134

MANCOVA for one way classification with homogeneity of regression coefficient vectors

G Mokesh Rayalu, J Ravisankar and G Y Mythili
Department of Mathematics, School of Advanced Sciences,VIT University, Vellore-632014, India

E-mail: mokesh.g@vit.ac.in

Abstract. The MANOVA and MANCOVA are the extensions of the univariate ANOVA and ANCOVA techniques to multidimensional or vector valued observations. The assumption of a Gaussian distribution has been replaced with the Multivariate Gaussian distribution for the vectors data and residual term variables in the statistical models of these techniques. The objective of MANCOVA is to determine if there are statistically reliable mean differences that can be demonstrated between groups later modifying the newly created variable. When randomization assignment of samples or subjects to groups is not possible, multivariate analysis of covariance (MANCOVA) provides statistical matching of groups by adjusting dependent variables as if all subjects scored the same on the covariates. In this research article, an extension has been made to the MANCOVA technique with more number of covariates and homogeneity of regression coefficient vectors is also tested.

1. Introduction

Suppose N observations vectors on 'm' study variables $y^{(1,2,.....m)}$ and 's' concomitant variables $x^{(1,2,.....s)}$) are distributed into 'k' groups of r_i observations.

Let $(x_{ij}^{(1)}, x_{ij}^{(2)}, \ldots, x_{ij}^{(s)}, y_{ij}^{(1)}, y_{ij}^{(2)}, \ldots, y_{ij}^{(m)})$ be the j[th] observation vector in the i[th] subclass with 's' covariates. $i = 1,2,3,4,\ldots,k;\ j = 1,2,3,4\ldots,r_i$

The MANCOVA model for one way classified data with 's' supplementary variables is given by

$$Y_{ij} = \mu + T_i + \sum_{h=1}^{s} \beta_h \left(x_{ij}^{(h)} - \bar{x}^{(h)} \right) + \in_{ij};$$

where $i = 1,2,3,4,\ldots,k;\ j = 1,2,3,4\ldots,r_i$ and $\in_{ij} \overset{i.i.d}{\sim} N(0,\Sigma)$ (1.1)

Here $\boldsymbol{Y}_{ij}$ is j[th] observation vector of main variables in i[th] subclass and

$\boldsymbol{y}_{ij}^{(u)}$ is j[th] observation on u[th] study variable $Y^{(u)}$ in i[th] subclass.

$\boldsymbol{x}_{ij}^{(h)}$ is j[th] observation on h[th] concomitant variable $X^{(h)}$ in i[th] subclass.

μ is general mean effect.

T_i is effect of i[th] subclass.

β_h is the vector of regression coefficients of Y on $X^{(h)}$.

Content from this work may be used under the terms of the Creative Commons Attribution 3.0 licence. Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.
Published under licence by IOP Publishing Ltd

$\beta_h = \begin{bmatrix} \beta_h^{(1)} \\ \beta_h^{(2)} \\ \vdots \\ \beta_h^{(m)} \end{bmatrix}$; where $\beta_h^{(u)}$ is the regression coefficient of $Y^{(u)}$ on $x^{(h)}$.

$\in_{ij}$'s are random error vectors.

$\bar{x}^{(h)} = \frac{x_{**}^{(h)}}{\sum_{i=1}^{k} r_i}; \; h = 1,2,\ldots,s$;

$x_{**}^{(h)}$ denotes the grand total of x values of h^{th} covariate and $\sum$ is variance covariance matrix.

Assumptions: Following are the assumptions under which MANCOVA technique can be Applied [3].

i. The observations are independent and the sample is completely random in MANCOVA
ii. In MANCOVA, explanatory variables are categorical variable and explained variable are scale variables .Covariates can be dichotomous, Continuous with ordinal.
iii. The explained variables cannot be too correlated
iv. The sample must be follow Multivariate Normality.
v. Between group variance are equal.
vi. The model for MANCOVA technique is linear model i.e., all the effects of different levels of several factors are additive in Nature.
vii. $X^{(1)}, X^{(2)}, \ldots, X^{(s)}$ are not influenced by the subclasses of the classification.

Null hypotheses:

$H_0 : T_1 = T_2 = \cdots = T_k = 0$ or $H_0 : \overline{T_1} = \overline{T_2} = \cdots = \overline{T_k} = 0$.

i.e., All the subclasses have the same effects and are not significant.

2. Expressions for the matrices of sum of squares and sum of products due to parameters

2.1 Matrices of sum of squares due to vector of main variables Y

Total sum of squares matrix

$$\left[T_{YY}\right]_{m\times m} = \sum_{i=1}^{k}\sum_{j=1}^{r_i} Y_{ij}Y_{ij}^T - \frac{Y_{**}Y_{**}^T}{\sum_{i=1}^{k} r_i} \tag{2.1}$$

Between subclasses sum of squares matrix

$$\left[A_{YY}\right]_{m\times m} = \sum_{i=1}^{k} \frac{Y_{i*}Y_{i*}^T}{r_i} - \frac{Y_{**}Y_{**}^T}{\sum_{i=1}^{k} r_i} \tag{2.2}$$

Error sum of squares matrix $\left[E_{YY}\right]_{m\times m} = \left[T_{YY}\right]_{m\times m} - \left[A_{YY}\right]_{m\times m}$ (2.3)

2.2 Sum of squares due to h^{th} ancillary variable $X^{(h)}$(h=1,2,…,s)

Total sum of squares

$$T_{x^{(h)}x^{(h)}} = \sum_{i=1}^{k}\sum_{j=1}^{r_i} x_{ij}^{(h)^2} - \frac{x_{**}^{(h)^2}}{\sum_{i=1}^{k} r_i} \tag{2.4}$$

Between subclasses sum of squares

$$A_{x^{(h)}*x^{(h)}} = \sum_{i=1}^{k} \frac{x_{i**}^{(h)^2}}{r_i} - \frac{x_{***}^{(h)^2}}{\sum_{i=1}^{k} r_i} \tag{2.5}$$

Error sum of squares $E_{x^{(h)}*x^{(h)}} = T_{x^{(h)}*x^{(h)}} - A_{x^{(h)}*x^{(h)}}$ (2.6)

2.3 Matrices of sum of products due to vector of main variables Y and each of the h[th] ancillary

2.4 variable X[(h)](h=1,2,…,s)

Total sum of products matrix

$$\left[T_{x^{(h)}Y}\right]_{1\times m} = \sum_{i=1}^{k}\sum_{j=1}^{r_i} x_{ij}^{(h)} Y_{ij}^{T} - \frac{x_{**}^{(h)} Y_{**}^{T}}{\sum_{i=1}^{k} r_i} \tag{2.7}$$

Between subclasses sum of products matrix

$$\left[A_{x^{(h)}Y}\right]_{1\times m} = \sum_{i=1}^{k} \frac{x_{i*}^{(h)} Y_{i*}^{T}}{r_i} - \frac{x_{**}^{(h)} Y_{**}^{T}}{\sum_{i=1}^{k} r_i} \tag{2.8}$$

Error sum of products matrix

$$\left[E_{x^{(h)}Y}\right]_{1\times m} = \left[T_{x^{(h)}Y}\right]_{1\times m} - \left[A_{x^{(h)}Y}\right]_{1\times m} \tag{2.9}$$

2.5 Sum of products with respect to covariates X[(e)]and X[(f)] (e ≠ f=1,2,…,s)

Total sum of products

$$T_{x^{(e)}x^{(f)}} = \sum_{i=1}^{k}\sum_{j=1}^{r_i} x_{ij}^{(e)} x_{ij}^{(f)} - \frac{x_{**}^{(e)} x_{**}^{(f)}}{\sum_{i=1}^{k} r_i} \tag{2.10}$$

Between subclasses sum of products

$$A_{x^{(e)}x^{(f)}} = \sum_{i=1}^{k} \frac{x_{i**}^{(e)} x_{i**}^{(f)}}{r_i} - \frac{x_{***}^{(e)} x_{***}^{(f)}}{\sum_{i=1}^{k} r_i} \tag{2.11}$$

Error sum of products

$$E_{x^{(e)}*x^{(f)}} = T_{x^{(e)}*x^{(f)}} - A_{x^{(e)}*x^{(f)}} \tag{2.12}$$

3. Testing the homogeneity of regression coefficient vectors[1]

Wilks' lambda statistic $\Lambda = \frac{|E|}{|A+E|} = \frac{|E|}{|T|}$ (3.1)

is corresponding to the LRT and this ratio of generalized variances can be used to test the above H_0. If this ratio is too small, null hypothesis is rejected.

For more than three groups and more than three variables, Bartlett has shown that if H_0 is true and $\sum r_i$ is large,

$$-\left(\sum_{i=1}^{k} r_i - g - s - \frac{a+1-(g-1)}{2}\right)\ln\Lambda = -\left(\frac{2\sum_{i=1}^{k} r_i - 2g - 2s - a + g - 2}{2}\right)\ln\Lambda = -\left(\frac{2\left(\sum_{i=1}^{k} r_i - s - 1\right)}{2} - \frac{a+g}{2}\right)\ln\Lambda$$

$$= -\left(\sum_{i=1}^{k} r_i - s - 1 - \frac{a+g}{2}\right)\ln\Lambda \sim \chi^2_{m(k-1)} \tag{3.2}$$

Reject H_0 at significance level α% l.o.s when $\sum r_i$ is large

The following three other multivariate test statistics are also used[2]:

Lawley-Hotelling trace = $tr[AE^{-1}]$ (3.3)

Pillai's trace = $tr[A(A+E)^{-1}]$ (3.4)

Roy's largest root = max eigen value of $E(A+E)^{-1}$ (3.5)

When there is a single non zero eigen value, the power of Roy's test is best and the power is large.

For testing the homogeneity of regression coefficients of Y on covariate $X^{(h)}$, the null hypothesis can be stated as that the vectors of slopes of regression lines of Y on $X^{(h)}$ are the same[4].

This hypothesis is assumed to be true in any application of the multivariate analysis of covariance[5].

To test for homogeneity of regression coefficient vectors, the following statistics are calculated:

Wilks' lambda statistic[6]

$$\Lambda^* = \frac{|T|}{\left|B^{(h)} - T + T\right|} = \frac{|T|}{\left|B^{(h)}\right|},$$

where $|T| = \det\left[T_{YY}^{adj}\right]_{m\times m}$ a $\left|B^{(h)}\right| = \det\left[B^{(h)}\right]_{m\times m} = \det\left[T_{x^{(h)}x^{(h)}} I_{m\times m} - \left[T_{YY}^{-1}\right]_{m\times m}\left[T_{x^{(h)}Y}^{T}\right]_{m\times s}\left[T_{x^{(h)}Y}\right]_{s\times m}\right]$ (3.6)

According to Bartlett, for more than three groups and more than three variables, if H_0 is true and $\sum r_i - k$ is large,

$$-\left(\sum_{i=1}^{k} r_i - 2g - \frac{a+1-(g-1)}{2}\right)\ln\Lambda^*$$

$$= -\left(\frac{2\sum_{i=1}^{k} r_i - 4g - a + g - 2}{2}\right)\ln\Lambda^* = -\left(\frac{2\left(\sum_{i=1}^{k} r_i - g - 1\right)}{2} - \frac{a+g}{2}\right)\ln\Lambda^* = -\left(\sum_{i=1}^{k} r_i - g - 1 - \frac{a+g}{2}\right)\ln\Lambda^* \sim \chi^2_{m(k-1)} \tag{3.7}$$

Rreject H_0 at significance level α when $\sum r_i - k$ is large [7].

where $\chi^2_{m(k-1)}(\alpha)$ is the upper $(100\alpha)^{th}$ percentile [8].

The following three other multivariate test statistics are also used:

Lawley-Hotelling trace = $tr[[B^{(h)}-T]T^{-1}]$ (3.8)

Pillai's trace = $tr[[B^{(h)}-T](B^{(h)})^{-1}]$ (3.9)

Roy's largest root = maximum eigen value of $T(B^{(h)})^{-1}$ (3.10)

When there is a single non zero eigen value, the power of Roy's test is best and the power is large.

4. Conclusions

The extension of MANCOVA technique with more number of ancillary variables and tests for homogeneity of vectors of regression coefficients has been developed in the present study.In the MANCOVA technique, the layout of the design can be changed at the testing stage, if necessary. To test the influences of subclasses on h^{th} covariate $X^{(h)}$; h=1,2,…,s. One may test the null hypothesis as follows:H_0: $X^{(h)}$ is not influenced or affected by the effects of the subclasses of the classification.The ANOVA for one way classified data can be carried out separately for h^{th} covariate $X^{(h)}$, so that the resulting test statistic infers whether it is influenced or not. If this is not significant then one may choose $X^{(h)}$ as concomitant variable for MANCOVA technique otherwise not considered as a covariate.

References

[1] D'Agostino R B and Sullivar L M 2005 *In:Encyclopedia of Biostatistics*, **5**,Wiley-Blackwell.
[2] Ferguson G A 1989 *Statistical Analysis in Psychology and Education* McGraw-Hill New York.
[3] Giri N C 2004 *Multivariate Statistical Analysis* Second Edition Marcel Dekker Inc New York.
[4] Johnson R A and Wichern D W 2007 *Applied Multivariate Statistical Analysis* Sixth Edition Pearson Prentice Hall Pearson Education Inc New Jersey.
[5] Mertler C A and Vannatta R A 2002 *Advanced and Multivariate Statistical Methods Practical Application and Interpretation*Second Edition, Pyrczak Publishing, Canada.
[6] Quinn G P and Keough M J 2002 *Experimental Design and Data Analysis for Biologists* Cambridge University Press Cambridge.
[7] Tabachnik B G and Fidell L S 2012 *Multivariate Statistics* Pearson Education.
[8] Wolfgang Karl Hardle and Leopold Simar 2015*Applied Multivariate Statistical Analysis*Fourth Edition Springer-Heidelberg New York.