Analysis of variance, multivariate (MANOVA)

Abstract: A designed experiment is set up in which the system studied is under the control of an investigator. The individuals, the treatments, the variables measured are decided by the investigator. In a clinical trial, patients satisfying some eligibility criteria are assigned at random, either to a new treatment or to a placebo. Patients are followed for some time and a few responses are recorded for each patient. In an agricultural experiment, a field is divided into plots of shape and size determined by the investigator. The plots are assigned at random one among a few treatments which could be, e.g., different fertilizers. The variables measured on each plot could be the yields of some crops. The general objective of a designed experiment is to assess the effects of different treatments on the responses. These effects are evaluated by statistical estimates and confidence intervals of the magnitude of the differences between treatments. The estimates should avoid biases and the random errors should be minimized as much as possible. The statistical methodology used to analyze such designed experiments in which there are several responses is termed MANOVA, an acronym for **M**ultivariate **AN**alysis **O**f **VA**riance.

Keywords: MANOVA; linear model; factor; experiment

Multivariate regression model and the general linear hypothesis

The linear model considered is $\mathbf{Y} = \mathbf{XB} + \mathbf{E}$, where $\mathbf{Y} : n \times p$ is the matrix of responses, $\mathbf{B} : k \times p$ is the matrix of unknown regression coefficients, and $\mathbf{X} : n \times k$ is a fixed and known design matrix of rank k. The rows of the error term $\mathbf{E}$ are independently distributed as multivariate normal $N_p(\mathbf{0}, \mathbf{\Sigma})$ with an unknown positive definite covariance matrix. The maximum likelihood estimators of the unknown parameters are $\hat{\mathbf{B}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ and $\hat{\mathbf{\Sigma}} = (\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}})'(\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}})/n$. The fundamental property of these estimates is that they form a set of sufficient statistics. Moreover, they are independently distributed as multivariate normal, $\hat{\mathbf{B}} \sim N\left(\mathbf{B}, (\mathbf{X}'\mathbf{X})^{-1} \otimes \mathbf{\Sigma}\right)$, and Wishart, $n\hat{\mathbf{\Sigma}} \sim W_p(n-k, \mathbf{\Sigma})$.

The general linear hypothesis is the hypothesis $H_0 : \mathbf{CB} = \mathbf{0}$, where the known matrix $\mathbf{C} : r \times k$ is of rank r. This null hypothesis is being tested against all alternatives in the multivariate regression model.

A canonical form

The canonical form due to Roy [8] and Anderson [1], see also Bilodeau and Brenner [3] or Eaton [5], is obtained by transforming the original response $\mathbf{Y}$ so that in the new model $\mathbf{X}$ becomes $\mathbf{X}_0 = \begin{pmatrix} \mathbf{I}_k \\ \mathbf{0} \end{pmatrix}$ and $\mathbf{C}$ reduces to $\mathbf{C}_0 = \begin{pmatrix} \mathbf{I}_r, \mathbf{0} \end{pmatrix}$. It can be shown that an equivalent problem in its canonical form is to test $H_0 : \mathbf{M}_1 = \mathbf{0}$ against $H_1 : \mathbf{M}_1 \neq \mathbf{0}$ based on $\mathbf{Z}_1 : r \times p \sim N(\mathbf{M}_1, \mathbf{I}_r \otimes \boldsymbol{\Sigma})$, $\mathbf{Z}_2 : (k-r) \times p \sim N(\mathbf{M}_2, \mathbf{I}_{k-r} \otimes \boldsymbol{\Sigma})$, and $\mathbf{Z}_3 : (n-k) \times p \sim N(\mathbf{0}, \mathbf{I}_{n-k} \otimes \boldsymbol{\Sigma})$, where $\mathbf{Z}_1$, $\mathbf{Z}_2$, and $\mathbf{Z}_3$ are independent.

The Wilks likelihood ratio test (LRT) rejects the null hypothesis for small values of

$$\Lambda = \frac{|\mathbf{Z}_3'\mathbf{Z}_3|^{n/2}}{|\mathbf{Z}_1'\mathbf{Z}_1 + \mathbf{Z}_3'\mathbf{Z}_3|^{n/2}}.$$

At this point, it is convenient to define the so called $U(p; m, n)$ distribution which is the distribution of $|\mathbf{W}_1| / |\mathbf{W}_1 + \mathbf{W}_2|$, where $\mathbf{W}_1 \sim W_p(n, \mathbf{I})$ is independently distributed of $\mathbf{W}_2 \sim W_p(m, \mathbf{I})$. Then, the likelihood ratio test has the distribution $\Lambda^{2/n} \sim U(p; r, n-k)$. It can be shown that $-[n - \frac{1}{2}(p - m + 1)] \log U(p; m, n)$ has a limiting chi-square distribution with pm degrees of freedom. This chi-square approximation can be adjusted by a C factor to get a test of exact significance level α. One computes the test statistic $U(p; m, n)$ and rejects the null hypothesis at significance α if

$$-\left[n - \frac{1}{2}(p - m + 1)\right] \log U(p; m, n) > C_{p,m,n-p+1}(\alpha) \cdot \chi^2_{pm}(\alpha).$$

Tables of the C factors can be found in Anderson [2] and more extended tables are available in Pillai and Gupta [7].

Other invariant tests

The problem in its canonical form is left invariant by the group of transformations

$$\mathbf{Z}_1 \mapsto \boldsymbol{\Gamma}_1\mathbf{Z}_1\mathbf{A}', \;\mathbf{Z}_2 \mapsto \boldsymbol{\Gamma}_2\mathbf{Z}_2\mathbf{A}' + \mathbf{N}, \;\mathbf{Z}_3 \mapsto \boldsymbol{\Gamma}_3\mathbf{Z}_3\mathbf{A}',$$

where $\boldsymbol{\Gamma}_1$, $\boldsymbol{\Gamma}_2$, and $\boldsymbol{\Gamma}_3$ are orthogonal, $\mathbf{A}$ is non singular, and $\mathbf{N}$ is arbitrary. The LRT is invariant with respect to this group of transformations. Let $t = \min(p, r)$, a maximal invariant is the set of non zero eigenvalues

$f_1 \geq \cdots \geq f_t$ of the matrix $\mathbf{Z}_1'\mathbf{Z}_1(\mathbf{Z}_3'\mathbf{Z}_3)^{-1}$. In the very special case $t = 1$, the LRT is uniformly most powerful invariant (UMPI). When $p = 1$ the LRT is equivalent to the usual F ratio, whereas when $r = 1$ it is equivalent to Hotelling's T^2 test. Both of these tests are UMPI. Unfortunately, when $t > 1$ there does not exist an UMPI test (even a locally MPI), see Fujikoshi [6].

Expressed in terms of the original MANOVA problem, $f_1 \geq \cdots \geq f_t$ are the t largest eigenvalues of

$$\left\{\hat{\mathbf{B}}'\mathbf{C}'[\mathbf{C}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}']^{-1}\mathbf{C}\hat{\mathbf{B}}\right\}(n\hat{\boldsymbol{\Sigma}})^{-1}$$

and the Wilks LRT rejects for small values of

$$\prod_{i=1}^{t}(1+f_i)^{-1} = \frac{\left|n\hat{\boldsymbol{\Sigma}}\right|}{\left|n\hat{\boldsymbol{\Sigma}} + \hat{\mathbf{B}}'\mathbf{C}'[\mathbf{C}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}']^{-1}\mathbf{C}\hat{\mathbf{B}}\right|}.$$

Other invariant tests proposed by popular statistical softwares are the Lawley-Hotelling trace test

$$\sum_{i=1}^{t} f_i = tr\left[\mathbf{Z}_1'\mathbf{Z}_1(\mathbf{Z}_3'\mathbf{Z}_3)^{-1}\right],$$

the test of Bartlett-Nanda-Pillai

$$\sum_{i=1}^{t} f_i(1+f_i)^{-1} = tr\left[\mathbf{Z}_1'\mathbf{Z}_1(\mathbf{Z}_1'\mathbf{Z}_1 + \mathbf{Z}_3'\mathbf{Z}_3)^{-1}\right],$$

and Roy's maximum root test f_1. Tables of critical values for these tests are available in Anderson [2].

Models of MANOVA

Completely randomized design with one factor

An experiment is set up to investigate the effects of a factor A, at levels $1, \ldots, I$, on p responses. A number n_i of experimental units are chosen at random and assigned to the treatment corresponding to the level i of the factor. There are in total $n = \sum_{i=1}^{I} n_i$ experimental units. The linear model used is

$$\mathbf{y}_{ij} = \boldsymbol{\mu} + \boldsymbol{\alpha}_i + \mathbf{e}_{ij},\ i = 1, \ldots, I;\ j = 1, \ldots, n_i,$$

where $\mathbf{y}_{ij} : p \times 1$ is the vector of the p responses observed on the experimental unit j at level i of the factor. The parameters represent an overall effect $\boldsymbol{\mu}$ and an effect $\boldsymbol{\alpha}_i$ due to factor A at level i. The error terms $\mathbf{e}_{ij}$ are supposed independent $N_p(\mathbf{0}, \boldsymbol{\Sigma})$. The constraint $\sum_{i=1}^{I} n_i \boldsymbol{\alpha}_i = \mathbf{0}$ is imposed to avoid over parametrization. The null hypothesis of interest is that of the absence of effect of factor A, $H_A : \boldsymbol{\alpha}_1 = \cdots = \boldsymbol{\alpha}_I = \mathbf{0}$. Let $\bar{\mathbf{y}}_{..}$ be the mean over all units and $\bar{\mathbf{y}}_{i.}$ be the mean over units having received factor A at level i. Wilks test is expressed in terms of sums of squares

$$\begin{aligned} SST &= \sum_{i=1}^{I} \sum_{j=1}^{n_i} \mathbf{y}_{ij} \mathbf{y}_{ij}' - n \bar{\mathbf{y}}_{..} \bar{\mathbf{y}}_{..}' \\ SS(A) &= \sum_{i=1}^{I} n_i \bar{\mathbf{y}}_{i.} \bar{\mathbf{y}}_{i.}' - n \bar{\mathbf{y}}_{..} \bar{\mathbf{y}}_{..}' \\ SSE &= SST - SS(A) \end{aligned}$$

and rejects the null hypothesis whenever $U(p; I-1, n-I) = |SSE|/|SSE + SS(A)|$ is small. Simultaneous $(1-\alpha)100\%$ confidence intervals on contrasts of treatment effects, $\sum_{i=1}^{I} \mathbf{a}' \boldsymbol{\alpha}_i b_i$ where $\mathbf{a} : p \times 1$ is arbitrary and the constants b_i satisfying $\sum_{i=1}^{I} b_i = 0$ define the contrast, can be constructed using the Bonferroni method or using Roy's maximum root test, see Srivastava [10]. These simultaneous confidence intervals can be adapted to other MANOVA models such as those in the subsequent sections.

Randomized complete block design with one factor

In a clinical trial, *a priori* information on variables, associated with the responses, such as sex, age, and stage of a disease, is used to group experimental units into homogeneous blocks. In the absence of effect of the factor, experimental units within a block should give similar responses. When the factor A has I levels, blocks of I homogeneous experimental units are formed. Within a block, the I levels of the factor are assigned at random to the I experimental units using, *e.g.*, a random permutation. The linear model with J blocks becomes

$$\mathbf{y}_{ij} = \boldsymbol{\mu} + \boldsymbol{\alpha}_i + \boldsymbol{\beta}_j + \mathbf{e}_{ij},\ i = 1, \ldots, I;\ j = 1, \ldots, J,$$

subject to the constraints $\sum_{i=1}^{I} \boldsymbol{\alpha}_i = \sum_{j=1}^{J} \boldsymbol{\beta}_j = \mathbf{0}$. The parameter $\boldsymbol{\mu}$ is an overall effect, $\boldsymbol{\alpha}_i$ is the effect due to factor A at level i and $\boldsymbol{\beta}_j$ is the effect of block j. The sums of squares required are

$$\begin{aligned} SST &= \sum_{i=1}^{I}\sum_{j=1}^{J} \mathbf{y}_{ij}\mathbf{y}'_{ij} - IJ\bar{\mathbf{y}}_{..}\bar{\mathbf{y}}'_{..} \\ SS(A) &= J\sum_{i=1}^{I} \bar{\mathbf{y}}_{i.}\bar{\mathbf{y}}'_{i.} - IJ\bar{\mathbf{y}}_{..}\bar{\mathbf{y}}'_{..} \\ SS(blocks) &= I\sum_{j=1}^{J} \bar{\mathbf{y}}_{.j}\bar{\mathbf{y}}'_{.j} - IJ\bar{\mathbf{y}}_{..}\bar{\mathbf{y}}'_{..} \\ SSE &= SST - SS(A) - SS(blocks). \end{aligned}$$

Wilks test for the hypothesis of absence of factor A effect, $H_A : \boldsymbol{\alpha}_1 = \cdots = \boldsymbol{\alpha}_I = \mathbf{0}$, is the test which rejects for small values of $U(p; I-1, (I-1)(J-1)) = |SSE|/|SSE + SS(A)|$. A similar test for the effect of blocking can be conducted but is usually of no intrinsic interest. The latin square design is another specialized blocking technique, see Cox and Reid [4] for a general discussion and Srivastava [10] for the statistical MANOVA analysis.

Balanced factorial design

In agriculture, the effects of three fertilizers, say nitrogen, potassium and potash, each applied at low or high quantities, on the yield of some crops is an example of a factorial experiment with two factors: fertilizer and quantity. A balanced factorial design consists of an equal number of replicates of all possible treatments; a treatment being any combination of the levels of the factors. Factorial experiments are usually preferred to doing two experiments investigating one factor at a time. They are more efficient for estimating main effects, which are the averaged effects of a single factor over all units. Interactions among factors can also be assessed in a factorial experiment but not from two one-at-a-time experiments, see Cox and Reid [4]. The linear model of a two factor balanced factorial design with interaction replicated r times is

$$\mathbf{y}_{ijk} = \boldsymbol{\mu} + \boldsymbol{\alpha}_i + \boldsymbol{\beta}_j + (\boldsymbol{\alpha\beta})_{ij} + \mathbf{e}_{ijk},\ i = 1, \ldots, I;\ j = 1, \ldots, J;\ k = 1, \ldots, r,$$

subject to the constraints $\sum_{i=1}^{I} \boldsymbol{\alpha}_i = \sum_{j=1}^{J} \boldsymbol{\beta}_j = \sum_{i=1}^{I} (\boldsymbol{\alpha\beta})_{ij} = \sum_{j=1}^{J} (\boldsymbol{\alpha\beta})_{ij} = \mathbf{0}$. The parameter $\boldsymbol{\mu}$ is an overall effect, $\boldsymbol{\alpha}_i$ is a main effect of factor A, $\boldsymbol{\beta}_j$ is a main effect of factor B, and $(\boldsymbol{\alpha\beta})_{ij}$ is an interaction effect between the two factors. The sums of squares required are

$$SST = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{r} \mathbf{y}_{ijk}\mathbf{y}'_{ijk} - IJr\bar{\mathbf{y}}_{...}\bar{\mathbf{y}}'_{...}$$
$$SS(A) = Jr\sum_{i=1}^{I} \bar{\mathbf{y}}_{i..}\bar{\mathbf{y}}'_{i..} - IJr\bar{\mathbf{y}}_{...}\bar{\mathbf{y}}'_{...}$$
$$SS(B) = Ir\sum_{j=1}^{J} \bar{\mathbf{y}}_{.j.}\bar{\mathbf{y}}'_{.j.} - IJr\bar{\mathbf{y}}_{...}\bar{\mathbf{y}}'_{...}$$
$$SS(Treatments) = r\sum_{i=1}^{I}\sum_{j=1}^{J} \bar{\mathbf{y}}_{ij.}\bar{\mathbf{y}}'_{ij.} - IJr\bar{\mathbf{y}}_{...}\bar{\mathbf{y}}'_{...}$$
$$SS(A \times B) = SS(Treatments) - SS(A) - SS(B)$$
$$SSE = SST - SS(Treatments).$$

The hypotheses of absence of A main effects $H_A : \boldsymbol{\alpha}_i = \mathbf{0}$, for all i, of absence of B main effects $H_B : \boldsymbol{\beta}_j = \mathbf{0}$, for all j, and of absence of interactions $H_{A\times B} : (\boldsymbol{\alpha\beta})_{ij} = \mathbf{0}$, for all i and j, can be tested using Wilks tests given respectively by

$$U(p; I-1, IJ(r-1)) = |SSE|/|SSE + SS(A)|$$
$$U(p; J-1, IJ(r-1)) = |SSE|/|SSE + SS(B)|$$
$$U(p; (I-1)(J-1), IJ(r-1)) = |SSE|/|SSE + SS(A \times B)|.$$

The computations of any MANOVA can be performed easily with the manova statement of either the SAS [9] proc anova (for balanced designs) or more generally the proc glm.

References

[1] Anderson T.W. (1958). *An Introduction to Multivariate Statistical Analysis.* John Wiley & Sons, New York.

[2] Anderson T.W. (1984). *An Introduction to Multivariate Statistical Analysis.* 2nd ed. John Wiley & Sons, New York.

[3] Bilodeau M., Brenner D. (1999). *Theory of Multivariate Statistics.* Springer-Verlag, New York.

[4] Cox D.R., Reid N. (2000). *The Theory of the Design of Experiments.* Chapman and Hall/CRC.

[5] Eaton M.L. (1983). *Multivariate Statistics, a Vector Space Approach.* John Wiley & Sons, New York.

[6] Fujikoshi Y. (1988). Comparison of powers of a class of tests for multivariate linear hypothesis and independence. *Journal of Multivariate Analysis* **26**, 48–58.

[7] Pillai K.C.S., Gupta A.K. (1969). On the exact distribution of Wilks's criterion, *Biometrika* **56**, 109–118.

[8] Roy S.N. (1957). *Some aspects of multivariate analysis.* John Wiley & Sons, New York.

[9] SAS Institute Inc., SAS 9.1.3 Help and Documentation, Cary, NC: SAS Institute Inc., 2000-2004.

[10] Srivastava M.S. (2002). *Methods of Multivariare Statistics.* John Wiley & Sons, New York.