Multivariate Statistical Analysis

Fall 2011

C. L. Williams, Ph.D.

Lecture 10 for Applied Multivariate Analysis

Outline

Computation of T^2

As we have already seen there are a number of ways to calculate Hotelling's T^2. In fact, there are other approaches as well depending on the type of multivariate analysis we want to conduct and what type of questions need answering.

$$
\begin{aligned}
T^2 &= (n_1 + n_2 - 2)\,\frac{1-\Lambda}{\Lambda} && \text{Wilk's } \Lambda \\
T^2 &= (n_1 + n_2 - 2)\,U^{(s)} && \text{Lawley-Hotelling's Lambda} \\
T^2 &= (n_1 + n_2 - 2)\,\frac{V^{(s)}}{1-V^{(s)}} && \text{Pillai's } V^{(s)} \\
T^2 &= (n_1 + n_2 - 2)\,\frac{\theta}{1-\theta} && \text{Roy's largest root } \theta
\end{aligned}
$$

each of these we will study in a little more detail in Chapter 6

Obtaining T^2 from multiple regression

Recall as we stated earlier, in the strictest sense, multiple regression is not a multivariate method. But we can use it to determine the value of a T^2, taking advantage of the nature of multiple regression and of T^2. This method also highlight the use of MANAOV in producing T^2. Consider the following. Let

$$\begin{aligned} w_i &= \frac{n_2}{n_1+n_2} \text{ for each of } \mathbf{y}_{11}, \mathbf{y}_{12}, \ldots, \mathbf{y}_{1n_1} \text{ in sample 1} \\ &= -\frac{n_1}{n_1+n_2} \text{ for each of } \mathbf{y}_{21}, \mathbf{y}_{22}, \ldots, \mathbf{y}_{2n_1} \text{ in sample 2} \end{aligned}$$

so that the prediction or estimating equation for the regression of the w_i on the y_i's.

$$\widehat{w}_i = b_0 + b_1 y_{i1} + b_2 y_{i2} + \cdots + b_p y_{ip}$$

It is easy enough to show that $\overline{w} = 0$ for all n_1+n_2 observations. and that the estimate for β_0, b_0 can be determined

$$b_0 = \overline{w} - b_1\overline{y}_1 - b_1\overline{y}_1 - b_2\overline{y}_2 - \cdots - b_p\overline{y}_p$$

so that

$$\begin{aligned} \widehat{w}_i &= \overline{w} + b_1\left(y_{i1} - \overline{y}_1\right) + b_2\left(y_{i2} - \overline{y}_2\right) + \cdots + b_p\left(y_{ip} - \overline{y}_p\right) \\ &= b_1\left(y_{i1} - \overline{y}_1\right) + b_2\left(y_{i2} - \overline{y}_2\right) + \cdots + b_p\left(y_{ip} - \overline{y}_p\right) \end{aligned}$$

Given the regression coefficients and the squared multiple correlation coefficient we can construct the T^2

$$T^2 = (n_1 + n_2 - 2)\frac{R^2}{1 - R^2}$$

as well as the discriminant function

$$\mathbf{a} = S_{pl}^{-1}(\mathbf{y}_1 - \mathbf{y}_2)$$

Paired Multivariate case

We made the assumption that **y** and **x** have a bivariate normal distribution, in which **y** and **x** are correlated. Here we assume **y** and **x** are correlated and have a multivariate normal distribution:

$$\begin{pmatrix} \mathbf{y}_1 \\ \mathbf{x}_1 \end{pmatrix} \sim MVN_{2p}\left[\begin{pmatrix} \boldsymbol{\mu}_y \\ \boldsymbol{\mu}_x \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_{yy} & \boldsymbol{\Sigma}_{yx} \\ \boldsymbol{\Sigma}_{xy} & \boldsymbol{\Sigma}_{xx} \end{pmatrix}\right]$$

Note the same null is being tested as in the two-sample case

$$
\begin{aligned}
H_0 : \boldsymbol{\mu}_d &= 0 \\
H_0 : \boldsymbol{\mu}_1 &= \boldsymbol{\mu}_2
\end{aligned}
$$

We can calculate:

$$
\begin{aligned}
\overline{\mathbf{d}} &= \frac{1}{n}\sum_{i=1}^{n}\mathbf{d}_i \\
\mathbf{S}_d &= \frac{1}{n-1}\sum_{i=1}^{n}\left(\mathbf{d}_i - \overline{\mathbf{d}}\right)\left(\mathbf{d}_i - \overline{\mathbf{d}}\right)'
\end{aligned}
$$

Paired-Sample test statistic

Which is used to calculate:

$$T^2 = \overline{\mathbf{d}}' \left(\frac{\mathbf{S}_d}{n} \right)^{-1} \overline{\mathbf{d}}$$

Tests for additional information

We wish to test the hypothesis that $\mathbf{x}_1$ and $\mathbf{x}_2$ are redundant for separating the two groups, that is, that the extra q variables do not contribute anything significant beyond the information already available in $\mathbf{y}_1$ and $\mathbf{y}_2$ for separating the groups. This is in the spirit of a full and reduced model test in regression. However, here we are working with a subset of dependent variables as contrasted to the subset of independent variables in the regression setting. Thus both $\mathbf{y}$ and $\mathbf{x}$ are subvectors of dependent variables. In this setting, the independent variables would be grouping variables 1 and 2 corresponding to $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$. We are not asking if the x's can significantly separate the two groups by themselves, but whether they provide additional separation beyond the separation already achieved by the y's.

$$T^2_{p+q} = \left(\frac{n_1 n_2}{n_1 + n_2}\right) \left[\begin{pmatrix} \overline{\mathbf{y}}_1 \\ \overline{\mathbf{x}}_1 \end{pmatrix} - \begin{pmatrix} \overline{\mathbf{y}}_2 \\ \overline{\mathbf{x}}_2 \end{pmatrix}\right] S_{pl}^{-1} \left[\begin{pmatrix} \overline{\mathbf{y}}_1 \\ \overline{\mathbf{x}}_1 \end{pmatrix} - \begin{pmatrix} \overline{\mathbf{y}}_2 \\ \overline{\mathbf{x}}_2 \end{pmatrix}\right]$$

whereas T_p^2 for the reduced set of p variables is

$$T_p^2 = \left(\frac{n_1 n_2}{n_1 + n_2} \right) (\bar{\mathbf{y}}_\mathbf{1} - \bar{\mathbf{y}}_\mathbf{2})' \, \mathbf{S}_{yy}^{-1} \, (\bar{\mathbf{y}}_\mathbf{1} - \bar{\mathbf{y}}_\mathbf{2})$$

Then the test statistic for the significance of the increase from T_p^2 to T_{p+q}^2 is given by

$$T^2(\mathbf{x}|\mathbf{y}) = (\nu - p)\frac{T_{p+q}^2 - T_p^2}{\nu + T_p^2}$$

$$F = \left(\frac{\nu - p - q + 1}{q}\right) \frac{T^2_{p+q} - T^2_p}{\nu + T^2_p}$$

$$F_{q,\nu-p-q+1} = \left(\frac{\nu - p - q + 1}{q}\right) \frac{R^2_{p+q} - R^2_p}{1 - R^2_{p+q}}$$

$$t^2(x|\mathbf{y}) = (\nu - p)\frac{T^2_{p+1} - T^2_p}{\nu + T^2_p}$$

If $\mathbf{y}$ is $MNV_p(\mu, \mathbf{\Sigma})$ and the variables in $\mathbf{y}$ are commensurate (measured in the same units and with approximately equal variances as, for example, in the probe word data in Table 3.5), we may wish to compare the means $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \ldots, \boldsymbol{\mu}_p$ in $\boldsymbol{\mu}$. This might be of interest when a measurement is taken on the same research unit at successive times. Such situations are often referred to as repeated measures design or growth curves, which are discussed in some generality in Sections 6.9 and 6.10. In the present section, we discuss one- and two-sample profile analysis. Profile analysis for several samples is covered in Section 6.8. We cover it here for simplicity.