# Appendix 14: Box’s *M* Test

Box’s *M* statistic is used to test for homogeneity of covariance matrices. The *j*th set of *r* dependent variables in the *i*th cell are $y'_{ij} = x'_{ij}\mathbf{B} + e'_{ij}$ where $e_{ij} \sim N_r\left(0, w_{ij}^{-1}\Sigma_i\right)$ for $i = 1, \ldots, g$ and $j = 1, \ldots, n_i$. The null hypothesis of the test for homogeneity of covariance matrices is $H_o : \Sigma_1 = \cdots = \Sigma_g$. Box (1949) derived a test statistic based on the likelihood-ratio test. The test statistic is called Box’s *M* statistic. For moderate to small sample sizes, an *F* approximation is used to compute its significance.

Box’s *M* statistic is not designed to be used in a linear model context;[1] therefore the observed cell means are used in computing the statistic.

## Notation

The following notation is used throughout this chapter, unless otherwise stated:

$g$ Number of cells with non-singular covariance matrices.

$n_i$ Number of cases in the *i*th cell.

$n$ Total sample size, $n = n_1 + \cdots + n_g$.

$\mathbf{y}_{ij}$ The *j*th set of dependent variables in the *i*th cell. A column vector of length *r*.

$w_{ij}$ Regression weight associated with $\mathbf{y}_{ij}$. It is assumed $w_{ij} > 0$.

[1] Although Anderson (1958, Section 10.2) mentioned that the population cell means can be expressed as linear combinations of parameters, he assumed that the combination coefficients are different for different cells, which is not the model assumed for GLM .

# Statistics

## Means

$$\bar{\mathbf{y}}_i = \sum\nolimits_{j=1}^{n_i} \mathbf{y}_{ij} \big/ n_i$$

## Cell Covariance Matrix

$$\mathbf{S}_i = \begin{cases} \sum\nolimits_{j=1}^{n_i} w_{ij} \left(\mathbf{y}_{ij} - \bar{\mathbf{y}}_i\right)\left(\mathbf{y}_{ij} - \bar{\mathbf{y}}_i\right)' \Big/ \left(n_i - 1\right) & \text{if } n_i > 1 \\ \mathbf{0} & \text{if } n_i \le 1 \end{cases}$$

## Pooled Covariance Matrix

$$\mathbf{S} = \begin{cases} \sum\nolimits_{i=1}^{g} (n_i - 1)\mathbf{S}_i \Big/ (n - g) & \text{if } n > g \\ \mathbf{0} & \text{if } n \le g \end{cases}$$

## Box's *M* Statistic

$$M = \begin{cases} (n-g)\log|\mathbf{S}| - \sum_{i=1}^{g} \left(n_i - 1\right)\log|\mathbf{S}_i| & \text{if } |\mathbf{S}| > 0 \\ \text{SYSMIS} & \text{if } |\mathbf{S}| \le 0 \end{cases}$$

## Significance

$$1-\mathrm{CDF.F}\left(\gamma M, f_1, f_2\right)$$

where CDF.F is the SPSS function for the cumulative *F* distribution and

$$f_1=(g-1)r(r+1)/2$$

$$\rho=1-\frac{2r^2+3r-1}{6(r+1)(g-1)}\left(\sum_{i=1}^{g}\frac{1}{(n_i-1)}-\frac{1}{(n-g)}\right)$$

$$\tau-\frac{(r-1)(r+2)}{6(g-1)}\left(\sum_{i=1}^{g}\frac{1}{(n_i-1)^2}-\frac{1}{(n-g)^2}\right)$$

$$f_2=\frac{f_1+2}{\left|\tau-(1-\rho)^2\right|}$$

$$\gamma=\frac{\left(\rho-f_1/f_2\right)}{f_1}$$

The significance is a system-missing value whenever the denominator is zero in the above expression.

# References

Anderson, T. W. 1958. *Introduction to multivariate statistical analysis*. New York: John Wiley & Sons, Inc.

Box, G. E. P., 1949. A general distribution theory for a class of likelihood criteria. *Biometrika*, 36: 317–346.

Seber, G. A. F. 1984. *Multivariate observations*. New York: John Wiley & Sons, Inc. (Section 9.2.6).