Matrix Formation in Univariate and Multivariate General Linear Models

Arwa A. Alkhalaf [a]
aalkhalaf@kau.edu.sa
[a] *Measurement, Evaluation and Research Methodology*
Department of Psychology, Faculty of Education
King Abdulaziz University, Jeddah, KSA

Abstract
This paper offers an overview of matrix formation and calculation techniques within the framework of General Linear Models (GLMs). It takes a sequential approach, beginning with a detailed exploration of matrix formation and calculation methods in regression analysis and univariate analysis of variance (ANOVA). Subsequently, it extends the discussion to cover multivariate analysis of variance (MANOVA). The primary objective of this study was to provide a clear and accessible explanation of the underlying matrices that play a crucial role in GLMs. Through linking, essentially different statistical methods, by fundamental principles and algebraic foundations that underpin the GLM estimation. Insights presented here aim to assist researchers, statisticians, and data analysts in enhancing their understanding of GLMs and their practical implementation in diverse research domains. This paper contributes to a better comprehension of the matrix-based techniques that can be extended to GLMs.

Keywords
Matrix formation, General Linear Models, Univariate Analysis of Variance (ANOVA), Multivariate Analysis of Variance (MANOVA)

1. Introduction

Modeling refers to the development of mathematical expressions that explains the behavior of a random variable of interest. It is aimed at describing how the mean of a dependent variable changes with changing conditions [1]. General linear modeling (GLM) involves solving algebraic equations that are complex, which can be conceptually simplified with the use of geometry [2]. Matrix formation of general linear models provides for easier and understandable calculations for large models and samples. This paper aims at summarizing the matrix formation and calculation of general linear models; starting with the regression and univariate analysis of variance approaches then multivariate analysis of variance.

The simplest linear model involves only one independent variable and states that the true mean of the dependent variable (Y) changes at a constant rate as the value of the independent variable (X) changes. The difference of an observation (Y_i) from its population mean E(Y) is taken into account by adding a random error (e). The functional relationship between the dependent variable Y, and independent variable X is the equation of a straight line, shown in Eq. (1).

Eq(1): $$Y = \beta_0 + \beta_1 X + e$$

Most models will use more than one independent variable to explain the behavior of the dependent variable. The linear model can be extended to include any number of independent variables, shown in Eq. (2).

Eq(2): $$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots + \beta_p X_p + e$$

The subscript notation on each X and β denote each independent variable and its regression coefficient. There are p independent variables and p + 1 parameters to be estimated. The least squares method of estimation requires that estimates of the p + 1 parameters are minimized, shown in Eq. (3).

Eq(3): $$SS_{Res} = \sum e_i^2 = \sum (Y_i - \hat{Y}_i)^2 = \sum (Y_i - \beta_0 - \beta_1 X_1 - \beta_2 X_2 - \beta_3 X_3 - \ldots - \beta_p X_p)^2$$

The estimated values of the parameter that minimize sums of squares of the residuals, denoted (SS_{Res}), are obtained by derivation of SS_{Res} with respect to each β in turn equal to zero. This gives p + 1 normal equations that must be solved simultaneously to obtain the least squares estimates of the parameters [1]. In the simplest form of a multiple linear model with two independent variables three equations must be solved simultaneously to obtain the estimates of β_0, β_1, and β_2, shown in Eq. (4-6).

Manuscript received April 5, 2024
Manuscript revised April 20, 2024
https://doi.org/10.22937/IJCSNS.2024.24.4.5

Eq(4):

$$\frac{\partial SS_{Res}}{\partial \beta_0} = -2\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_1 - \beta_2 X_2) = 0$$

$$\sum_{i=1}^{n} Y_i = n\beta_0 + \beta_1 \sum_{i=1}^{n} X_1 + \beta_2 \sum_{i=1}^{n} X_2$$

Eq(5):

$$\frac{\partial SS_{Res}}{\partial \beta_1} = -2\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_1 - \beta_2 X_2)\, X_1 = 0$$

$$\sum_{i=1}^{n} X_1 Y_i = \beta_0 \sum_{i=1}^{n} X_1 + \beta_1 \sum_{i=1}^{n} X_1^2 + \beta_2 \sum_{i=1}^{n} X_1 X_2$$

Eq(6):

$$\frac{\partial SS_{Res}}{\partial \beta_2} = -2\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_1 - \beta_2 X_2)\, X_2 = 0$$

$$\sum_{i=1}^{n} X_1 Y_i = \beta_0 \sum_{i=1}^{n} X_1 + \beta_1 \sum_{i=1}^{n} X_1 X_2 + \beta_2 \sum_{i=1}^{n} X_2^2$$

It is clear that estimating parameters through algebraic methods becomes increasingly difficult as the number of dependent and independent variables increases. For this reason, matrices are used to estimate parameters in complicated general linear models.

2. Matrices in Regression

Regression provides the basic machinery that all general linear models are based upon. Four matrices are needed to express the regression model: 1) the observed dependent variable matrix, which is a n × 1 column vector, 2) the observed independent variable, which is a n × p+1 matrix where the first column consists of ones, 3) the parameters β matrix, which is a p+1 × 1 column vector, and 4) the n × 1 column vector of random errors. The linear model can be written as shown in Eq. (7).

Eq(7):

$$Y = X\beta + E$$

$$\begin{bmatrix} y_i \\ y_2 \\ \vdots \\ y_i \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & x_{13} & \dots & x_{1p} \\ 1 & x_{21} & x_{22} & x_{23} & \dots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{i1} & x_{i2} & x_{i3} & \dots & x_{ip} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & x_{n3} & \dots & x_{np} \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_i \\ \vdots \\ b_n \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_i \\ \vdots \\ e_n \end{bmatrix}$$

Each column of X contains the values for an independent variable. The elements of a row of X are the coefficients on the corresponding parameters in β. The ordinary least square (OLS) estimate B of β minimizes SS_{Res}, such that under OLS assumptions it can simply be proved that $B = (X'X)^{-1}X'Y$. Using simple matrix calculations the B matrix elements can be estimated. By looking at the elements of *X'X* and *X'Y*. Equation (8) illustrates how each row corresponds to its estimated parameter, and Eq. (9-10) shows the OLS equivalent. In simple regression the first row corresponds to B_1 parameter, and is equal to $\frac{\sum_{i=1}^{n}(x_{i1}-x_1)(y_i-\bar{Y})}{\sum_{i=1}^{n}(x_{i1}-x_1)^2}$.

Eq(8):

$$X'X = \begin{bmatrix} n & \sum x_{i1} & \sum x_{i2} & \dots & \sum x_{ij} & \dots & \sum x_{ip} \\ \sum x_{i1} & \sum x_{i1}^2 & \sum x_{i1}x_{i2} & \dots & \sum x_{i1}x_{ij} & \dots & \sum x_{i1}x_{ip} \\ \sum x_{i2} & \sum x_{i1}x_{i2} & \sum x_{i2}^2 & \dots & \sum x_{i2}x_{ij} & \dots & \sum x_{i2}x_{ip} \\ \vdots & \vdots & \vdots & \dots & \vdots & \dots & \vdots \\ \sum x_{ij} & \vdots & \vdots & \dots & \sum x_{ij}^2 & \dots & \sum x_{ij}x_{ip} \\ \vdots & \vdots & \vdots & \dots & \vdots & \dots & \vdots \\ \sum x_p & \dots & \dots & \dots & \dots & \dots & \sum x_p^2 \end{bmatrix}$$

$$X'Y = \begin{bmatrix} \sum y_i \\ \sum x_{i1}y_i \\ \sum x_{i2}y_i \\ \vdots \\ \sum x_{ij}y_i \\ \vdots \\ \sum x_{ip}y_i \end{bmatrix}$$

Eq(9):

$$X'X = \begin{bmatrix} n & \sum(x_{i1}-x_1) & \dots & \sum(x_{ip}-x_p) \\ \sum(x_{i1}-x_1) & \sum(x_{i1}-x_1)^2 & \dots & \sum(x_{i1}-x_1)(x_{ip}-x_p) \\ \vdots & \vdots & \dots & \vdots \\ \sum(x_{ip}-x_p) & \sum(x_{i1}-x_1)(x_{ip}-x_p) & \dots & \sum(x_{ip}-x_p)^2 \end{bmatrix}$$

Eq(10):

$$\boldsymbol{X'Y} = \begin{bmatrix} \sum(\boldsymbol{y_i} - \overline{\boldsymbol{Y}}) \\ \sum(\boldsymbol{x_{i1}} - \boldsymbol{x_1})(\boldsymbol{y_i} - \overline{\boldsymbol{Y}}) \\ \vdots \\ \sum(\boldsymbol{x_{ip}} - \boldsymbol{x_p})(\boldsymbol{y_1} - \overline{\boldsymbol{Y}}) \end{bmatrix}$$

3. Matrices in univariate analysis of variance (ANOVA)

Understanding regression facilitates transition to univariate and multivariate analysis of variance. The general linear model does not change with the change of the experiment design, where it is comprised of Data = Model + Error [3]. However, the research questions determine the type of analysis that is best suited for a study. Univariate analysis of variance builds on the regression matrix equation with one major difference. The independent variables in regression should include at least one continuous variable [4], whereas in analysis of variance the independent variable is a constructed matrix that determines group inclusion.

In multiple regressions, a number of models that include different sets of variables are compared. Analysis of variance (ANOVA) is used to understand the variability in the models and choose the best fitting model. ANOVA in regression compares two models, the first with the complete set of independent variables (Y=XB+E) to a model with no independent variables (the intercept-only model $Y=B_0$) [4]. Table 1 shows the ANOVA in matrix formation.

Table 1. ANOVA table in matrix form.

	Sum of Squares (SS)	Degrees of Freedom	Mean Square (MS)	F Statistic
SS_{reg}[a]	$B'X'Y - \left(\frac{1}{n}\right)Y'Y$	$p-1$	$MS_{reg} = \frac{SS_{reg}}{p-1}$	$\frac{MS_{reg}}{MS_e}$
SS_E	$Y'Y - B'X'Y$	$n-p$	$MS_e = \frac{SS_e}{n-p}$	
SS_{total}	$Y'Y - \left(\frac{1}{n}\right)Y'Y$	$n-1$		

[a] Table notations: reg (between subjects), E (within subjects), total (total subjects), p (number of groups), n (sample size)

ANOVA with two factors is modeled as $Y_{ij} = \mu + \alpha + \beta + \alpha\beta + e_{ij}$, where α is the mean for effect of level *i* of factor A such that *i* = 1,2,3, … , a (if there are *a* levels of factor A), β is the mean for effect of level *j* of factor B such that j= 1, 2, 3,…, b (if that are *b* levels of factor B), and μ is the grand mean or the intercept. In ANOVA we are interested in the contribution of all levels of both factors on any given observation. The design matrix is the means of which we are able to model ANOVA levels, and compare means all the while using the general linear model.

We can recall the matrix notation of multiple regression, Y = Xβ + E. In ANOVA, Y is the observations column vector, X represents the design matrix, β is the mean column vector that includes the means for each level of each factor, and E is the column vector representing the difference between the observed and predicted scores. This makes the calculations for any size ANOVA simple. Table 2 shows an example of a 2 by 3 factorial ANOVA with 2 subjects (Y_i) in each cell.

Table 2. ***Example of a 2 by 3 factorial ANOVA***.

	A_1	A_2	A_3
B_1	Y_1, Y_2	Y_3, Y_4	Y_5, Y_6
B_2	Y_7, Y_8	Y_9, Y_{10}	Y_{11}, Y_{12}

The design matrix is essential in understanding matrix formation in univariate ANOVA. The design matrix can be represented in many different forms depending on the method of coding that is used (i.e. dummy, contrast or effect coding). The results of the analysis do not change with different coding systems, however the interpretation of the results must be reflective of the type of coding [5]. For this specific example, the integral role of dummy coding on the linear variation of ANOVA is demonstrated, where factor A has two levels and factor B has one level. The vectors in the matrix shown in Eq. (11) are denoted as following, Y_i is the observed dependent variable vector, X_0 is the vector for the grand mean or the constant of the linear model, X_1 is the dummy code vector for B, X_2 is the dummy code vector for level 1 for A (A_1), X_3 is the dummy code vector for level 2 for A (A_2), X_4

is the dummy code vector for the interaction between B and level 1 for A (BA_1), and X_5 is the dummy code vector for the interaction between B and level 2 for A (BA_2).

Eq(11):

$$\mathbf{Y=X\beta+E}$$

$$[\boldsymbol{Y_i}] = [\boldsymbol{X_0} \quad \boldsymbol{X_1} \quad \boldsymbol{X_2} \quad \boldsymbol{X_3} \quad \boldsymbol{X_4} \quad \boldsymbol{X_5}]\ [\boldsymbol{\beta}] + [\boldsymbol{E}]$$

$$\begin{bmatrix} \boldsymbol{Y_1} \\ \boldsymbol{Y_2} \\ \boldsymbol{Y_3} \\ \boldsymbol{Y_4} \\ \boldsymbol{Y_5} \\ \boldsymbol{Y_6} \\ \boldsymbol{Y_7} \\ \boldsymbol{Y_8} \\ \boldsymbol{Y_9} \\ \boldsymbol{Y_{10}} \\ \boldsymbol{Y_{11}} \\ \boldsymbol{Y_{12}} \end{bmatrix} = \begin{bmatrix} \mathbf{1} & \mathbf{1} & \mathbf{1} & \mathbf{0} & \mathbf{1} & \mathbf{0} \\ \mathbf{1} & \mathbf{1} & \mathbf{1} & \mathbf{0} & \mathbf{1} & \mathbf{0} \\ \mathbf{1} & \mathbf{1} & \mathbf{0} & \mathbf{1} & \mathbf{0} & \mathbf{1} \\ \mathbf{1} & \mathbf{1} & \mathbf{0} & \mathbf{1} & \mathbf{0} & \mathbf{1} \\ \mathbf{1} & \mathbf{1} & \mathbf{-1} & \mathbf{-1} & \mathbf{-1} & \mathbf{-1} \\ \mathbf{1} & \mathbf{1} & \mathbf{-1} & \mathbf{-1} & \mathbf{-1} & \mathbf{-1} \\ \mathbf{1} & \mathbf{-1} & \mathbf{1} & \mathbf{0} & \mathbf{-1} & \mathbf{0} \\ \mathbf{1} & \mathbf{-1} & \mathbf{1} & \mathbf{0} & \mathbf{-1} & \mathbf{0} \\ \mathbf{1} & \mathbf{-1} & \mathbf{0} & \mathbf{1} & \mathbf{0} & \mathbf{-1} \\ \mathbf{1} & \mathbf{-1} & \mathbf{0} & \mathbf{1} & \mathbf{0} & \mathbf{-1} \\ \mathbf{1} & \mathbf{-1} & \mathbf{-1} & \mathbf{-1} & \mathbf{-1} & \mathbf{1} \\ \mathbf{1} & \mathbf{-1} & \mathbf{-1} & \mathbf{-1} & \mathbf{-1} & \mathbf{1} \end{bmatrix} \begin{bmatrix} \boldsymbol{\mu} \\ \boldsymbol{B} \\ \boldsymbol{A_1} \\ \boldsymbol{A_2} \\ \boldsymbol{BA_1} \\ \boldsymbol{BA_2} \end{bmatrix} + \begin{bmatrix} \boldsymbol{E_1} \\ \boldsymbol{E_2} \\ \boldsymbol{E_3} \\ \boldsymbol{E_4} \\ \boldsymbol{E_5} \\ \boldsymbol{E_6} \\ \boldsymbol{E_7} \\ \boldsymbol{E_8} \\ \boldsymbol{E_9} \\ \boldsymbol{E_{10}} \\ \boldsymbol{E_{11}} \\ \boldsymbol{E_{12}} \end{bmatrix}$$

From Table 1 we know how SS between subjects (corresponds to SS_{reg}) and SS within subjects (corresponds to SS_E) are calculated in matrix form. However, the matrix formation in regression needs a little more adjustment to be applied in ANOVA. Most importantly, the regression matrix formation does not distinguish proportions of the effect that are attributable to A, B, or their interaction. To calculate the contribution of the variance for each effect through matrices, we recalculate the regression for 3 reduced models [4]. The reduced models are calculated by dropping certain columns in the design matrix to produce a new design matrix that is a subset of the original one. Because of that the corresponding predictors will not be calculated. For example, if we dropped the interaction columns, shown in Eq. (11), BA_1 and BA_2, we would be deleting the predictors containing information about the interaction but keeping the predictors and sources of variance SS_{AB} between factors A and B. Similarly by dropping the B columns, parameters and sources of variance $SS_{B\ \&\ BA}$ between factors A and interaction BA will be estimated; and by dropping A parameters related to B and BA will be estimated. Now that we calculated all possible subsets using the formulas in Table 1 we can find the sums of squares for each main effect by subtraction as follows:

- $SS_{BA} = SS_{total} - SS_{A\ \&\ B}$
- $SS_A = SS_{total} - SS_{B\ \&\ BA}$
- $SS_B = SS_{total} - SS_{A\ \&\ BA}$

To compare models, the F statistic then can be computed by dividing the effect of interest by the mean square error (MSE) for the full model, taking into account the degrees of freedom (df) for each source of variance. For example, to examine the effect of factor A, the F statistic can be calculated as shown in Eq. (12).

Eq(12):

$$\mathbf{F_{(df\ full\ model,\ df\ reduced\ model)}} = \frac{\boldsymbol{SSA}}{\boldsymbol{df_{full\ model} - df_{reduced\ model}}} / \boldsymbol{MSE}$$

Matrices in multivariate analysis of variance (MANOVA)

Dependent variables in multivariate analysis should be inter-correlated to form a system of variables that are of interest. MANOVA research questions address whether an overall effect of an independent variable exists. This system should have a conceptual link in order for the effects to have substantive meaning. MANOVA is the generalized form of ANOVA, where the model contains a matrix of dependent variables (Y), a design matrix (X), a parameter matrix (β) and an error matrix (E), such that it follows the same general linear model, Y = Xβ + E. Each observed score is a matrix of quantitative scores on the dependent variables.

The calculation differences between MANOVA and ANOVA relates to the nature of sums of squares, where it is a scalar in ANOVA but a matrix in MANOVA. A sums of squares and cross product matrix is the same as sums of squares in ANOVA and are calculated by taking the squared differences from each observed score and its appropriate mean [6]. Calculating MANOVA is simple and straightforward and requires basic knowledge of matrix algebra, however it gains complexity as the sample size for each dependent variable increases and as the number of dependent variable increases. As in ANOVA, three basic sums of squares and cross product matrices are needed to test a hypothesis. Similar to a one-way ANOVA, a one-way MANOVA requires calculating SS_{total},

$SS_{between}$, and SS_{within}, where $SS_{total} = SS_{between} + SS_{within}$.

SS_{within} can be calculated by taking the sum of the sums of squares and cross product for each dependent variable, such that $SS_{within} = W_1 + W_2 + \ldots + W_j$, where j = 1, … ,n is a subscript that represents the number of levels or groups. SS_{within} is a square matrix of a rank equal to the number of dependent variables. For example, if there are i = 1, … , n dependent variables the W_j matrix is shown in Eq. (13). The subscript k represents the observed score, where k = 1, … , n, and y_{kij} is the observed score in the ith dependent variable and jth group, $\bar{Y}_{ij}$ is the mean of the ith variable in the jth group.

Eq(13):

$$\boldsymbol{W_i} = \begin{bmatrix} \boldsymbol{ss_{11}} & \boldsymbol{ss_{12}} & \ldots & \boldsymbol{ss_{1i}} \\ \boldsymbol{ss_{21}} & \boldsymbol{ss_{22}} & \ldots & \boldsymbol{ss_{2i}} \\ \vdots & \vdots & \ldots & \vdots \\ \boldsymbol{ss_{i1}} & \ldots & \ldots & \boldsymbol{ss_{ii}} \end{bmatrix}$$

where,

$$\boldsymbol{ss_{ii}}\ (\boldsymbol{diagonal}) = \sum_i \left(\boldsymbol{y_{kij}} - \boldsymbol{Y_{ii}}\right)^2$$

$$\boldsymbol{ss_{i(i-1)}}\ (\boldsymbol{off-diagonal}) = \sum_i \left(\boldsymbol{y_{k(i-1)j}} - \boldsymbol{\bar{Y}_{(i-1)j}}\right)\left(\boldsymbol{y_{kij}} - \boldsymbol{\bar{Y}_{ij}}\right)$$

Similar to SS_{within}, the $SS_{between}$ matrix is square of rank equal to the number of dependent variables. The elements on the diagonal of the matrix are equal to $\boldsymbol{b_{ii} = \Sigma_j\, n_j\, (\bar{Y}_{ij} - \bar{Y}_i)^2}$, and the off diagonal elements are equal to $\boldsymbol{b_{(i-1)i} = b_{i(i-1)} = \Sigma_j\, n_j\, (\bar{Y}_{(i-1)j} - \bar{Y}_i)\,(\bar{Y}_{ij} - \bar{Y}_i)}$, where n_j is the number of subjects in group j, $\bar{Y}_{ij}$ is the mean of variable i in group j, and $\bar{Y}_i$ is the mean of variable i.

Let us extend this example to a two-way MANOVA. In this case, $SS_{between}$ must be partitioned to take into account the variance that is attributed to each independent variable [6]. Sums of squares and cross product of between-subjects is the summation of the effects of each independent variable and their interaction. To make the calculation of sums of squares and cross product easier, Tabachnik and Fidell (2007) identified a group of matrices that are of a rank equal to the number of dependent variables. For example, let us consider a two-way MANOVA that has two independent variables A (two levels) and B (two levels), and two dependent variables Y_1 and Y_2, elements of these matrices are means as shown in Table 4.

Table 4. ***Important MANOVA matrices.***

A_1	A_2	B_1	B_2	Grand Mean (GM)
$\bar{Y}_{11.}$ (mean for DV1 in A1) $\bar{Y}_{21.}$ (mean for DV2 in A1)	$\bar{Y}_{12.}$ (mean for DV1 in A2) $\bar{Y}_{22.}$ (mean for DV2 in A2)	$\bar{Y}_{1.1}$ (mean for DV1 in B1) $\bar{Y}_{2.1}$ (mean for DV2 in B1)	$\bar{Y}_{1.2}$ (mean for DV1 in B2) $\bar{Y}_{2.2}$ (mean for DV2 in B2)	$\bar{Y}_1$ (mean for DV1 on all levels) $\bar{Y}_2$ (mean for DV1 on all levels)

For this case $SS_{total} = SS_A + SS_B + SS_{AB} + SS_{within}$ where,

- $SS_A = n_k \sum_k (A_k - GM)(A_k - GM)'$, k is the levels of A, k = 1, 2 with degrees of freedom (df_a = k - 1).
- $SS_B = n_m \sum_m (B_m - GM)(B_m - GM)'$, m is the levels of B, m = 1, 2 with degrees of freedom (df_b = m - 1).
- $SS_{AB} = [n_{km} \sum_k \sum_m (A_k B_m - GM)(A_k B_m - GM)'] - SS_A - SS_B$,with degrees of freedom ($df_{ab} = df_a \times df_b$).
- $SS_{within} = \sum_i \sum_k \sum_m (y_{ikm} - A_k B_m)(y_{ikm} - A_k B_m)'$, i = 1, … , n is the observed score and with degrees of freedom (df_{error} = a × b × (n-1)).
- $SS_{total} = \sum_i \sum_k \sum_m (y_{ikm} - GM)(y_{ikm} - GM)'$, with degrees of freedom (df_{total} = a×b×n - 1).

The null hypothesis tests whether the means across levels of effects and interactions are equal. For a two-way MANOVA there are three null hypothesis one that tests the effect of A: H_0 $\mu_{a1}=\mu_{a2}$, the second tests the effects of B: H_0 $\mu_{b1}=\mu_{b2}$, and the last tests the effects of the interaction: H_0 $\mu_{ab1}=\mu_{ab2}$. The alternative hypothesis states that the means are not equal, in the case where there is more than one level in a independent variable the alternative hypothesis states that at least one mean is different. Wilks' lambda and other statistics (such as Hotelling's T and Roy's largest root) are used to test the main effects and interactions. For the purpose of this paper, Wilks' lambda will be examined. Wilks' lambda is the "ratio of the determinant of the error cross-product matrix

to the determinant of the sum of the error and effect cross-product matrices" (p. 260, Tabachnik & Fidell, 2007), as shown in Eq. (14). To evaluate the significance of Wilks' lambda a statistic that approximately follows an F-distribution and closely fits lambda is calculated as in Eq. (15). The approximate F statistic is tested for significance using the regular F-tables at a chosen significance level.

Eq(14):

$$\Lambda = Det(SS_{within})/Det(SS_{effect} + SS_{within})$$

Eq(15):

$$Approximate\ F_{(df1,df2)} = \left(\frac{1-y}{y}\right)\left(\frac{df_2}{df_1}\right)$$

where $y = \Lambda^{1/s}$, and s = min (number of parameters estimated p, df_{effect})

$df_1 = p \times df_{effect}$

$df_2 = s\,[df_{error} - (p - df_{effect} + 1)/2] - (df_1 - 2)/2$

To illustrate a two-way MANOVA, a dataset extracted from [7] is used as shown in Table 5. This modified (extracted from a larger dataset) dataset is based on a study that examines the uses of program evaluation. A measure was designed to extract perceptions of stakeholders in a project surrounding the usefulness of program evaluations. The original measure contains 73 items that elicit responses on use of evaluation findings, use of evaluation process, level of stakeholder involvement and factors that effect uses of the evaluation. The items are measured on a 5-point Likert scale, that are treated here as continuous [8]. For the purpose of this paper two variables which measure evaluation findings are considered as dependent variables:

1. I feel the project was enhanced after the first year of evaluation feedback (Inst1).
2. I feel the project was enhanced after the first evaluation analysis feedback (Inst2).

There are also two grouping variables gender (2 levels) and power (3 levels) as shown in Table 5.

Table 5. ***Small-sample data for illustration of MANOVA.***

	Power					
	Principle Investigator (PI)		Post-doc Fellow (PDF)	Trainee		
	Inst1	Inst2	Inst1	Inst1		
			Inst2	Inst2		
Female	1	2	2	3	3	3
	3	2	2	3	4	4
	4	4	1	2	2	3
Male	2	3	3	3	3	3
	2	2	2	3	3	3
	1	3	4	3	4	4

The first step is to find the mean matrices, as shown in Table 6. The second step is to fill out the MANOVA table. This needs basic knowledge of matrix algebra, or the use of a matrix calculator, as shown in Table 7. The third and last step is to calculate Wilks' lambda and approximate F-statistic to test significance for each effect, as shown in Table 8. With the use of Wilks' criterion, the instrumental use of findings in the evaluation of the Working on Walls project were significantly affected by gender $F(2,11) = 0.322$, $p < 0.05$, but not by power or the interaction between power and gender.

Table 6. ***First step in MANOVA: Calculating matrices***

G_{female}	G_{male}	P_{PI}	P_{PDF}	P_T	GM
2.44	2.67	1.44	1.56	2.11	1.77
2.67	3	1.67	1.89	2.22	1.93

Table 7. ***Sums of square (SS) and cross product (CP) matrices and their corresponding degrees of freedom (df).***

Sources of variance	SS and CP	df
SS_{gender}	[,1] [,2] [1,]13.39 14.27 [2,]14.27 15.23	1
SS_{power}	[,1] [,2] [1,] 1.53 1.15 [2,] 1.15 0.92	2
$SS_{gender \times power}$	[,1] [,2] [1,] 6.01 2.09 [2,] 2.09 2.81	2
SSwithin	[,1] [,2] [1,] 10.68 6.37 [2,] 6.37 7.37	12
SStotal	[,1] [,2] [1,] 31.62 23.88 [2,] 23.88 26.33	17

Table 8. ***Test of significance of effects.***

Source of variance	Wilks lambda	df1,df2	Approximate-F	Significance
SSG	0.322	2,11	11.58	0.002
SSP	0.854	4,22	0.4523	0.769
SSGP	0.388	4,22	3.33	0.282

4. Conclusion

This paper is dedicated to understanding the intricacies of matrix formation and computation within the domains of univariate and multivariate analysis of variance. While the fundamental mathematical procedures underlying modeling remain consistent, variations emerge concerning statistical assumptions, diagnostic assessments, and the nature of data. By extending the mathematical techniques employed in the simplest form of regression, known as Ordinary Least Squares (OLS), to the broader spectrum of general linear modeling, we establish a seamless link between general and generalized linear models. Understanding the matrices employed to represent the general linear model necessitates a fundamental grasp of matrix algebra, rendering parameter estimation formulas more accessible. As model complexity increases, manual calculation of matrix algebraic expressions becomes progressively challenging. Fortunately, advances of computer software has eliminated this challenge. In pursuit of our objective, this paper successfully clarifies the interconnections between diverse analytical approaches tailored to essentially unique research inquiries.

This paper has unveiled the intrinsic connection between regression and GLMs, highlighting the significance of matrices as the computational backbone of these models. The bridge constructed between regression models and univariate and multivariate applications underscores the versatility and adaptability of the generalized linear model framework.

The univariate analysis of variance and regression methods discussed in this paper laid the groundwork for comprehending the matrix algebra involved in GLMs. Subsequently, the extension to multivariate analysis of variance allowed for a broader and more encompassing perspective on statistical calculations. It is essential that a solid grasp of matrix operations and their integration into GLMs is invaluable. The insights shared herein not only contribute to a deeper understanding of the theoretical foundations but also provide practical utility for researchers in their quest to extract meaningful information from data.

References

[1] Rawlings, J., Dickey, D., & Pantula, S. (1998). *Applied regression analysis : A research tool.* Springer.

[2] Herr, D. G. (1980). On the history of the use of geometry in the general linear model. *The American Statistician, 34*(1), 43.

[3] Zumbo, B. (2012). *The analysis of data arising from an experiment or quasi-experiment* . Unpublished manuscript.

[4] Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple Regression/Correlation analysis for the behavioral sciences* (Third Edition ed.). Mahwah, New Jersey: Lawrence Erlbuam Associates, Inc.

[5] Olive, D. (2012). MANOVA. *Robust multivariate analysis* (pp. 169)

[6] Tabachnik, B. G., & Fidell, L. S. (2007). *Using multivariate statistics* (Sixth ed.). New Jersey, United States: Pearson Education, Inc.

[7] Alkhalaf, A. (2012). The relationship among Process Use, Findings use and Stakeholder Involvement in Evaluation. [Master's Thesis, University of British Columbia].

[8] Robitizsch, A. (2020). Why Ordinal Variables Can (Almost) Always Be Treated as Continuous Variables: Clarifying Assumptions of Robust Continuous and Ordinal Factor Analysis Estimation Methods. Frontiers in Education, 5, https://www.frontiersin.org/articles/10.3389/feduc.2020.589965.