# ANOVA

## Model and Matrix Computations

### Notation

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

| | |
|---|---|
| $N$ | Number of cases |
| $F$ | Number of factors |
| $CN$ | Number of covariates |
| $k_i$ | Number of levels of factor $i$ |
| $Y_k$ | Value of the dependent variable for case $k$ |
| $Z_{jk}$ | Value of the $j$th covariate for case $k$ |
| $w_k$ | Weight for case $k$ |
| $W$ | Sum of weights of all cases |

### The Model

A linear model with covariates can be written in matrix notation as

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{ZC} + \mathbf{e} \tag{1}$$

where

| | |
|---|---|
| $\mathbf{Y}$ | $N \times 1$ vector of values of the dependent variable |
| $\mathbf{X}$ | Design matrix $(N \times p)$ of rank $q < p$ |
| $\beta$ | Vector of parameters $(p \times 1)$ |
| $\mathbf{Z}$ | Matrix of covariates $(N \times CN)$ |
| $\mathbf{C}$ | Vector of covariate coefficients $(CN \times 1)$ |
| $\mathbf{e}$ | Vector of error terms $(N \times 1)$ |

## Constraints

To reparametrize equation **(1)** to a full rank model, a set of non-estimable conditions is needed. The constraint imposed on non-regression models is that all parameters involving level 1 of any factor are set to zero.

For regression model, the constraints are that the analysis of variance parameters estimates for each main effect and each order of interactions sum to zero. The interaction must also sum to zero over each level of subscripts.

For a standard two way ANOVA model with the main effects $\alpha_i$ and $\beta_j$, and interaction parameter $\gamma_{ij}$, the constraints can be expressed as

$$\begin{array}{ll} \alpha_1 = \beta_1 = \gamma_{1j} = \gamma_{i1} = 0 & \text{non – regression} \\ \alpha_\bullet = \beta_\bullet = \gamma_{i\bullet} = \gamma_{\bullet j} = 0 & \text{regression} \end{array}$$

where $\bullet$ indicates summation.

## Computation of Matrices

### $\mathbf{X'X}$

#### Non-regression Model

The $\mathbf{X'X}$ matrix contains the sum of weights of the cases that contain a particular combination of parameters. All parameters that involve level 1 of any of the factors are excluded from the matrix. For a two-way design with $k_1 = 2$ and $k_2 = 3$, the symmetric matrix would look like the following:

$$\begin{array}{c|ccccc} & \alpha_2 & \beta_2 & \beta_3 & \gamma_{22} & \gamma_{23} \\ \hline \alpha_2 & N_{2\bullet} & N_{22} & N_{23} & N_{22} & N_{23} \\ \beta_2 & & N_{\bullet 2} & 0 & N_{22} & 0 \\ \beta_3 & & & N_{\bullet 3} & 0 & N_{23} \\ \gamma_{22} & & & & N_{22} & 0 \\ \gamma_{23} & & & & & N_{23} \end{array}$$

The elements $N_{i\bullet}$ or $N_{\bullet j}$ on the diagonal are the sums of weights of cases that have level $i$ of $\alpha$ or level $j$ of $\beta$. Off-diagonal elements are sums of weights of cases cross-classified by parameter combinations. Thus, $N_{\bullet 3}$ is the sum of weights of cases in level 3 of main effect $\beta_3$, while $N_{22}$ is the sum of weights of cases with $\alpha_2$ and $\beta_2$.

### Regression Model

A row of the design matrix $\mathbf{X}$ is formed for each case. The row is generated as follows:

If a case belongs to one of the 2 to $k_i$ levels of factor *i*, a code of 1 is placed in the column corresponding to the level and 0 in all other $k_i - 1$ columns associated with factor *i*. If the case belongs in the first level of factor *i*, –1 is placed in *all* the $k_i - 1$ columns associated with factor *i*. This is repeated for each factor. The entries for the interaction terms are obtained as products of the entries in the corresponding main effect columns. This vector of dummy variables for a case will be denoted as $d(i), i = 1, \ldots, NC$, where $NC$ is the number of columns in the reparametrized design matrix. After the vector $\mathbf{d}$ is generated for case *k*, the *ij*th cell of $\mathbf{X'X}$ is incremented by $d(i)d(j)w_k$, where $i = 1, \ldots, NC$ and $j \geq i$.

### Checking and Adjustment for the Mean

After all cases have been processed, the diagonal entries of $\mathbf{X'X}$ are examined. Rows and columns corresponding to zero diagonals are deleted and the number of levels of a factor is reduced accordingly. If a factor has only one level, the analysis will be terminated with a message. If the first specified level of a factor is missing, the first non-empty level will be deleted from the matrix for non-regression model. For regression designs, the first level cannot be missing. All entries of $\mathbf{X'X}$ are subsequently adjusted for means.

The highest order of interactions in the model can be selected. This will affect the generation of $\mathbf{X'X}$. If none of these options is chosen, the program will generate the highest order of interactions allowed by the number of factors. If sub-matrices corresponding to main effects or interactions in the reparametrized model are not of full rank, a message is printed and the order of the model is reduced accordingly.

## Cross-Product Matrices for Continuous Variables

Provisional means algorithm are used to compute the adjusted-for-the-means cross-product matrices.

### Matrix of Covariates $\mathbf{Z'Z}$

The covariance of covariates *m* and *l* after case *k* has been processed is

$$\mathbf{Z'Z}_{ml}(k) = \mathbf{Z'Z}_{ml}(k-1) + \frac{w_k\left(W_k Z_{lk} - \sum_{j=1}^{k} w_j Z_{lj}\right)\left(W_k Z_{mk} - \sum_{j=1}^{k} w_j Z_{mj}\right)}{W_k W_{k-1}}$$

where $W_k$ is the sum of weights of the first *k* cases.

### The Vector $\mathbf{Z'Y}$

The covariance between the *m*th covariate and the dependent variable after case *k* has been processed is

$$\mathbf{Z'Y}_m(k) = \mathbf{Z'Y}_m(k-1) + \frac{w_k\left(W_k Y_k - \sum_{j=1}^{k} w_j Y_j\right)\left(W_k Z_{mk} - \sum_{j=1}^{k} w_j Z_{mj}\right)}{W_k W_{k-1}}$$

### The Scalar $\mathbf{Y'Y}$

The corrected sum of squares for the dependent variable after case *k* has been processed is

$$\mathbf{Y'Y}(k) = \mathbf{Y'Y}(k-1) + \frac{w_k\left(W_k Y_k - \sum_{j=1}^{k} w_j Y_j\right)^2}{W_k W_{k-1}}$$

### The Vector $\mathbf{X'Y}$

$\mathbf{X'Y}$ is a vector with *NC* rows. The *i*th element is

$$\mathbf{X'Y}_i = \sum_{k=1}^{N} Y_k w_k \delta_k \text{ ,}$$

where, for non-regression model, $\delta_k = 1$ if case *k* has the factor combination in column *i* of $\mathbf{X'X}$; $\delta_k = 0$ otherwise. For regression model, $\delta_k = d(i)$, where $d(i)$ is the dummy variable for column *i* of case *k*. The final entries are adjusted for the mean.

### Matrix $\mathbf{X'Z}$

The (*i*, *m*)th entry is

$$\mathbf{X'Z}_{im} = \sum_{k=1}^{N} Z_{mk} w_k \delta_k$$

where $\delta_k$ has been defined previously. The final entries are adjusted for the mean.

# Computation of ANOVA Sum of Squares

The full rank model with covariates

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{ZC} + \mathbf{e}$$

can also be expressed as

$$\mathbf{Y} = \mathbf{X}_k\mathbf{b}_k + \mathbf{X}_m\mathbf{b}_m + \mathbf{ZC} + \mathbf{e}$$

where **X** and **b** are partitioned as

$$\mathbf{X} = \left[\mathbf{X}_k \,|\, \mathbf{X}_m\right] \textbf{ and } \beta = \left[\frac{\mathbf{b}_k}{\mathbf{b}_m}\right].$$

The normal equations are then

$$\begin{bmatrix} \mathbf{Z}'\mathbf{Z} & \mathbf{Z}'\mathbf{X}_k & \mathbf{Z}'\mathbf{X}_m \\ \mathbf{X}_k'\mathbf{Z} & \mathbf{X}_k'\mathbf{X}_k & \mathbf{X}_k'\mathbf{X}_m \\ \mathbf{X}_m'\mathbf{Z} & \mathbf{X}_m'\mathbf{X}_k & \mathbf{X}_m'\mathbf{X}_m \end{bmatrix} \begin{bmatrix} \hat{\mathbf{C}} \\ \hat{\mathbf{b}}_k \\ \hat{\mathbf{b}}_m \end{bmatrix} = \begin{bmatrix} \mathbf{Z}'\mathbf{Y} \\ \mathbf{X}_k'\mathbf{Y} \\ \mathbf{X}_m'\mathbf{Y} \end{bmatrix} \qquad \textbf{(2)}$$

The normal equations for any reduced model can be obtained by excluding those entries from equation **(2)** corresponding to terms that do not appear in the reduced model.

Thus, for the model excluding $\mathbf{b}_m$,

$$\mathbf{Y} = \mathbf{X}_k\mathbf{b}_k + \mathbf{ZC} + \mathbf{e}$$

the solution to the normal equation is:

$$\begin{bmatrix} \tilde{\mathbf{C}} \\ \tilde{\mathbf{b}}_k \end{bmatrix} = \begin{bmatrix} \mathbf{Z}'\mathbf{Z} & \mathbf{Z}'\mathbf{X}_k \\ \mathbf{X}_k'\mathbf{Z} & \mathbf{X}_k'\mathbf{X}_k \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{Z}'\mathbf{Y} \\ \mathbf{X}_k'\mathbf{Y} \end{bmatrix} \qquad \textbf{(3)}$$

The sum of squares due to fitting the complete model (explained *SS*) is

$$R(\mathbf{C}, \mathbf{b}_k, \mathbf{b}_m) = \left[\hat{\mathbf{C}}', \hat{\mathbf{b}}_k', \hat{\mathbf{b}}_m'\right] \begin{bmatrix} \mathbf{Z}'\mathbf{Y} \\ \mathbf{X}_k'\mathbf{Y} \\ \mathbf{X}_m'\mathbf{Y} \end{bmatrix} = \hat{\mathbf{C}}'\mathbf{Z}'\mathbf{Y} + \hat{\mathbf{b}}_k'\mathbf{X}_k'\mathbf{Y} + \hat{\mathbf{b}}_m'\mathbf{X}_m'\mathbf{Y}$$

For the reduced model, it is

$$R(\mathbf{C}, \mathbf{b}_k) = \left[\tilde{\mathbf{C}}', \tilde{\mathbf{b}}_k'\right] \begin{bmatrix} \mathbf{Z}'\mathbf{Y} \\ \mathbf{X}_k'\mathbf{Y} \end{bmatrix} = \tilde{\mathbf{C}}'\mathbf{Z}'\mathbf{Y} + \tilde{\mathbf{b}}_k'\mathbf{X}_k'\mathbf{Y}$$

The residual (unexplained) sum of squares for the complete model is $RSS = \mathbf{Y}'\mathbf{Y} - R(\mathbf{C}, \mathbf{b}_k, \mathbf{b}_m)$ and similarly for the reduced model. The total sum of squares is $\mathbf{Y}'\mathbf{Y}$. The reduction in the sum of squares due to including $\mathbf{b}_m$ in a model that already includes $\mathbf{b}_k$ and $\mathbf{C}$ will be denoted as $R(\mathbf{b}_m | \mathbf{C}, \mathbf{b}_k)$. This can also be expressed as

$$R(\mathbf{b}_m | \mathbf{C}, \mathbf{b}_k) = R(\mathbf{C}, \mathbf{b}_k, \mathbf{b}_m) - R(\mathbf{C}, \mathbf{b}_k)$$

There are several ways to compute $R(\mathbf{b}_m|\mathbf{C},\mathbf{b}_k)$. The sum of squares due to the full model, as well as the sum of squares due to the reduced model, can each be calculated, and the difference obtained (Method 1).

$$R(\mathbf{b}_m|\mathbf{C},\mathbf{b}_k) = \hat{\mathbf{C}}'\mathbf{Z}'\mathbf{Y} + \hat{\mathbf{b}}'_k\mathbf{X}'_k\mathbf{Y} + \hat{\mathbf{b}}'_m\mathbf{X}'_m\mathbf{Y} - \tilde{\mathbf{C}}'\mathbf{Z}'\mathbf{Y} - \tilde{\mathbf{b}}'_k\mathbf{X}'_k\mathbf{Y}$$

A sometimes computationally more efficient procedure is to calculate

$$R(\mathbf{b}_m|\mathbf{C},\mathbf{b}_k) = \hat{\mathbf{b}}'_m\mathbf{T}_m^{-1}\hat{\mathbf{b}}_m$$

where $\hat{\mathbf{b}}_m$ are the estimates obtained from fitting the full model and $\mathbf{T}_m$ is the partition of the inverse matrix corresponding to $\mathbf{b}_m$ (Method 2).

$$\begin{bmatrix} \mathbf{Z}'\mathbf{Z} & \mathbf{Z}'\mathbf{X}_k & \mathbf{Z}'\mathbf{X}_m \\ \mathbf{X}'_k\mathbf{Z} & \mathbf{X}'_k\mathbf{X}_k & \mathbf{X}'_k\mathbf{X}_m \\ \mathbf{X}'_m\mathbf{Z} & \mathbf{X}'_m\mathbf{X}_k & \mathbf{X}'_m\mathbf{X}_m \end{bmatrix}^{-1} = \begin{bmatrix} \mathbf{T}_c & \mathbf{T}_{ck} & \mathbf{T}_{cm} \\ \mathbf{T}_{kc} & \mathbf{T}_k & \mathbf{T}_{km} \\ \mathbf{T}_{mc} & \mathbf{T}_{mk} & \mathbf{T}_m \end{bmatrix}$$

# Model and Options

## Notation

Let $\mathbf{b}$ be partitioned as

$$\mathbf{b} = \begin{bmatrix} \mathbf{M} \\ \mathbf{D} \end{bmatrix} = \begin{bmatrix} \mathbf{m}_1 \\ \vdots \\ \mathbf{m}_F \\ \hline \mathbf{d}_1 \\ \vdots \\ \mathbf{d}_{F-1} \end{bmatrix}$$

where

| | |
|---|---|
| $\mathbf{M}$ | Vector of main effect coefficients |
| $\mathbf{m}_i$ | Vector of coefficients for main effect $i$ |
| $\mathbf{m}^{(i)}$ | $\mathbf{M}$ excluding $\mathbf{m}_i$ |
| $\mathbf{M}^{i*}$ | $\mathbf{M}$ including only $\mathbf{m}_1$ through $\mathbf{m}_{i-1}$ |
| $\mathbf{D}$ | Vector of interaction coefficients |
| $\mathbf{d}_k$ | Vector of $k$th order interaction coefficients |
| $\mathbf{d}_{k_i}$ | Vector of coefficients for the $i$th of the $k$th order interactions |
| $\mathbf{D}^{(k)}$ | $\mathbf{D}$ excluding $\mathbf{d}_k$ |
| $\mathbf{D}^{k*}$ | $\mathbf{D}$ including only $\mathbf{d}_1$ through $\mathbf{d}_{k-1}$ |
| $\mathbf{d}_k^{(i)}$ | $\mathbf{d}_k$ excluding $\mathbf{d}_{k_i}$ |
| $\mathbf{C}$ | Vector of covariate coefficients |
| $c_i$ | Covariate coefficient |
| $\mathbf{C}^{(i)}$ | $\mathbf{C}$ excluding $c_i$ |
| $\mathbf{C}^{i*}$ | $\mathbf{C}$ including only $c_1$ through $c_{i-1}$ |

## Models

Different types of sums of squares can be calculated in ANOVA.

## Sum of Squares for Type of Effects

| | Covariates | Main Effects | Interactions |
|---|---|---|---|
| Experimental and Hierarchical | $R(\mathbf{C})$ | $R(\mathbf{M}\mid\mathbf{C})$ | $R\left(\mathbf{d}_k \mid \mathbf{C},\mathbf{M},\mathbf{D}^{k*}\right)$ |
| Covariates with Main Effects | $R(\mathbf{C},\mathbf{M})$ | $R(\mathbf{C},\mathbf{M})$ | $R\left(\mathbf{d}_k \mid \mathbf{C},\mathbf{M},\mathbf{D}^{k*}\right)$ |
| Covariates after Main Effects | $R(\mathbf{C}\mid\mathbf{M})$ | $R(\mathbf{M})$ | $R\left(\mathbf{d}_k \mid \mathbf{C},\mathbf{M},\mathbf{D}^{k*}\right)$ |
| Regression | $R(\mathbf{C}\mid\mathbf{M},\mathbf{D})$ | $R(\mathbf{M}\mid\mathbf{C},\mathbf{D})$ | $R\left(\mathbf{d}_k \mid \mathbf{C},\mathbf{M},\mathbf{D}^{k*}\right)$ |

All sums of squares are calculated as described in the introduction. Reductions in sums of squares $(R(\mathbf{A}|\mathbf{B}))$ are computed using Method 1. Since all cross-product matrices have been corrected for the mean, all sums of squares are adjusted for the mean.

## Sums of Squares Within Effects

| | Covariates | Main Effects | Interactions |
|---|---|---|---|
| Default Experimental | $R\left(c_i \mid \mathbf{C}^{(i)}\right)$ | $R\left(\mathbf{m}_i \mid \mathbf{C},\mathbf{M}^{(i)}\right)$ | $R\left(\mathbf{d}_{k_i} \mid \mathbf{C},\mathbf{M},\mathbf{D}^{k*},\mathbf{d}_k^{(i)}\right)$ |
| Covariates with Main Effects | $R\left(c_i \mid \mathbf{M},\mathbf{C}^{(i)}\right)$ | $R\left(\mathbf{m}_i \mid \mathbf{C},\mathbf{M}^{(i)}\right)$ | same as default |
| Covariates after Main Effects | $R\left(c_i \mid \mathbf{M},\mathbf{C}^{(i)}\right)$ | $R\left(\mathbf{m}_i \mid \mathbf{M}^{(i)}\right)$ | same as default |
| Regression | $R\left(c_i \mid \mathbf{M},\mathbf{C}^{(i)},\mathbf{D}\right)$ | $R\left(m_i \mid \mathbf{M}^{(i)},\mathbf{C},\mathbf{D}\right)$ | $R\left(\mathbf{d}_{k_i} \mid \mathbf{C},\mathbf{M},\mathbf{D}^{(k_i)}\right)$ |
| Hierarchical | $R\left(c_i \mid \mathbf{C}^{i*}\right)$ | $R\left(\mathbf{m}_i \mid \mathbf{C},\mathbf{M}^{i*}\right)$ | same as default |
| Hierarchical and Covariates with Main Effects or Hierarchical and Covariates after Main Effects | $R\left(c_i \mid \mathbf{C}^{i*},\mathbf{M}\right)$ | $R\left(\mathbf{m}_i \mid \mathbf{M}^{i*}\right)$ | same as default |

Reductions in sums of squares are calculated using Method 2, except for specifications involving the Hierarchical approach. For these, Method 1 is used. All sums of squares are adjusted for the mean.

## Degrees of Freedom

### Main Effects

$$df_M = \sum_{i=1}^{F} (k_i - 1)$$

### Main Effect *i*

$$\left(k_i - 1\right)$$

### Covariates

$$df_c = CN$$

### Covariate *i*

$$\mathbf{1}$$

### Interactions $\mathbf{d}_r$

$df_r =$ number of linearly independent columns corresponding to interaction $\mathbf{d}_r$ in $\mathbf{X}'\mathbf{X}$

### Interactions $\mathbf{d}_{r_i}$

$\boldsymbol{df}$ = number of independent columns corresponding to interaction $\mathbf{d}_{r_i}$ in $\mathbf{X}'\mathbf{X}$

### Model

$$df_{Model} = df_M + df_c + \sum_{r=1}^{F-1} df_r$$

#### Residual

$$W-1-df_{Model}$$

#### Total

$$\boldsymbol{W-1}$$

# Multiple Classification Analysis

## Notation

| | |
|---|---|
| $Y_{ijk}$ | Value of the dependent variable for the *k*th case in level *j* of main effect *i* |
| $n_{ij}$ | Sum of weights of observations in level *j* of main effect *i* |
| $k_i$ | Number of nonempty levels in the *i*th main effect |
| $W$ | Sum of weights of all observations |

## Basic Computations

#### Mean of Dependent Variable in Level *j* of Main Effect *i*

$$\overline{Y}_{ij}=\sum_{k=1}^{n_{ij}}Y_{ijk}\Big/n_{ij}$$

#### Grand Mean

$$\overline{Y}=\sum_{i}\sum_{j}\sum_{k}Y_{ijk}\Big/W$$

### Coefficient Estimates

The computation of the coefficient for the main effects only model $\left(b_{ij}\right)$ and coefficients for the main effects and covariates only model $\left(\tilde{b}_{ij}\right)$ are obtained as previously described.

## Calculation of the MCA Statistics (Andrews, et al., 1973)

### Deviations

For each level of each main effect, the following are computed:

### Unadjusted Deviations

The unadjusted deviation from the grand mean for the *j*th level of the *i*th factor:

$$m_{ij} = \overline{Y}_{ij} - \overline{Y}$$

### Deviations Adjusted for the Main Effects

$$m_{ij}^1 = b_{ij} - \sum_{j=2}^{k_i} b_{ij} n_{ij} \Big/ W \text{, where } b_{i1} = 0 \text{.}$$

### Deviations Adjusted for Main Effects and Covariates (Only for Models with Covariates)

$$m_{ij}^2 = \tilde{b}_{ij} - \sum_{j=2}^{k_i} \tilde{b}_{ij} n_{ij} \Big/ W \text{, where } \tilde{b}_{i1} = 0 \text{.}$$

### ETA and Beta Coefficients

For each main effect *i*, the following are computed:

$$\text{ETA}_i = \sqrt{\sum_{j=2}^{k_i} n_{ij} \left( \bar{Y}_{ij} - \bar{Y} \right)^2 \Big/ \mathbf{Y}'\mathbf{Y}}$$

### Beta Adjusted for Main Effects

$$\text{Beta}_i = \sqrt{\sum_{j=2}^{k_i} n_{ij} \left( m_{ij}^1 \right)^2 \Big/ \mathbf{Y}'\mathbf{Y}}$$

### Beta Adjusted for Main Effects and Covariates

$$\text{Beta}_i = \sqrt{\sum_{j=2}^{k_i} n_{ij} \left( m_{ij}^2 \right)^2 \Big/ \mathbf{Y}'\mathbf{Y}}$$

### Squared Multiple Correlation Coefficients

Main effects model

$$R_m^2 = \frac{R(\mathbf{M})}{\mathbf{Y}'\mathbf{Y}} .$$

Main effects and covariates model

$$R_{mc}^2 = \frac{R(\mathbf{M}, \mathbf{C})}{\mathbf{Y}'\mathbf{Y}} .$$

The computations of $R(\mathbf{M})$, $R(\mathbf{M},\mathbf{C})$, and $\mathbf{Y}'\mathbf{Y}$ are outlined previously.

# Unstandardized Regression Coefficients for Covariates

Estimates for the **C** vector, which are obtained the first time covariates are entered into the model, are printed.

# Cell Means and Sample Sizes

Cell means and sample sizes for each combination of factor levels are obtained from the $\mathbf{X}'\mathbf{Y}$ and $\mathbf{X}'\mathbf{X}$ matrices prior to correction for the mean.

$$\overline{Y_i} = \frac{(\mathbf{X}'\mathbf{Y})_i}{(\mathbf{X}'\mathbf{X})_{ii}} \qquad i = 1,\ldots,CN$$

Means for combinations involving the first level of a factor are obtained by subtraction from marginal totals.

### Matrix Inversion

The Cholesky decomposition (Stewart, 1973) is used to triangularize the matrix. If the tolerance is less than $10^{-5}$, the matrix is considered singular.

# References

Andrews, F., Morgan, J., Sonquist, J., and Klem, L. 1973. *Multiple classification analysis*. Ann Arbor, Mich.: University of Michigan at Ann Arbor.

Searle, S. R. 1966. *Matrix algebra for the biological sciences.* New York: John Wiley & Sons, Inc.

Searle, S. R. 1971. *Linear models*. New York: John Wiley & Sons, Inc.

Stewart, G. W. 1973. *Introduction to matrix computations*. New York: Academic Press.