ผนวก ค

Algorithms ของ โปรแกรม SPSS

Descriptives Algorithms

Descriptives computes univariate statistics—including the mean,standard deviation, minimum, and maximum—for numeric variables.

Notation

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

X_i	Value of the variable for case i
w_i	Weight for case i
N	Number of cases
W_i	Sum of the weights for the first i cases
$\overline{X}_i$	Mean for the first i cases

Moments

Moments about the mean are calculated recursively using a provisional Means algorithm (Spicer,1972):

$$W_j = \sum_{i=1}^{j} w_i$$

$$v_j = \frac{w_j}{W_j}\left(X_j - \overline{X}_{j-1}\right)$$

$$M_j^4 = M_{j-1}^4 - 4v_j M_{j-1}^3 + 6v_j^2 M_{j-1}^2 + \left(\frac{W_j^2 - 3w_j W_{j-1}}{w_j^3}\right) v_j^4 W_{j-1} W_j$$

$$M_j^3 = M_{j-1}^3 - 3v_j M_{j-1}^2 + \frac{W_j W_{j-1}}{w_j^2}(W_j - 2w_j) v_j^3$$

$$M_j^2 = M_{j-1}^2 + \frac{W_j W_{j-1}}{w_j} v_j^2$$

$$\overline{X}_j = \overline{X}_{j-1} + v_j$$

$$W_0 = \overline{X}_0 = M_0^2 = M_0^3 = M_0^4 = 0$$

After the last observation has been processed,

W_N = sum of weights for all cases

$\overline{X}_N$ = mean

$$M_N^r = \sum_{i=1}^{N} w_i \left(X_i - \overline{X}\right)^r$$

Basic Statistics

Mean

$$\overline{X}_N$$

Variance

$$S^2 = M_N^2/(W_N - 1)$$

Standard Deviation

$$S = \sqrt{S^2}$$

Standard Error

$$S_{\overline{X}} = \frac{S}{\sqrt{W_N}}$$

Minimum

$$\min_j X_j$$

Maximum

$$\max_j X_j$$

Sum

$$\overline{X}_N W_N$$

Skewness and Standard Error of Skewness

$$g_1 = \frac{W_N M_N^3}{(W_N-1)(W_N-2)S^3} \quad se(g_1) = \sqrt{\frac{6W_N(W_N-1)}{(W_N-2)(W_N+1)(W_N+3)}}$$

If $W_N \leq 2$ or $S^2 < 10^{-20}$, g_1 and its standard error are not calculated.

Kurtosis (Bliss, 1967, p. 144) and Standard Error of Kurtosis

$$g_2 = \frac{W_N(W_N+1)M_N^4 - 3M_N^2 M_N^2 (W_N-1)}{(W_N-1)(W_N-2)(W_N-3)S^4} \quad se(g_2) = \sqrt{\frac{4(W_N^2-1)(SE(g_1))^2}{(W_N-3)(W_N+5)}}$$

If $W_N \leq 3$ or $S^2 < 10^{-20}$, g_2 and its standard error are not calculated.

Z-Scores

$$Z_i = \frac{X_i - \overline{X}_N}{S}$$

If X_i is missing or $S \le 0$, Z_i is set to the system missing value.

References

Bliss, C. I. 1967. Statistics in biology, Volume 1. New York: McGraw-Hill.

Spicer, C. C. 1972. Algorithm AS 52: Calculation of power sums of deviations about the mean. Applied Statistics, 21, 226–227.

ANOVA Algorithms

This chapter describes the algorithms used by the ANOVA procedure.

Model and Matrix Computations

Notation

The following notation is used throughout this section 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 jth 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

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{ZC} + \mathbf{e} \quad (1)$$

where

Y	$N \times 1$ vector of values of the dependent variable
X	Design matrix $(N \times p)$ of rank $q < p$
β	Vector of parameters $p \times 1$
Z	Matrix of covariates $(N \times CN)$
C	Vector of covariate coefficients $(CN \times 1)$
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 α_i and β_j, and interaction parameter γ_{ij}, the constraints can be expressed as

$$\alpha_1 = \beta_1 = \gamma_{1j} = \gamma_{i1} = 0 \quad \text{non} - \text{regression}$$
$$\alpha_\bullet = \beta_\bullet = \gamma_{i\bullet} = \gamma_{\bullet j} = 0 \quad \text{regression}$$

where • indicates summation.

Computation of Matrices

$X'X$

Non-regression Model

The $X'X$ 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:

	α_2	β_2	β_3	γ_{22}	γ_{23}
α_2	$N_{2\bullet}$	N_{22}	N_{23}	N_{22}	N_{23}
β_2		$N_{\bullet 2}$	0	N_{22}	0
β_3			$N_{\bullet 3}$	0	N_{23}
γ_{22}				N_{22}	0
γ_{23}					N_{23}

The elements $N_{i\bullet}$ or $N_{\bullet j}$ on the diagonal are the sums of weights of cases that have level i of a or level j of β. Off-diagonal elements are sums of weights of cases crossclassified by parameter combinations. Thus, $N_{\bullet 3}$ is the sum of weights of cases in level 3 of main effect β_3, while is the sum of weights of cases with α_2 and β_2.

Regression Model

A row of the design matrix **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 **d** is generated for case *k*, the *ij*th cell of $X'X$ 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 $X'X$ 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 $X'X$ are subsequently adjusted for means.

The highest order of interactions in the model can be selected. This will affect the generation of $X'X$ 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 Z'Z

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

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

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

The Vector Z'Y

The covariance between the mth covariate and the dependent variable after case k has been processed is

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

The Scalar Y'Y

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

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

The Vector X'Y

X'Y is a vector with NC rows. The ith element is

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

where, for non-regression model, $\delta_k = 1$ if case k has the factor combination in column i of $X'X$; $\delta_k = 0$ 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 X'Z

The (i, m)th entry is

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

where δ_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

$$\mathbf{X} = [\mathbf{X}_k|\mathbf{X}_m] \text{ and } \beta = \begin{bmatrix} \mathbf{b}_k \\ \mathbf{b}_m \end{bmatrix}.$$

The normal equations are then

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

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 b_m ,

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

the solution to the normal equation is:

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

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

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

The residual (unexplained) sum of squares for the complete model is $RSS = Y'Y - R(C, b_k, b_m)$ and similarly for the reduced model. The total sum of squares is $Y'Y$. The reduction in the sum of squares due to including in a model that already includes b_k and C will be denoted as $R(b_m | C, b_k)$. 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 $R(b_m | C, b_k)$. 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).

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

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 $\hat{b}_m$ are the estimates obtained from fitting the full model and T_m is the partition of the inverse matrix corresponding to 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 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

M	Vector of main effect coefficients
m_i	Vector of coefficients for main effect *i*
$m^{(i)}$	M excluding m_i
m^{i*}	M including only m_1 through m_i -1
D	Vector of interaction coefficients
d_k	Vector of *k*th order interaction coefficients
d_{ki}	Vector of coefficients for the *i*th of the *k*th order interactions
$D^{(k)}$	excluding d_k
D^{k*}	including only d_1 through d_k -1
$d_k^{(i)}$	d_k excluding d_{ki}

C	Vector of covariate coefficients
c_i	Covariate coefficient
$C^{(i)}$	C excluding c_i
C^{i*}	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	$R(\mathbf{C})$	$R(\mathbf{M}\|\mathbf{C})$	$R(\mathbf{d}_k\|\mathbf{C}, \mathbf{M}, \mathbf{D}^{k*})$
Covariates with Main Effects	$R(\mathbf{C}, \mathbf{M})$	$R(\mathbf{C}, \mathbf{M})$	$R(\mathbf{d}_k\|\mathbf{C}, \mathbf{M}, \mathbf{D}^{k*})$
Covariates after Main Effects	$R(\mathbf{C}\|\mathbf{M})$	$R(\mathbf{M})$	$R(\mathbf{d}_k\|\mathbf{C}, \mathbf{M}, \mathbf{D}^{k*})$
Regression	$R(\mathbf{C}\|\mathbf{M}, \mathbf{D})$	$R(\mathbf{M}\|\mathbf{C}, \mathbf{D})$	$R(\mathbf{d}_k\|\mathbf{C}, \mathbf{M}, \mathbf{D}^{k*})$

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.

Sum of Squares Within Effects

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

$$(k_i - 1)$$

Covariates

$$df_c = CN$$

Covariate i

1

Interactions

Interactions d_r:

df_r = number of linearly independent columns corresponding to interaction d_r in $X'X$

Interactions d_{ri} :

df = number of independent columns corresponding to interaction d_{ri} in $X'X$

Model

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

Residual

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

Total

$$\mathbf{W - 1}$$

Multiple Classification Analysis

Notation

Y_{ijk} Value of the dependent variable for the kth 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 ith main effect

W Sum of weights of all observations

Basic Computations

Mean of Dependent Variable in Level j of Main Effect i

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

Grand Mean

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

Coefficient Estimates

The computation of the coefficient for the main effects only model (b_{ij}) and coefficients for the main effects and covariates only model $(\tilde{b}_{ij})$ 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 jth level of the ith factor:

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

Deviations Adjusted for the Main Effects

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

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

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

ETA and Beta Coefficients

For each main effect i, the following are computed:

$$ETA_i = \sqrt{\sum_{j=2}^{k_i} n_{ij} \left(\overline{Y}_{ij} - \overline{Y}\right)^2 / \mathbf{Y}'\mathbf{Y}}$$

Beta Adjusted for Main Effects

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

Beta Adjusted for Main Effects and Covariates

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

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 $X'Y$ and $X'X$ matrices prior to correction for the mean.

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

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., J. Morgan, J. Sonquist, and L. Klein. 1973. Multiple classification analysis, 2nd ed.

Ann Arbor: University of Michigan.

Searle, S. R. 1966. Matrix algebra for the biological sciences. NewYork: JohnWiley & 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.

REGRESSION Algorithms

This procedure performs multiple linear regression with five methods for entry and removal of variables. It also provides extensive analysis of residual and influential cases. Caseweight (CASEWEIGHT) and regression weight (REGWGT) can be specified in the model fitting.

Notation

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

y_i	Dependent variable for case with variance
c_i	Caseweight for case i; if CASEWEIGHT is not specified
g_i	Regression weight for case i; if REGWGT is not specified
l	Number of distinct cases
w_i	$c_i g_i$
W	$\sum_{i=1}^{l} w_i$
P	Number of independent variables
C	Sum of caseweights: $\sum_{i=1}^{l} c_i$
x_{ki}	The kth independent variable for case i
$\bar{X}_k$	Sample mean for the *k*th independent variable: $\bar{X}_k = \left(\sum_{i=1}^{l} w_i x_{ki} \right) \Big/ W$
$\bar{Y}$	Sample mean for the dependent variable: $\bar{Y} = \left(\sum_{i=1}^{l} w_i y_i \right) \Big/ W$
h_i	Leverage for case *i*
$\tilde{h}_i$	$\frac{g_i}{w} + h_i$
S_{kj}	Sample covariance for X_k and X_j
S_{yy}	Sample variance for *Y*
S_{ky}	Sample covariance for X_k and *Y*
P^*	Number of coefficients in the model. $p^* = p$ if the intercept is not included; otherwise $p^* = p + 1$

R The sample correlation matrix for $X_1,\ldots,X_p$ and Y

Descriptive Statistics

$$\mathbf{R} = \begin{bmatrix} r_{11} & \cdots & r_{1p} r_{1y} \\ r_{21} & \cdots & r_{2p} r_{2y} \\ \cdot & \cdots & \cdot \; \cdot \\ r_{y1} & \cdots & r_{yp} r_{yy} \end{bmatrix}$$

where

$$r_{kj} = \frac{S_{kj}}{\sqrt{S_{kk} S_{jj}}}$$

and

$$r_{yk} = r_{ky} = \frac{S_{ky}}{\sqrt{S_{kk} S_{yy}}}$$

The sample mean $\overline{X}_i$ and covariance S_{ij} are computed by a provisional means algorithm. Define $W_k = \sum_{i=1}^{k} w_i$ cumulative weight up to case *k*.

then

$$\overline{X}_{i(k)} = \overline{X}_{i(k-1)} + \left(x_{ik} - \overline{X}_{i(k-1)}\right) \frac{w_k}{W_k}$$

where

$$\overline{X}_{i(1)} = x_{i1}$$

If the intercept is included,

$$C_{ij(k)} = C_{ij(k-1)} + \left(x_{ik} - \overline{X}_{i(k-1)}\right)\left(x_{jk} - \overline{X}_{j(k-1)}\right)\left(w_k - \frac{w_k^2}{W_k}\right)$$

where

$$C_{ij(1)} = 0$$

Otherwise,

$$C_{ij(k)} = C_{ij(k-1)} + w_k x_{ik} x_{jk}$$

where

$$C_{ij(1)} = w_1 x_{i1} x_{j1}$$

The sample covariance S_{ij} is computed as the final C_{ij} divided by C-1.

Sweep Operations (Dempster, 1969)

For a regression model of the form

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_p X_{pi} + e_i$$

sweep operations are used to compute the least squares estimates b of β and the associated regression statistics. The sweeping starts with the correlation matrix R. Let $\tilde{R}$ be the new matrix produced by sweeping on the kth row and column of R. The elements of $\tilde{R}$ are

$$\tilde{r}_{kk} = \frac{1}{r_{kk}}$$
$$\tilde{r}_{ik} = \frac{r_{ik}}{r_{kk}}, \quad i \neq k$$
$$\tilde{r}_{kj} = -\frac{r_{kj}}{r_{kk}}, \quad j \neq k$$

and

$$\tilde{r}_{ij} = \frac{r_{ij}r_{kk} - r_{ik}r_{kj}}{r_{kk}}, \quad i \neq k, j \neq k$$

If the above sweep operations are repeatedly applied to each row of R_{11} in

$$\mathbf{R} = \begin{pmatrix} \mathbf{R}_{11} & \mathbf{R}_{12} \\ \mathbf{R}_{21} & \mathbf{R}_{22} \end{pmatrix}$$

where R_{11} contains independent variables in the equation at the current step, the result is

$$\tilde{\mathbf{R}} = \begin{pmatrix} \mathbf{R}_{11}^{-1} & -\mathbf{R}_{11}^{-1}\mathbf{R}_{12} \\ \mathbf{R}_{21}\mathbf{R}_{11}^{-1} & \mathbf{R}_{22} - \mathbf{R}_{21}\mathbf{R}_{11}^{-1}\mathbf{R}_{12} \end{pmatrix}$$

The last row of

$$\mathbf{R}_{21}\mathbf{R}_{11}^{-1}$$

contains the standardized coefficients (also called BETA), and

$$\mathbf{R}_{22} - \mathbf{R}_{21}\mathbf{R}_{11}^{-1}\mathbf{R}_{12}$$

can be used to obtain the partial correlations for the variables not in the equation, controlling for the variables already in the equation. Note that this routine is its own inverse; that is, exactly the same operations are performed to remove a variable as to enter a variable.

Variable Selection Criteria

Let r_{ij} be the element in the current swept matrix associated with X_i and X_j. Variables are entered or removed one at a time. X_k is eligible for entry if it is an independent variable not currently in the model with

$r_{kk} \geq t$ (tolerance with a default of 0.0001)

and also, for each variable that is currently in the model,

$$\left(r_{jj} - \frac{r_{jk}r_{kj}}{r_{kk}}\right)t \leq 1$$

The above condition is imposed so that entry of the variable does not reduce the tolerance of variables already in the model to unacceptable levels.

The F-to-enter value for X_k is computed as

$$F - to - enter_k = \frac{(C - p^* - 1)V_k}{r_{yy} - V_k}$$

with 1 and C-p^*-1 degrees of freedom, where is the number of coefficients currently in the model and

$$V_k = \frac{r_{yk}r_{ky}}{r_{kk}}$$

The *F*-to-remove value for X_k is computed as

$$F - to - remove_k = \frac{(C - p^*)|V_k|}{r_{yy}}$$

with 1 and C-p^* degrees of freedom.

Methods for Variable Entry and Removal

Five methods for entry and removal of variables are available. The selection process is repeated until the maximum number of steps (MAXSTEP) is reached or no more independent variables qualify for entry or removal. The algorithms for these five methods are described in the following sections.

Stepwise

If there are independent variables currently entered in the model, choose X_k such that F – to – remove$_k$ is minimum. X_k is removed if F – to – remove$_k$ < F_{out} (default = 2.71) or, if probability criteria are used, P(F – to – remove$_k$) > F_{out} (default = 0.1). If the inequality does not hold, no variable is removed from the model.

If there are no independent variables currently entered in the model or if no entered variable is to be removed, choose X_k such that F – to – enter$_k$ is maximum. X_k is entered if F – to – enter$_k$ > F_{in} (default = 3.84) or, if P(F – to – enter$_k$) < F_{in} (default = 0.05). If the inequality does not hold, no variable is entered.

At each step, all eligible variables are considered for removal and entry.

Forward

This procedure is the entry phase of the stepwise procedure.

Backward

This procedure is the removal phase of the stepwise procedure and can be used only after at least one independent variable has been entered in the model.

Enter (Forced Entry)

Choose X_k such r_{kk} that is maximum and enter X_k .Repeat for all variables to be entered.

Remove (Forced Removal)

Choose X_k such r_{kk} that is minimum and remove X_k .Repeat for all variables to be removed.

Statistics

The following statistics are available.

Summary

For the summary statistics, assume p independent variables are currently entered in the equation, of which a block of q variables have been entered or removed in the current step.

Multiple R

$$R = \sqrt{1 - r_{yy}}$$

R Square

$$R^2 = 1 - r_{yy}$$

Adjusted R Square

$$R_{adj}^2 = R^2 - \frac{(1 - R^2)p}{C - p^*}$$

R Square Change (when a block of q independent variables was added or removed)

$$\Delta R^2 = R_{current}^2 - R_{previous}^2$$

F Change and Significance of F Change

$$\Delta F = \begin{cases} \frac{\Delta R^2 (C - p^*)}{q(1 - R_{current}^2)} & \text{for the addition of independent variables} \\ \frac{\Delta R^2 (C - p^* - q)}{q(R_{previous}^2 - 1)} & \text{for the removal of independent variables} \end{cases}$$

the degrees of freedom for the addition are q and $C - p^*$, while the degrees of freedom for the removal are q and $C - p^* - q$.

Residual Sum of Squares

$$SS_e = r_{yy}(C - 1)S_{yy}$$

with degrees of freedom $C - p^*$.

Sum of Squares Due to Regression

$$SS_R = R^2(C-1)S_{yy}$$

with degrees of freedom p.

ANOVA Table

Analysis of Variance	*df*	*Sum of Squares*	*Mean Square*
Regression	p	SS_R	$(SS_R)/p$
w	$C-p^*$	SS_e	$(SS_e)/(C-p^*)$

Standard Error of Estimate

Also known as the standard error of regression, this is simply the square root of the mean square residual from the ANOVA table, or $\sqrt{(SS_e)/(C-)p^*}$.

Variance-Covariance Matrix for Unstandardized Regression Coefficient Estimates

A square matrix of size p with diagonal elements equal to the variance, the below diagonal elements equal to the covariance, and the above diagonal elements equal to the correlations:

$$var(b_k) = \frac{r_{kk} r_{yy} S_{yy}}{S_{kk}(C-p^*)}$$

$$cov(b_k, b_j) = \frac{r_{kj} r_{yy} S_{yy}}{\sqrt{S_{kk} S_{jj}}(C-p^*)}$$

$$cor(b_k, b_j) = \frac{r_{kj}}{\sqrt{r_{kk} r_{jj}}}$$

Selection Criteria

The following selection criteria are available.

Akaike Information Criterion (AIC)

$$AIC = C \ln\left(\frac{SS_e}{C}\right) + 2p^*$$

Amemiya's Prediction Criterion (PC)

$$PC = \frac{\left(1 - R^2\right)(C + p^*)}{C - p^*}$$

Mallow's CP

$$CP = \frac{SS_e}{\hat{\sigma}^2} + 2p^* - C$$

where $\hat{\sigma}^2$ is the mean square error from fitting the model that includes all the variables in the variable list.

Schwarz Bayesian Criterion (SBC)

$$SBC = C \ln\left(\frac{SS_e}{C}\right) + p^* \ln(C)$$

Collinearity

The following measures of collinearity are available.

Variance Inflation Factors

$$VIF_i = \frac{1}{r_{ii}}$$

Tolerance

$$\text{Tolerance}_i = r_{ii}$$

Eigenvalues

The eigenvalues of scaled and uncentered cross-product matrix for the independent variables in the equation are computed by the QL method (Wilkinson and Reinsch, 1971).

Condition Indices

$$\eta_k = \frac{\max \lambda_j}{\lambda_k}$$

Variance-Decomposition Proportions

Let

$$\mathbf{v}_i = (v_{i1}, \ldots, v_{ip})$$

be the eigenvector associated with eigenvalue λ_i . Also, let

$$\Phi_{ij} = v_{ij}^2/\lambda_i \text{ and } \Phi_j = \sum_{i=1}^{p} \Phi_{ij}$$

The variance-decomposition proportion for the jth regression coefficient associated with the ith component is defined as

$$\pi_{ij} = \Phi_{ij}/\Phi_j$$

Statistics for Variables in the Equation

The following statistics are computed for each variable in the equation.

Regression Coefficient

$$b_k = \frac{r_{yk}\sqrt{S_{yy}}}{\sqrt{S_{kk}}} \text{ for } k = 1, \ldots, p$$

The standard error of b_k is computed as

$$\hat{\sigma}_{b_k} = \sqrt{\frac{r_{kk} r_{yy} S_{yy}}{S_{kk}(C - p^*)}}$$

95% confidence interval for coefficient

$$b_k \pm \hat{\sigma}_{b_k} t_{0.975, C-p^*}$$

If the model includes the intercept, the intercept is estimated as

$$b_0 = \overline{y} - \sum_{k=1}^{p} b_k \overline{X}_k$$

The variance of b_0 is estimated by

$$\hat{\sigma}_{b_0}^2 = \frac{(C-1) r_{yy} S_{yy}}{C(C-p^*)} + \sum_{k=1}^{p} \overline{X}_k^2 \hat{\sigma}_{b_k}^2 + 2 \sum_{k=j+1}^{p} \sum_{j=1}^{p-1} \overline{X}_k \overline{X}_j est.\text{cov}(b_k, b_j)$$

Beta Coefficients

$$Beta_k = r_{yk}$$

The standard error of $Beta_k$ is estimated by

$$\hat{\sigma}_{Beta_k} = \sqrt{\frac{r_{yy} r_{kk}}{C - p^*}}$$

F-test for $Beta_k$

$$F = \left(\frac{Beta_k}{\hat{\sigma}_{Beta_k}}\right)^2$$

with 1 and $C - p^*$ degrees of freedom.

Part Correlation

$$Part - Corr(X_k) = \frac{r_{yk}}{\sqrt{r_{kk}}}$$

Partial Correlation

$$Partial - Corr(X_k) = \frac{r_{yk}}{\sqrt{r_{kk} r_{yy} - r_{yk} r_{ky}}}$$

Statistics for Variables Not in the Equation

The following statistics are computed for each variable not in the equation.

Standardized regression coefficient Beta if predictor enters the equation at the next step

$$Beta_k^* = \frac{r_{yk}}{r_{kk}}$$

The F-test for $Beta_k$

$$F = \frac{(C - p^* - 1) r_{yk}^2}{r_{kk} r_{yy} - r_{yk}^2}$$

with 1 and $C - p^*$ degrees of freedom

Partial Correlation

$$Partial(X_k) = \frac{r_{yk}}{\sqrt{r_{yy} r_{kk}}}$$

Tolerance

$$\text{Tolerance}_k = r_{kk}$$

Minimum tolerance among variables already in the equation if predictor enters at the next step is

$$\min_{1 \le j \le p} \left(\frac{1}{r_{jj} - \left(r_{kj} r_{jk}\right)/r_{kk}}, r_{kk} \right)$$

Residuals and Associated Statistics

There are 19 temporary variables that can be added to the active system file. These variables can be requested with the RESIDUAL subcommand.

Centered Leverage Values

For all cases, compute

$$h_i = \begin{cases} \frac{g_i}{(C-1)} \sum_{j=1}^{p} \sum_{k=1}^{p} \frac{\left(X_{ji} - \overline{X}_j\right)\left(X_{ki} - \overline{X}_k\right) r_{jk}}{\sqrt{S_{jj} S_{kk}}} & \text{if intercept is included} \\ \frac{g_i}{(C-1)} \sum_{j=1}^{p} \sum_{k=1}^{p} \frac{X_{ji} X_{ki} r_{jk}}{\sqrt{S_{jj} S_{kk}}} & \text{otherwise} \end{cases}$$

For selected cases, leverage is ; for unselected case i with positive caseweight, leverage is

$$h'_i = \begin{cases} g_i \left[\left(\frac{1}{W} + h_i\right) / \left(1 + \frac{1}{W} + h_i\right) - \frac{1}{W+1} \right] & \text{if intercept is included} \\ h_i / (1 + h_i / g_i) & \text{otherwise} \end{cases}$$

Unstandardized Predicted Values

$$\hat{Y}_i = \begin{cases} \sum_{k=1}^{p} b_k X_{ki} & \text{if no intercept} \\ b_0 + \sum_{k=1}^{p} b_k X_{ki} & \text{otherwise} \end{cases}$$

Unstandardized Residuals

$$e_i = Y_i - \hat{Y}_i$$

Standardized Residuals

$$ZRESID_i = \begin{cases} \frac{e_i}{s} & \text{if no regression weight is specified} \\ \text{SYSMIS} & \text{otherwise} \end{cases}$$

where s is the square root of the residual mean square.

Standardized Predicted Values

$$ZPRED_i = \begin{cases} \frac{\hat{Y}_i - \overline{Y}}{sd} & \text{if no regression weight is specified} \\ \text{SYSMIS} & \text{otherwise} \end{cases}$$

where *sd* is computed as

$$sd = \sqrt{\sum_{i=1}^{l} \frac{c_i \left(\hat{Y}_i - \overline{Y}\right)^2}{C - 1}}$$

Studentized Residuals

$$SRES_i = \begin{cases} \frac{e_i/s}{\sqrt{(1-\tilde{h}_i)/g_i}} & \text{for selected cases with } c_i > 0 \\ \frac{e_i/s}{\sqrt{(1+\tilde{h}_i)/g_i}} & \text{otherwise} \end{cases}$$

Deleted Residuals

$$DRESID_i = \begin{cases} e_i/\left(1 - \tilde{h}_i\right) & \text{for selected cases with } c_i > 0 \\ e_i & \text{otherwise} \end{cases}$$

Studentized Deleted Residuals

$$SDRESID_i = \begin{cases} \frac{DRESID_i}{s^*_{(i)}} & \text{for selected cases with } c_i > 0 \\ \frac{e_i}{s\sqrt{(1+\bar{h}_i)/g_i}} & \text{otherwise} \end{cases}$$

where

$$s^*_{(i)} = \frac{1}{\sqrt{C - p^* - 1}} \sqrt{\frac{(C - p^*)s^2}{1 - \bar{h}_i} - DRESID_i^2}$$

Adjusted Predicted Values

$$ADJPRED_i = Y_i - DRESID_i$$

DfBeta

$$DFBETA_i = b - b(i) = \frac{g_i e_i \left(\mathbf{X}'\mathbf{W}\mathbf{X}\right)^{-1} \mathbf{X}_i^t}{1 - \bar{h}_i}$$

where

$$\mathbf{X}_i^t = \begin{cases} (1, X_{1i}, \ldots, X_{pi}) & \text{if intercept is included} \\ (X_{1i}, \ldots, X_{pi}) & \text{ottherwise} \end{cases}$$

and $\mathbf{W} = diag(w_1, \ldots, w_l)$.

This is only computed for selected cases with case weight greater than or equal to 1.

Standardized DfBeta

$$SDBETA_{ij} = \frac{b_j - b_j(i)}{s_{(i)}\sqrt{(\mathbf{X}^t\mathbf{W}\mathbf{X})^{-1}_{jj}}}$$

where $b_j \quad b_j(i)$ **is the** j**th component of b** $\quad$ **b**(i), **and**

$$s_{(i)} = s^*_{(i)}\sqrt{1 - \bar{h}_i}$$

This is only computed for selected cases with case weight greater than or equal to 1.

DfFit

$$DFFIT_i = \mathbf{X}_i[\mathbf{b} - \mathbf{b}(i)] = \frac{\bar{h}_i e_i}{1 - \bar{h}_i}$$

This is only computed for selected cases with case weight greater than or equal to 1.

Standardized DfFit

$$SDFIT_i = \frac{DFFIT_i}{s_{(i)}\sqrt{\bar{h}_i}}$$

This is only computed for selected cases with case weight greater than or equal to 1.

Covratio

$$COVRATIO_i = \left(\frac{s_{(i)}}{s}\right)^{2p^*} \times \frac{1}{1 - \bar{h}_i}$$

This is only computed for selected cases with case weight greater than or equal to 1.

Mahalanobis Distance

For selected cases with $c_i > 0$,

$$MAHAL_i = \begin{cases} (C-1)h_i & \text{if intercept is included} \\ Ch_i & \text{otherwise} \end{cases}$$

For unselected cases with $c_i > 0$

$$MAHAL_i = \begin{cases} Ch'_i & \text{if intercept is included} \\ (C+1)h'_i & \text{otherwise} \end{cases}$$

Cook's Distance (Cook, 1977)

For selected cases with $c_i > 0$

$$COOK_i = \begin{cases} \left(DRESID_i^2 \bar{h}_i g_i\right) / \left[s^2(p+1)\right] & \text{if intercept is included} \\ \left(DRESID_i^2 h_i g_i\right) / \left(s^2 p\right) & \text{otherwise} \end{cases}$$

For unselected cases with $c_i > 0$

$$COOK_i = \begin{cases} \left(DRESID_i^2\left(h'_i + \frac{1}{W}\right)\right)/\left[\bar{s}^2(p+1)\right] & \text{if intercept is included} \\ \left(DRESID_i^2 h'_i\right)/\left(\bar{s}^2 p\right) & \text{otherwise} \end{cases}$$

where h'_i is the leverage for unselected case i, and $\bar{s}^2$ is computed as

$$\bar{s}^2 = \begin{cases} \frac{1}{C-p}\left[SS_e + e_i^2\left(1 - h'_i - \frac{1}{1+W}\right)\right] & \text{if intercept is included} \\ \frac{1}{C-p+1}\left[SS_e + e_i^2\left(1 - h'_i\right)\right] & \text{otherwise} \end{cases}$$

Standard Errors of the Mean Predicted Values

For all the cases with positive caseweight,

$$SEPRED_i = \begin{cases} s\sqrt{\tilde{h}_i/g_i} & \text{if intercept is included} \\ s\sqrt{h_i/g_i} & \text{otherwise} \end{cases}$$

95% Confidence Interval for Mean Predicted Response

$$LMCIN_i = \hat{Y}_i - t_{0.975,C-p} \cdot SEPRED_i$$
$$UMCIN_i = \hat{Y}_i + t_{0.975,C-p} \cdot SEPRED_i$$

95% Confidence Interval for a Single Observation

$$LICIN_i = \begin{cases} \hat{Y}_i - t_{0.975,C-p} \cdot s\sqrt{\left(\tilde{h}_i + 1\right)/g_i} & \text{if intercept is included} \\ \hat{Y}_i - t_{0.975,C-p} s\sqrt{(h_i + 1)/g_i} & \text{otherwise} \end{cases}$$

$$UICIN_i = \begin{cases} \hat{Y}_i + t_{0.975,C-p} \cdot s\sqrt{\left(\tilde{h}_i + 1\right)/g_i} & \text{if intercept is included} \\ \hat{Y}_i + t_{0.975,C-p} s\sqrt{(h_i + 1)/g_i} & \text{otherwise} \end{cases}$$

Durbin-Watson Statistic

$$DW = \frac{\sum_{i=2}^{l} (\ddot{e}_i - \ddot{e}_{i-1})^2}{\sum_{i=1}^{l} c_i \ddot{e}_i^2}$$

where $\ddot{e}_i = e_i\sqrt{g_i}$.

Note: the Durbin-Watson statistic cannot be computed if there are fractional case weights. Even with integer case weights, the formula is only valid if the case weights represent contiguous case replications in the original sample.

Partial Residual Plots

The scatterplots of the residuals of the dependent variable and an independent variable when both of these variables are regressed on the rest of the independent variables can be requested in the RESIDUAL branch. The algorithm for these residuals is described in (Velleman and Welsch, 1981).

Missing Values

By default, a case that has a missing value for any variable is deleted from the computation of the correlation matrix on which all consequent computations are based. Users are allowed to change the treatment of cases with missing values.

References

Cook, R. D. 1977. Detection of influential observations in linear regression. Technometrics, 19, 15–18.

Dempster, A. P. 1969. Elements of Continuous Multivariate Analysis. Reading,MA: Addison-Wesley.

Velleman, P. F., and R. E. Welsch. 1981. Efficient computing of regression diagnostics. American Statistician, 35, 234–242.

Wilkinson, J. H., and C. Reinsch. 1971. Linear Algebra. In: Handbook for Automatic Computation, Volume II, J. H. Wilkinson, and C. Reinsch, eds. NewYork: Springer-Verlag.