REGRESSION

```
REGRESSION VARIABLES={varlist   }
                     {ALL       }
                     {(COLLECT)}

 [/STATISTICS=[DEFAULTS**] [R] [COEFF] [ANOVA] [OUTS]
              [ZPP] [CHA] [CI] [F] [BCOV] [SES] [TOL]
              [COND] [XTX] [HISTORY] [END] [LINE] [ALL]]

 [/CRITERIA=[DEFAULTS**] [TOLERANCE({0.01  })] [MAXSTEPS({2v})]
                                    {value}             {n }
            [{PIN({0.05  })}] [{POUT({0.1   })}]
             {    {value} }    {     {value}
             {FIN({3.84  })}   {FOUT({2.71  })
             {    {value} }    {     {value}

 [/{NOORIGIN**}]
   {ORIGIN    }

 /DEPENDENT=varlist

 /[METHOD]={STEPWISE [= varlist]     } [/METHOD=...]
           {FORWARD [=varlist]       }
           {BACKWARD [=varlist]      }
           {ENTER [=varlist]         }
           {REMOVE=varlist           }
           {TEST=(varlist) (varlist)}

 [DESCRIPTIVES=[DEFAULTS] [MEAN] [STDDEV] [CORR]
               [VARIANCE] [XPROD] [SIG] [N] [BADCORR]
               [COV]      [ALL]   [NONE**]]

 [/SELECT={ALL**                    }]
          {varname relation value}

 [/MISSING={LISTWISE**       } [INCLUDE]]
           {PAIRWISE         }
           {MEANSUBSTITUTION}

 [/WIDTH={value or SET**}]
         {n             }
```

**Default if subcommand is omitted.

Example:

```
REGRESSION VARIABLES=XVAR1 TO XVAR7,YVAR/ DEPENDENT=YVAR
/METHOD=ENTER XVAR1 XVAR2 XVAR3/METHOD=ENTER XVAR4 TO XVAR7
/METHOD=STEPWISE.
```

Overview

Procedure REGRESSION calculates multiple regression equations and associated statistics and plots. Several methods for selecting variables into the equation are available. Statistics for analyzing residuals and influential observations are also available. Several types of plots, including partial residual plots, can be displayed.

Defaults

The minimum specification for REGRESSION is a list of variables from which a correlation matrix is computed (VARIABLES), a dependent variable which begins building of an equation (DEPENDENT), and a method for selecting blocks of independent variables for the equation (METHOD). For each block of variables selected, the default display includes statistics on the equation (including R^2 and analysis of variance), on the variables in the equation (including regression coefficients), and on variables being considered that are not in the equation. The default display uses the width specified on SET. By default, all cases in the active file with valid values for all the variables named on

the VARIABLES subcommand are used to compute the correlation matrix on which the regression equations are based. The default equations include a constant (intercept).

Tailoring

The options for procedure REGRESSION can be grouped into logically related operations or specifications.

Global-Control Subcommands. These optional subcommands can be named only once and determine which variables and cases are included in the analysis. DESCRIPTIVES requests descriptive statistics on the variables in the analysis. SELECT estimates the model based on a subset of cases. MISSING specifies the treatment of cases with missing values.

Equation-Control Subcommands. These optional subcommands control the calculation and display of each equation. STATISTICS controls the statistics displayed as the equations are built. CRITERIA specifies the criteria used by the variable selection method, and ORIGIN specifies whether regression is through the origin.

Display Format. The WIDTH subcommand controls the width of the display for REGRESSION only.

Analysis of Residuals. The optional subcommands that analyze and plot residuals are described in REGRESSION: Residuals.

Writing and Reading Matrices. The optional subcommands that read and write matrix materials are described in REGRESSION: Matrix Materials.

Syntax

- The minimum specification is three subcommands and their specifications: VARIABLES, which lists the variables that will be used to compute the correlation matrix; DEPENDENT, which initiates equation(s) and specifies at least one dependent variable; and METHOD, which specifies the method to be used in selecting independent variables. If no variable list is specified on METHOD, all variables named on VARIABLES but not on DEPENDENT will be considered for selection.
- The VARIABLES subcommand can be used only once and must be specified before the DEPENDENT and METHOD subcommands.
- More than one DEPENDENT subcommand can be used.
- A DEPENDENT subcommand must be followed immediately by a METHOD subcommand. More than one METHOD subcommand can follow a DEPENDENT subcommand.
- The MISSING, DESCRIPTIVES, and SELECT subcommands may precede or follow the VARIABLES subcommand they modify. If any of these subcommands appears more than once on the REGRESSION command, the last one of each type is in effect for the entire REGRESSION command. These subcommands must not be placed between the DEPENDENT and METHOD subcommands.
- The last CRITERIA, STATISTICS, and ORIGIN subcommands encountered before a DEPENDENT subcommand are in effect for all equations built from that DEPENDENT subcommand. These subcommands must not be placed between the DEPENDENT and METHOD subcommands.
- The WIDTH subcommand can appear anywhere. The last WIDTH subcommand encountered is in effect for all REGRESSION displays.
- All subcommands must be separated by slashes.

Operations

- REGRESSION is a procedure and causes the data to be read.
- REGRESSION calculates a correlation matrix that includes all variables named on the VARIABLES subcommand. All equations requested on the REGRESSION command are calculated from the same correlation matrix.
- The MISSING, DESCRIPTIVES, and SELECT subcommands control the calculation of the correlation matrix and associated displays.
- The DEPENDENT subcommand and the METHOD subcommands that follow it control the building of equations.

- An equation is built for each variable named on a DEPENDENT subcommand using the same independent variables and methods.
- Multiple METHOD subcommands operate on the equations defined by the preceding DEPENDENT subcommand.
- All independent variables that pass the tolerance criterion are candidates for entry (see CRITERIA Subcommand).
- If the width set is less than 132, all statistics requested may not be displayed. The WIDTH subcommand within REGRESSION allows you to increase the display width and obtain all the statistics available.

Limitations

- The number of variables that can be named on the VARIABLES subcommand depends on the memory available.

Example

```
REGRESSION VARIABLES=XVAR1 TO XVAR7,YVAR DEPENDENT=YVAR
 METHOD=ENTER XVAR1 XVAR2 XVAR3/METHOD=ENTER XVAR4 TO XVAR7
 METHOD=STEPWISE.
```

- VARIABLES requests that variables XVAR1 to XVAR7 and YVAR be made available for analysis by REGRESSION.
- DEPENDENT defines a single equation, with YVAR as the dependent variable.
- The first METHOD subcommand requests that XVAR1 to XVAR3 be entered into the equation.
- The second METHOD subcommand requests that XVAR4 to XVAR7 be added to the equation containing XVAR1 to XVAR3.
- The last METHOD subcommand requests that the entire equation be evaluated using the stepwise method.

VARIABLES Subcommand

The required VARIABLES subcommand names all the variables in the analysis with either a variable list or a keyword.

- The minimum specification is a list of two variables or the keyword ALL or (COLLECT). There is no default variable list.
- There must be one and only one VARIABLES subcommand, and it must precede any DEPENDENT or METHOD subcommands.
- All variables named on the DEPENDENT and METHOD subcommands must be named or implied on the VARIABLES subcommand.
- You can name any user-defined variable in the active file.
- You can use the TO keyword in the variable list to refer to consecutive variables on the active file.
- The order of variables in the correlation matrix constructed by REGRESSION is the same as their order on the VARIABLES subcommand.
- If you use the keyword (COLLECT), the order of variables in the correlation matrix is the order in which they are first referred to on the DEPENDENT and METHOD subcommands.

You can specify either of the following keywords instead of a variable list:

ALL *Include all user-defined variables in the active file.*

(COLLECT) *Include all variables named on the DEPENDENT and METHOD subcommands.* If (COLLECT) is used, the METHOD subcommands must have variable lists.

Example

```
REGRESSION VARIABLES=(COLLECT)/DEPENDENT=YVAR
/METHOD=STEP BVAR CVAR DVAR EVAR FVAR GVAR
/METHOD=ENTER HVAR
/DEPENDENT=HVAR/METHOD=ENTER GVAR.
```

- (COLLECT) requests that the correlation matrix include YVAR, BVAR, CVAR, DVAR, EVAR, FVAR, GVAR, and HVAR.

- The first DEPENDENT subcommand defines a single equation in which YVAR is the dependent variable.
- The first METHOD subcommand requests that the block of variables BVAR to GVAR be considered for inclusion using a stepwise procedure.
- The second METHOD subcommand defines a second variable block consisting of HVAR. HVAR is entered if it satisfies the default tolerance criterion.
- A second DEPENDENT subcommand requests an equation in which HVAR is the dependent variable.
- GVAR will be entered into this equation if it satisfies the default tolerance criterion.

DEPENDENT Subcommand

The required DEPENDENT subcommand specifies a list of variables and requests that an equation be built for each.

- The minimum specification is a single variable. There is no default variable list.
- You can specify more than one DEPENDENT subcommand. Each must be followed by at least one METHOD subcommand.
- All variables named on a DEPENDENT subcommand must have been previously named or implied on the VARIABLES subcommand.
- The keyword TO on a DEPENDENT subcommand refers to the order variables are named on the VARIABLES subcommand. If VARIABLES=(COLLECT) was specified, TO refers to the order of variables on the active file.
- If the DEPENDENT subcommand names more than one variable, an equation is built for each using the same independent variable(s) and methods. Thus, no variable named on the DEPENDENT subcommand can be specified as an independent variable on associated METHOD subcommands.

METHOD Subcommand

The required METHOD subcommand specifies a variable selection method and names a block of variables to be evaluated using that method.

- The minimum specification is a method keyword and, for some methods, a list of variables. The actual keyword METHOD may be omitted.
- The default variable list for methods FORWARD, BACKWARD, STEPWISE, and ENTER consists of all variables named on the VARIABLES subcommand that are not named on the preceding DEPENDENT subcommand.
- There is no default variable list for the REMOVE and TEST methods.
- If VARIABLES=(COLLECT) is specified, you cannot use the default variable list and must name the variables.
- The keyword TO in a variable block named on METHOD refers to the order variables are named on the VARIABLES subcommand. If VARIABLES=(COLLECT) was specified, TO refers to the order of variables on the active file.
- At least one METHOD subcommand must follow each DEPENDENT subcommand.
- When you specify more than one method for a single DEPENDENT subcommand, the methods are cumulative.

The available stepwise methods are

BACKWARD *Backward elimination.* Variables in the block are considered for removal. At each step, the variable with the largest probability-of-*F* value is removed, provided that the value is larger than POUT (see CRITERIA Subcommand). If no variables are in the equation when BACKWARD is specified, all independent variables are first entered.

FORWARD *Forward entry.* Variables in the block are added to the equation one at a time. At each step, the variable not in the equation with the smallest probability of *F* is entered if the value is smaller than PIN (see CRITERIA Subcommand).

STEPWISE *Stepwise selection.* If there are independent variables already in the equation, the variable with the largest probability of F is removed if this value is larger than POUT. The equation is recomputed omitting the removed variable and the evaluation process is repeated until no more independent variables can be removed. Then, the independent variable not in the equation with the smallest probability of F is entered if this value is smaller than PIN. Then all variables in the equation are again examined for removal. This process continues until no variables in the equation need to be removed and no variables not in the equation are eligible for entry, or until the maximum number of steps has been reached (see CRITERIA Subcommand).

The methods that enter or remove the entire variable block in a single step are

ENTER *Forced entry.* All variables in the block are entered in a single step in order of decreasing tolerance. If the order in which variables are entered is important, use multiple METHOD=ENTER subcommands.

REMOVE *Forced removal.* All variables are removed in a single step. REMOVE requires a variable list.

TEST *Test indicated subsets of independent variables using* R^2 *change and its test of significance.* This method first builds a model that includes all variables named on that METHOD=TEST subcommand, adding any variables that are not already in the equation. Variables already in the equation are not removed. Test statistics are printed for the subsets, usually in addition to any other statistics that may have been requested. Test subsets are specified in parentheses. A variable can be used in more than one subset, and each subset can include any number of variables. Variables named on TEST remain in the equation when the METHOD is completed.

Example

```
REGRESSION VARIABLES=XVAR1 TO XVAR7,YVAR DEPENDENT YVAR
/METHOD=STEPWISE/METHOD=ENTER.
```

- STEPWISE applies the stepwise procedure to variables XVAR1 to XVAR7.
- Any variables not in the equation when the STEPWISE method is complete will be forced into the equation with ENTER.

Example

```
REGRESSION VARIABLES=(COLLECT)/DEPENDENT=ZVAR
/METHOD=TEST(QVAR15 TO QVAR20)(QVAR15,PVAR)
/METHOD=ENTER XVAR.
```

- The VARIABLES=(COLLECT) specification assembles a correlation matrix that includes all variables named on the DEPENDENT and METHOD subcommands. QVAR15 TO QVAR20 refers to all variables between and including QVAR15 and QVAR20 on the active file.
- REGRESSION first builds the full equation of all the variables named on the first METHOD subcommand: ZVAR regressed on QVAR15 to QVAR20 and PVAR. For each set of test variables, the R^2 change, F, probability, sums of squares, and degress of freedom are displayed.
- XVAR is added to the equation by the second METHOD subcommand. Variables QVAR15 to QVAR20 and PVAR are in the equation when this subcommand is executed.

STATISTICS Subcommand

The optional STATISTICS subcommand controls the display of statistics for the equation and for the independent variables.

- The minimum specification is simply the subcommand keyword.
- If the STATISTICS subcommand is omitted or if it is included with no keywords, R, ANOVA, COEFF, and OUTS are displayed (see below).
- If any statistics are specified on STATISTICS, only statistics specifically requested are displayed.

- The statistical displays for stepwise methods (BACKWARD, FORWARD, and STEPWISE) are sometimes different from those for methods that enter and remove blocks of variables. In particular, some statistical displays are not produced for method TEST if the equation has not changed since the last variable block.
- Method TEST always produces its own display in addition to other statistics.
- A STATISTICS subcommand affects any equations that are subsequently defined and remains in effect until overridden by another STATISTICS subcommand.
- A STATISTICS subcommand may not be placed between the DEPENDENT and METHOD subcommands.
- If the width is set to less than 132, some requested statistics may not be displayed.

Global Statistics

DEFAULTS *R, ANOVA, COEFF, and OUTS.* These are displayed if the STATISTICS subcommand is omitted or if it is specified without keywords.

ALL *Display all statistics except F, LINE, and END.*

Equation Statistics

R *Multiple* R. Includes R^2, adjusted R^2, and standard error of the estimate.

ANOVA *Analysis of variance table.* Includes regression and residual sum of squares, mean square, F, and probability of F.

CHA *Change in* R^2. Includes the change in R^2 between steps, F at the end of each step and its probability, and F for the equation and its probability. For stepwise methods (BACKWARD, FORWARD, and STEPWISE), these statistics are displayed at the end of each step. For other methods, the statistics are displayed for the variable block.

BCOV *Variance-covariance matrix for unstandardized regression coefficients.* Matrix has covariances below the diagonal, correlations above the diagonal, and variances on the diagonal.

XTX *Sweep matrix.*

COND *Condition number bounds.* Displays the upper and lower bounds for the condition number of the submatrix of the sweep matrix, which contains independent variables already entered. (See Berk, 1977.)

Statistics for the Independent Variables

COEFF *Regression coefficients.* Includes regression coefficients (B), standard errors of the coefficients, standardized regression coefficients (beta), t, and two-tailed probability of t.

OUTS *Statistics for variables not yet in the equation* that have been named on METHOD subcommands for the equation. Displays beta, t, two-tailed probability of t, and minimum tolerance of the variable if it were the only variable entered next.

ZPP *Zero-order, part, and partial correlation.*

CI *95% confidence interval for the unstandardized regression coefficient.*

SES *Approximate standard error of the standardized regression coefficients.* (See Meyer & Younger, 1976.)

TOL *Tolerance.* Displays tolerance for variables in the equation and, for variables not in the equation, the tolerance each variable would have if it were the only variable entered next.

F *F value for* B *and its probability.* Displayed instead of the t value.

Step Summary Statistics

The full summary line displayed by keywords LINE, HISTORY, and END includes R, R^2, adjusted R^2, F, probability of F, R^2 change, F of the change, probability of R^2 change, and statistics on variables added or removed. For stepwise methods (BACKWARD, FORWARD, and STEPWISE), the statistics refer to each step. For other methods (ENTER, REMOVE, and TEST), the statistics refer to the entire variable block. If other statistics are requested, the summary line may not be produced for a block that does not entail steps.

LINE *Display a single summary line for each step for step methods.* The default or requested statistics are displayed at the end of each method block for all methods.

HISTORY *Display a final summary report.* For stepwise methods, the report includes a summary line for each step. For ENTER and REMOVE, the report includes a summary line for each method. A summary line is displayed for TEST only if the equation changes. If HISTORY is the only statistic requested, COEFF is displayed for the final equation.

END *Display one summary line per step for step methods and one summary line per variable block for other methods.* The summary line is displayed for TEST only if the equation changes. Other default or requested statistics are displayed at the completion of the last METHOD subcommand for the equation.

CRITERIA Subcommand

The optional CRITERIA subcommand controls the statistical criteria used in building the regression equations. The way in which these criteria are used depends on the method specified on the METHOD subcommand. The default criteria are noted in the description of each CRITERIA keyword below.

- The minimum specification is a criterion keyword and its arguments, if any.
- If the CRITERIA subcommand is omitted or included with no specifications, the default criteria are in effect.
- Only default criteria that are changed explicitly on the CRITERIA subcommand are altered.
- A CRITERIA subcommand affects any subsequent DEPENDENT and METHOD subcommands and remains in effect until overridden by another CRITERIA subcommand.
- The CRITERIA subcommand may not be placed between the DEPENDENT and METHOD subcommands.

Tolerance and Minimum Tolerance Criteria

Variables must pass both tolerance and minimum tolerance tests in order to enter and remain in a regression equation. Tolerance is the proportion of the variance of a variable in the equation that is not accounted for by other independent variables in the equation. The minimum tolerance of a variable not in the equation is the smallest tolerance any variable already in the equation would have if the variable being considered were included in the analysis.

If a variable passes the tolerance criteria, it is further tested according to the method in effect. These criteria are controlled by the CRITERIA subcommand or defaults in effect for the equations that the DEPENDENT subcommand builds.

Testing Independent Variables

The ENTER, REMOVE, and TEST methods use only the TOLERANCE criterion. The stepwise methods (BACKWARD, FORWARD, and STEPWISE) differ in the way they use criteria.

- BACKWARD selects variables according to the probability of *F*-to-remove (keyword POUT). Specify FOUT to use *F*-to-remove.
- FORWARD selects variables according to the probability of *F*-to-enter (keyword PIN). Specify FIN to use *F*-to-enter.
- STEPWISE uses both PIN and POUT (or FIN and FOUT) as criteria. If the criterion for entry (PIN or FIN) is less stringent than the criterion for removal (POUT or FOUT), the same variable may cycle in and out until the maximum number of steps is reached. If PIN is larger than POUT or FIN is smaller than FOUT, REGRESSION adjusts POUT or FOUT and issues a warning.

DEFAULTS *PIN(0.05), POUT(0.10), and TOLERANCE(0.01).* These are the defaults if the CRITERIA subcommand has not been specified. If criteria have been changed, keyword DEFAULTS restores these defaults.

PIN(value) *Probability of* F-*to-enter*. The default value is 0.05. Keywords PIN and FIN are mutually exclusive. If more than one is used, the last mentioned is in effect.

FIN(value) F-*to-enter*. If no value is specified, the value defaults to 3.84. Keywords PIN and FIN are mutually exclusive. If more than one is used, the last mentioned is in effect.

POUT(value) *Probability of* F-*to-remove*. The default value is 0.10. Keywords POUT and FOUT are mutually exclusive. If more than one is used, the last mentioned is in effect.

FOUT(value) F-*to-remove*. If no value is specified, the value defaults to 2.71. Keywords POUT and FOUT are mutually exclusive. If more than one is used, the last mentioned is in effect.

TOLERANCE(value) *Tolerance*. The default value is 0.01. If the tolerance chosen is very low, REGRESSION issues a warning.

MAXSTEPS(n) *Maximum number of steps*. The value of MAXSTEPS is the sum of the maximum number of steps over each method for the equation. The default values are

BACKWARD or FORWARD methods: the number of variables meeting PIN/POUT or FIN FOUT criteria.

STEPWISE method: twice the number of independent variables.

Example

```
REGRESSION VARIABLES=XVAR1 TO XVAR7,YVAR
/CRITERIA=PIN(.1) POUT(.15) TOL(.001)
/DEPENDENT=YVAR/METHOD=FORWARD
/CRITERIA=DEFAULTS
/DEPENDENT=YVAR/METHOD=STEPWISE.
```

- The first CRITERIA subcommand relaxes the default criteria for entry and removal while the FORWARD method is used.
- The second CRITERIA subcommand reestablishes the defaults for the second equation.

ORIGIN and NOORIGIN Subcommands

The optional ORIGIN and NOORIGIN subcommands control whether or not the constant is suppressed.

- The minimum specification is simply the ORIGIN or NOORIGIN subcommand. There are no additional specifications.
- ORIGIN and NOORIGIN must be specified before the DEPENDENT and METHOD subcommands they modify.
- ORIGIN and NOORIGIN are mutually exclusive.
- If neither ORIGIN nor NOORIGIN is specified, NOORIGIN is the default and all equations include a constant term (intercept).
- ORIGIN requests regression through the origin. The constant term is suppressed. Once specified, ORIGIN remains in effect until NOORIGIN is requested.

Example

```
REGRESSION VAR=(COL)
/DEP=VARA/METHOD=ENTER VARB
/ORIGIN/DEP=VARA/METHOD=ENTER VARB
/NOORIGIN/DEP=VARB/METHOD=ENTER VARC.
```

- The keyword (COLLECT) builds a correlation matrix that includes VARA, VARB, and VARC.
- The REGRESSION command requests three equations. The first regresses VARA on VARB and includes a constant term because the default (NOORIGIN) is in effect. The second regresses VARA on VARB and suppresses the constant (ORIGIN). The third regresses VARB on VARC and includes a constant term because NOORIGIN has been specified.
- This example takes advantage of spelling permitted by three-character truncation of keywords.

DESCRIPTIVES Subcommand

By default, descriptive statistics are not displayed. Use the optional DESCRIPTIVES subcommand to request the display of correlation and descriptive statistics for variables in the correlation matrix.

- The minimum specification is simply the command keyword.
- If DESCRIPTIVES is specified without keywords, MEAN, STDDEV, and CORR are displayed.
- If DESCRIPTIVES is included and any keywords are specified, only those statistics specifically requested are displayed.
- Descriptive statistics are displayed only once for all variables named or implied on VARIABLES.
- Descriptive statistics are based on all valid cases for each variable if PAIRWISE or MEANSUBSTITUTION has been specified on MISSING. Otherwise, only cases included in the computation of the correlation matrix are included in the calculation of the descriptive statistics.

NONE *Turn off all descriptive statistics.* This is the default if the subcommand is omitted.

DEFAULTS *MEAN, STDDEV, and CORR.* Same as specifying DESCRIPTIVES without specifications.

MEAN *Variable means.*

STDDEV *Variable standard deviations.*

VARIANCE *Variable variances.*

CORR *Correlation matrix.*

SIG *One-tailed probabilities of the correlation coefficients.*

BADCORR *Display the correlation matrix only if some coefficients cannot be computed.*

COV *Covariance matrix.*

XPROD *Cross-product deviations from the mean.*

N *Numbers of cases used to compute correlation coefficients.*

ALL *Display all descriptive statistics.*

Example

```
REGRESSION DESCRIPTIVES=DEFAULTS SIG COV
/VARIABLES=XVAR1 TO XVAR7,YVAR
/DEPENDENT=YVAR/METHOD=ENTER XVAR1/METHOD=ENTER XVAR2.
```

- The variable means, variable standard deviations, correlation matrix, one-tailed probabilities of the correlation coefficients, and covariance matrix are displayed.
- Statistics are displayed for all variables named on VARIABLES, even though only variables YVAR, XVAR1, and XVAR2 are used to build the equations.
- XVAR1 is entered into the equation by the first METHOD subcommand; XVAR2 is entered by the second METHOD subcommand.

SELECT Subcommand

By default, all cases on the active file are considered for inclusion in REGRESSION. Use the optional SELECT subcommand to include a subset of cases in the correlation matrix and resulting regression statistics.

- The minimum specification is a logical expression or the keyword ALL (the default if the subcommand is omitted).
- Do not include the variable named on SELECT on the VARIABLES subcommand.
- The logical expression on SELECT is of the form

  ```
  /SELECT=varname relation value
  ```

  where the relation can be EQ, NE, LT, LE, GT, or GE.
- Only cases for which the logical expression on SELECT is true are included in the calculation of the correlation matrix and regression statistics. All other cases, including those with missing values for the variable named on SELECT, are unselected.

- SELECT displays statistics describing the fit of the model estimated among the unselected cases.
- By default, residuals and predicted values are calculated and reported separately for both selected and unselected cases (see REGRESSION: Residuals).
- Cases deleted from the active file with the SELECT IF, PROCESS IF, or SAMPLE commands are not passed to REGRESSION. Such cases are not reviewed by the SELECT subcommand and are not included among either the selected or unselected cases.
- The display of the values of the variable named on SELECT is controlled by that variable's format (see DATA LIST).

Example

```
REGRESSION SELECT SEX EQ 'BOYS'
/VARIABLES=XVAR1 TO XVAR7,YVAR/DEPENDENT=YVAR
/METHOD=STEP/RESIDUALS=NORMPROB.
```

- Only cases with the value BOYS for SEX are included in the correlation matrix calculated by REGRESSION.
- Separate normal probability plots are displayed for boys and girls (see REGRESSION: Residuals).

MISSING Subcommand

By default, a case that has a user- or system-missing value for any variable named or implied on the VARIABLES subcommand is omitted from the computation of the correlation matrix on which all analyses are based. Use the optional MISSING subcommand to change the treatment of cases with missing values.

- The minimum specification is a keyword specifying a missing-value treatment.

The available keywords are

LISTWISE *Delete cases with missing values listwise.* Only cases with valid values on all variables named on the current VARIABLES subcommand are used. If INCLUDE is also specified, only cases with system-missing values are deleted listwise. LISTWISE is the default if the MISSING subcommand is omitted.

PAIRWISE *Delete cases with missing values pairwise.* Each correlation coefficient is computed using cases with complete data for the pair of variables correlated. If INCLUDE is also specified, only cases with system-missing values are deleted pairwise.

MEANSUBSTITUTION *Replace missing values with the variable mean.* All cases are included and the substitutions are treated as valid observations. If INCLUDE is also specified, user-missing values are included in the computation of the means. Mean substitution affects the computation of the correlation matrix and the calculation of predicted values and residuals.

INCLUDE *Include cases with user-missing values.* All user-missing values are treated as valid values by the methods LISTWISE, PAIRWISE, and MEANSUBSTITUTION. Including user-missing values affects the computation of the correlation matrix and the calculation of predicted values and residuals.

Example

```
REGRESSION  VARIABLES=XVAR1 TO XVAR7,YVAR/DEPENDENT=YVAR
/METHOD=STEP
/MISSING=MEANS.
```

- Missing values are replaced with the means of the variables when the correlation matrix is calculated.
- The MEANS keyword takes advantage of spelling permitted by three-character truncation.

WIDTH Subcommand

The optional WIDTH subcommand controls the width of the display within the REGRESSION procedure.

- The minimum specification is an integer between 72 and 132.
- The default display uses the width specified on SET. The width specified on the WIDTH subcommand within REGRESSION overrides the width on SET for the REGRESSION display only.
- The WIDTH subcommand can appear anywhere.
- If more than one WIDTH subcommand is included, the last WIDTH specified will be in effect for the display.
- If the width is less than 132, some statistics may not be displayed.

References

Berk, K. N. 1977. Tolerance and condition in regression computation. *Journal of the American Statistical Association* 72:863-66.

Meyer, L. S., and M. S. Younger. 1976. Estimation of standardized coefficients. *Journal of the American Statistical Association* 71:154-57.

REGRESSION: Matrix Materials

```
REGRESSION [READ=[DEFAULTS] [MEAN] [STDDEV]
           [VARIANCE] {CORR} [N]]
                      {COV }

 [/WRITE=[DEFAULTS] [MEAN] [STDDEV]
         [VARIANCE] [CORR] [COV]
         [N] [NONE**]]

 /VARIABLES=varlist/DEPENDENT=varlist/METHOD=method
```

**Default if subcommand is omitted.

Example:

```
REGRESSION  VARIABLES=AGE TO SUICIDE DESCRIPTIVES
/WRITE
/DEP=SUICIDE/METHOD=ENTER
/RESIDUALS.
```

Overview

Procedure REGRESSION can both read and write matrix materials. The matrix materials REGRESSION reads can be written by REGRESSION or other SPSS/PC procedures such as CORRELATION, or they can be entered from other sources if the appropriate keyword specifications are used. Such matrix materials can be processed more quickly than cases. Matrix materials to be read can be in either fixed or freefield format but must conform to certain format and record specifications (see DATA LIST: Matrix Materials).

The subcommands for writing and reading matrix materials can be used in addition to the REGRESSION subcommands described in REGRESSION.

WRITE Subcommand

Use the WRITE subcommand to write the matrix materials used in the REGRESSION computations to an external file.

- The WRITE subcommand can be specified anywhere except between the DEPENDENT subcommand and the METHOD subcommands that define an equation.
- If the WRITE subcommand is included without specifications, REGRESSION writes a vector of means, a vector of standard deviations, a correlation matrix, and the number of cases.
- If any keyword specifications are given, only those materials specified are written.
- Matrix materials are written for all variables named on the VARIABLES subcommand.
- The order of variables in vectors and matrices is the same as the order in which variables are named on the VARIABLES subcommand.
- If (COLLECT) is used on the VARIABLES subcommand, the order of the variables in the matrix materials is the order in which they are first named on REGRESSION. Use keyword CORRELATION on the DESCRIPTIVES subcommand to display a listing of the matrix you write.
- Only one WRITE subcommand should be used. If more than one WRITE appears, the last one encountered will be in effect.
- Each type of matrix material written begins on a new record. Eight 10-column fields are used on each record.
- REGRESSION displays a format table describing the contents and format of the file it has written.
- The VARIABLES, SELECT, ORIGIN or NOORIGIN, and MISSING subcommands in effect for the REGRESSION procedure also affect the materials that are written.

- Matrix materials are written to the results file named on the SET command (by default, SPSS.PRC).
- If the results file named on SET is not empty when the WRITE subcommand is executed, its contents will be overwritten.

The following keywords can be specified on WRITE. These keywords should be specified in the order they are listed below.

DEFAULTS *Includes MEAN, STDDEV, CORR, and N.* This is the default if the WRITE subcommand is used without specifications.

MEAN *Write a vector of means.*

STDDEV *Write a vector of standard deviations.*

VARIANCE *Write a vector of variances.*

CORR *Write a correlation matrix.*

COV *Write a covariance matrix.*

N *Write out the n's of cases used to compute correlation coefficients.* When the MISSING=LISTWISE is in effect, the number of cases is written as one item on the last record. When PAIRWISE or MEANSUBSTITUTION is specified on MISSING, a matrix of *n*'s is written.

NONE *Turn off previous WRITE specifications.*

Example

```
SET RESULTS='GSS82.MAT'.
DATA LIST FILE='GSS82.DAT'/AGE 5-6 INCOME 7-13
ANOMIE1 TO ANOMIE7 14-20 SUICIDE 21.
REGRESSION  VARIABLES=AGE TO SUICIDE/DESCRIPTIVES
/WRITE
/DEP=SUICIDE/METHOD=ENTER
/RESIDUALS.
```

- The SET command sets the results file to the external file GSS82.MAT in the current directory.
- The DATA LIST command specifies variable names and column locations for raw data in file GSS82.DAT.
- The VARIABLES subcommand on REGRESSION requests that all variables in the active file between and including AGE and SUICIDE be included in the computation of the correlation matrix for the regression.
- The DESCRIPTIVES subcommand requests descriptive statistics for all variables in the analysis.
- The WRITE subcommand requests that default matrix materials be written to the external file specified on the SET command. When the writing is complete, GSS82.MAT will contain a vector of means, a vector of standard deviations, a correlation matrix, and the number of cases, in that order. All variables from AGE to SUICIDE will be included in the matrix materials.
- A regression of SUICIDE on all other variables implied by the VARIABLES subcommand will be computed and the default display produced (see REGRESSION).
- The RESIDUALS subcommand displays a histogram of the standardized residuals, a normal probability plot of the standardized residuals, the Durbin-Watson test statistic, and a listing of the 10 worst outliers based on the absolute value of the standardzied residuals (see REGRESSION: Residuals).

READ Subcommand

Use the READ subcommand to read matrix materials.

- There can be only one READ subcommand.
- The READ subcommand cannot be specified between the DEPENDENT and METHOD subcommands.
- If the READ subcommand is used without specifications, REGRESSION assumes that the matrix materials have the default structure: a vector of means, a vector of standard deviations, a correlation matrix, and the number of cases.

- If any keywords are specified on READ, only those matrix materials specified will be expected.
- The matrix materials you read must permit regression statistics to be calculated.
- You must read either a correlation or covariance matrix.
- When you specify READ on REGRESSION. you must first specify a DATA LIST command that points to the file containing the matrix materials and names the variables that will be read (see DATA LIST: Matrix Materials).
- All matrix materials read must be in a single input file.
- The order in which variables are named on the VARIABLES subcommand must be the order of the variables in each vector or matrix that is read.
- If a correlation matrix is the only matrix material to be read. an N command must be included to specify the number of cases. Only standardized coefficients will be available.

*If more than one kind of matrix material is present. the matrix materials must be arranged in the input file in the following order: the vector of means. the vector of standard deviations, the vector of variances. the correlation or covariance matrix, and the n's of cases.

- The specification on the MISSING and ORIGIN or NOORIGIN subcommands should agree with the options in effect when the matrix was written.
- The descriptive statistics available with the DESCRIPTIVES subcommand depend on which matrix materials are read.
- The RESIDUALS, CASEWISE. SCATTERPLOT, and PARTIALPLOT subcommands are not available when matrix materials are read.
- The (COLLECT) keyword in the VARIABLES subcommand is not allowed if the READ subcommand is used.

DEFAULTS ***MEAN. STDDEV, CORR. and*** N. If you specify READ without specifications. the input file must contain these materials in this order. Matrix materials written by REGRESSION are in this default format.

MEAN — *The matrix is preceded by a vector of means.*

STDDEV — *The matrix is preceded by a vector of standard deviations.*

VARIANCE — *The matrix is preceded by a vector of variances.*

CORR — *Correlation matrix*. Alternative to keyword COV.

COV — *Covariance matrix*. Alternative to keyword CORR. A covariance matrix is not allowed if pairwise deletion of missing values is specified.

N — *The number of cases used to compute correlation coefficients follows the matrix.* If the MISSING subcommand specifies MEANSUBSTITUTION or PAIRWISE, a symmetric matrix of n's is expected. If the MISSING subcommand specifies LISTWISE or INCLUDE, all coefficients are based on the same number of cases and a single number is expected. If a single number of cases is expected, it will be read from the first IO columns of the last record of the matrix file if the keyword N is specified on the READ subcommand. If the keyword N is not specified. the *n* specified on the N command is used. If an *n* is read from the matrix file and an *n* is specified on the N command. the *n* read from the matrix tile is used.

Example

```
DATA LIST FIXED MATRIX FILE='GSS82.MAT'
  /ACE INCOME ANOMIE1 TO ANOMIE7 SUICIDE.
REGRESSION READ
/VARIABLES=AGE INCOME ANOMIE1 TO ANOMIE7 SUICIDE
/DEPENDENT=SUICIDE/METHOD=ENTER ANOMIE1 TO ANOMIE7.
```

- The DATA LIST command specifies that the matrix input should be read from the file GSS82.MAT and names the variable in the file. The names of the variables on the DATA LIST command will be entered in the dictionary of the active file.

- The READ subcommand on REGRESSION requests that matrix materials be read and used for the procedure. Because no keyword specifications are given and the default listwise treatment of missing values is in effect, REGRESSION expects a vector of means, a vector of standard deviations, a correlation matrix, and a single *n* of cases.
- The VARIABLES subcommand names the variables in the order in which they appear in the vectors and matrix to be read.
- The DEPENDENT subcommand defines an equation in which SUICIDE is the dependent variable.
- The METHOD subcommand requests that the variables ANOMIE1 to ANOMIE7 be entered into the equation using the ENTER method.
- The variables AGE and INCOME are not used in the equation but must be named on the VARIABLES subcommand so that the locations of all variables in the matrix file are identified accurately.