WIPP PA

User's Manual

for

STEPWISE, Version 2.20

Document Version 1.00

WPO # 27768

NOVEMBER 1, 1995

TABLE OF CONTENTS

1.0 INTRODUCTION

This document is the User's Manual for the regression-analysis code STEPWISE (version number 2.20) in the context of its use in the 1996 Waste Isolation Pilot Project (WIPP) Compliance Certification Application (CCA) Performance Assessment (PA). STEPWISE receives input data files from two WIPP-PA sources, as follows: (1) sampled input parameter values are transferred from LHS by LHS2STEP, and (2) calculated release data that correspond to those input data are transferred from CCDFCALC by CCD2STEP. STEPWISE relates those two data files by performing various regression-analysis computations and reporting the results tabularly.

This user's manual identifies STEPWISE's sponsor and its expert consultant (Section 1). It describes the code's WIPP-PA purposes and functions (Section 2), provides recommended user training (Section 3), outlines the code's mathematical basis and numerical methods (Section 4), its capabilities and limitations (Section 5), describes user interactions (Section 6), input files (Section 7), error messages (Section 8), and output files (Section 9), and provides examples of relevant input and output files in its Appendices.

1.1 Software Identifier:

Code Name:	STEPWISE: A program to perform regression analyses on selected (normally uncertain) WIPP PA independent and dependent variables in connection with parameter sensitivity studies.
WIPP Prefix:	STP
Version Number:	2.20
Date:	05/30/96
Platform:	FORTRAN 77 for OpenVMS AXP, version 6.1, on a DEC Alpha

1.2 Points of Contact:

Sponsor: Amy P. Gilkey
New Mexico Engineering Research Institute
2201 Buena Vista Drive, SE
Albuquerque, NM 87106
Voice: (505) 766-9629
Fax: (505) 766-9125

Consultant: James W. Garner
Piru Associates, Inc.
2201 Buena Vista Drive, SE
Albuquerque, NM 87106
Voice: (505) 766-9629
Fax: (505) 766-9125

2.0 FUNCTIONAL REQUIREMENTS

This section lists STEPWISE's WIPP-relevant Functional Requirements as taken from the code's Requirements Document (known as the RD/VVP). They are as follows:

R.1 - STEPWISE can develop a regression model between calculated response and input variables using a stepwise regression procedure (STEPWISE).

R.2 - STEPWISE can develop a regression model between calculated response and input variables using a backward elimination procedure (BACKWARD).

R.3 - STEPWISE can develop a regression model between calculated response and input variables using a forward procedure, which is a special case of the stepwise regression procedure.

R.4 - STEPWISE can perform regression procedures on the ranks of the data (RANK).

R.5 - STEPWISE allows the user to force variables to appear in the regression model (FORCE).

R.6 - STEPWISE allows the user to drop observations (input vectors) from the analysis (DROP_OBS).

R.7 - STEPWISE allows the user to request the predicted error sum of squares (PRESS) to assist in model selection.

R.8 - STEPWISE allows the user to create scatter plots of the input variables (SCATTER).

R.9 - STEPWISE allows the user to define new variables which are a function of the input variables. These new variables are defined through the use of an equation which may contain other variables, constants, operators (+,-,*,/,^), and functions (ABS, LOG, LOG10, EXP, SQRT, IFLT0, IFEQ0, IFGT0) (TRANSFORM).

3.0 REQUIRED USER TRAINING AND/OR BACKGROUND

To exercise STEPWISE, users should have (1) basic knowledge of OpenVMS and Digital Command Language, and (2) access to the WIPP cluster of Alpha computers with an OpenVMS AXP (version 6.1) operating system, or their functional equivalents.

Because STEPWISE is primarily a regression-analysis code used to assess WIPP-PA parameter sensitivities, there is no scientific or regulatory reason to manipulate its input data. Thus, users need not have any special knowledge of the data files involved. However, to gain an overall understanding of the role and implications of STEPWISE's results within the WIPP PA process, it would be helpful if users had (1) a basic understanding of introductory statistical theory, including regression analysis and, specifically, the (a) stepwise, (b) forward and (c) backward-elimination procedures, (2) a basic overview understanding of the WIPP PA process and especially of the uncertain physical parameters that are used, the uncertainty-sampling methods that are used, and the application of input-data vectors in WIPP PAs (WIPP PA Dept. [Volumes 1 through 5], 1992a, 1992b, 1992c, 1992d, and 1992e), and (3) an operational familiarity with and general understanding of STEPWISE's principal upstream codes (LHS2STEP and CCD2STEP), which are described elsewhere in the WIPP PA software QA documentation (WPO # 27931 and 27937).

4.0 DESCRIPTION OF THE MODEL AND METHODS

STEPWISE neither models physical phenomena nor solves differential equations that model physical phenomena. Rather, it performs various statistical analyses to assess sensitivities to input variables of results calculated by other WIPP codes that do model physical phenomena. The STEPWISE code offers the analyst numerous plotting and regression-model building techniques that measure the strength of the relationship, if one exists, between input variables and calculated results.

STEPWISE originated as a basic regression code developed by K.E. Kemp, at the Statistics Department of Kansas State University. The code was further developed by Iman et al. (1980) and has since served as a regression-analysis code within Sandia National Laboratories.

4.1 Description of the Model

A complex hydrological model like the WIPP code system employs many geophysical and human-behavioral variables whose numerical values are known with varying degrees of certainty. Those whose values have a strong influence on results are of greater concern in WIPP PAs, and means must be devised (a) to identify them and (b) to quantify the strength of their influence. Perhaps the simplest tool for that purpose is the scatter plot. It is a plot of (i) sampled numerical values of an uncertain physical variable used in the computations (the independent variable) versus (ii) the calculated results (the dependent variable). If little or no relationship exists between an independent-dependent variable pair, their scatter plot will resemble a random distribution of points. However, if a notable relationship exists between them, the plotted points will cluster and exhibit a recognizable form.

Scatter plots are invaluable in identifying a relationship, but they do not quantify the intensity of that relationship. To accomplish that, a measuring tool is required. It is provided by linear regression modeling. Through regression models, it is possible to identify the variable or variables that contribute most to the variability observed in the results computed for WIPP PAs. However, due to the large number of independent variables and the large sample sizes encountered in WIPP PAs, it is inefficient to develop a regression model on every possible variable and then search for the ones having the greatest effect on results (the "best" or "most significant" subset of the original variables). Therefore some optimal regression method allowing a more systematic search is called for.

The STEPWISE code offers the user three options to identify the subset of uncertain input variables having the greatest influence on output variability. The three methods are: (1) the backward-elimination method, (2) the forward regression method (a special case of the stepwise regression method), (3) and the stepwise regression method. (1) The backward-elimination procedure begins with all the independent variables being included in the initial model. Backward elimination then drops variables progressively from the low-influence end of the variable list until the only remaining variables are those whose influence is significant at or above some user-specified level. (2) The forward regression procedure, which is a special case of the stepwise regression procedure, begins by regressing on the one independent variable in the model that explains the most variation in the dependent variable, as measured by its having the largest sample correlation coefficient with the dependent variable (see Section 4.2.6). It then progressively adds variables to the regression model in order of their additional contribution to the variability of the results as measured by their partial correlation (see Section 4.2.8) with the dependent variable. (3) The stepwise regression procedure uses this same forward procedure with an additional feature. At each step of the analysis, the variables presently in the regression model are reexamined regarding their current contribution to the regression sum of squares. The variable contributing least to the regression is dropped from the analysis, providing its contribution is below a user-specified level of significance. This procedure continues until only variables that are significant at or above some prespecified level remain in the model.

In addition to the three regression-model-building methods described above, the user may (a) force certain variables to appear in the final model, (b) perform regression analysis on the ranks of the data rather than on their numerical values, and (c) create new variables that are user-specified functions of the original variables. The code also allows the user opportunities to output (d) intermediate steps of the regression analysis, (e) information associated with the input data, and (f) an analysis of the residuals associated with the final model.

Users desiring a more thorough introduction to the underlying statistical concepts associated with the code are referred to the texts by Neter, Waserman, and Kutner (1983), or Draper and Smith (1981). For information concerning the use of the STEPWISE code in WIPP PA sensitivity analyses, the user is referred to SAND reports documenting the sensitivity analysis task in earlier iterations of the WIPP PA (WIPP PA Dept, 1992d, Volume 4; and Helton, et al., 1991).

4.2 Description of the Methods

4.2.1 Scatter Plots

Scatter plots are perhaps the simplest tools used in sensitivity analyses. They are plots of the various (sampled) numerical values of input variables used in a computation versus the results calculated using those numerical values for the input variables. Suppose, for example, a model is characterized by P input variables $X_1,..., X_P$. For example, X_1 might be the solubility in brine of an inventoried radioisotope, X_2 might be the partition coefficient for that isotope in Culebra clays, and so on. Suppose the resultant output of the computational model, Y, is total integrated release at the 5-km boundary 10,000 years after decommissioning.

Because the P input variables, X_j, where $j = 1,..., P$ are all uncertain, the WIPP sampling code LHS[*] (WIPP PA Dept [Volume 1], 1992a) assigns to each X_j a set of N numerical values that span its entire range of plausible variability. In the interest of creating a well-posed, solvable problem, the condition $N \geq P$ represents a minimum sample size. Therefore, each of the P input variables will be assigned N different numerical values, which will be denoted X_{ij}, $i = 1,..., N$,

[*] The Latin hypercube sampling (LHS) technique provides an effective experimental design for random sampling of input parameters. It assures that the entire range of variability of all uncertain parameters is well characterized.

$j = 1,...,P$. For each running of the model with each different input variable vector, a different result Y_i is obtained. A plot of the various calculated values of Y_i against all N sampled values of, say, the j^{th} input variable, X_j, is called a scatter plot of Y versus X_j.

If little or no relationship exists between the sampled values of the selected input variable and the corresponding calculated results, the points of the scatter plot will be distributed more or less randomly over the plotting field. Alternatively, the existence of a well-defined relationship between the sampled values of an input variable and their corresponding calculated results would produce an organized distribution or clustering of the points in some sort of recognizable and/or describable pattern. In cases where model results are wholly dominated by one or two input parameters, scatter plots can reveal the relationship between inputs and outputs. They are especially effective in revealing nonlinear dependencies, threshold behaviors, variable interactions, and other dependencies that are invaluable in understanding model behaviors and sensitivities. However, scatter plots are basically qualitative in nature, and quantitative measures are preferred for performance assessments. Examples of STEPWISE's quantitative approach to sensitivity analysis are given in the subsections that follow.

4.2.2 Linear Regression Analysis

Sensitivity analyses carried out on Latin hypercube sampled data include the construction of linear regression models, which approximate the behavior of results from a computer-based modeling system. Suppose, as above, a computer model is characterized by P uncertain input variables $X_1,...,X_P$ and leads to an output result Y. Because the input variable values are uncertain, the model is run with N different LHS-sampled sets of observations, or input vectors. Each observation vector is numerically different from every other observation vector and produces a family of N generally different output results Y_i. These data can be fit to a regression model of the form

$$\hat{Y} = b_o + \sum_{j=1}^{P} b_j X_j \ ,$$

where the method of least squares is applied to determine the coefficients b_j, which are called "ordinary regression coefficients." This equation represents the line* that best fits the scatter-plot data in the sense that the sum of the squared differences between the actual scatter-plot data and the values given by the regression line are minimized for the case of the chosen coefficients, b_j. Clearly, the ordinary regression coefficients may be thought of as the partial derivatives of the regression model with respect to the input variables. However, these ordinary regression coefficients are dimensional quantities and their numerical values are strongly dependent on the units in which the variables are expressed. For example, if Y is a release measure and X_3 is a permeability value, then the numerical value of b_3 would decrease by roughly seven orders of magnitude if the units of X_3 were changed from m/s to cm/day. Moreover, the various

* If $P = 1$ (i.e., there are two regression coefficents, b_0 and b_1), the equation represents a line. If $P = 2$, the equation represents a plane. For $P \geq 3$, a space having more than three dimensions is required to plot the equation.

different elements of the set b_j, i.e., for different values of j, have different dimensions one from another. Therefore, the elements of b_j cannot be directly compared one to another. Consequently, the b_j do not provide satisfactory direct measures of the relative importance of the P different input variables. To achieve a direct comparison between regression coefficients, the input variables must be transformed linearly to form new variables called "standardized variables", which are discussed below.

4.2.3 Standardized Variables

The problem associated with different input variables being measured in different units can be eliminated by introducing "standardized" (dimensionless) variables, which are centered on the means of their samples and normalized by the standard deviations of their samples. STEPWISE defines standardized variables, denoted by an asterisk, as: $X^* = (X - \bar{x})/s_x$ for each of the P input variables, where $\bar{x}$ and s_x are the means and standard deviations, respectively, for each sample variable, either independent or dependent. In terms of standardized variables, the previous regression model reduces to the standardized form,

$$\hat{Y}^* = \sum_{j=1}^{P} b_j^* X_j^* .$$

The coefficients, b_j^*, for $j = 1,..., P$, are known as standardized regression coefficients (SRCs). They are more useful for sensitivity studies than ordinary regression coefficients in that they are dimensionless, and they vary over roughly the same range of numerical values. Therefore, they can be compared one to another to provide a direct quantitative measure of the relative importance of the various input variables, providing the input variables are statistically independent. Thus, if b_3^* is numerically greater than b_4^*, then Y is more sensitive to variations in X_3 than it is to variations in X_4. As always, the reliability of such results is conditional on the degree to which the actual relationship between Y and the $X_1,...,X_P$ is described by the regression model. That topic is addressed in the next subsection.

4.2.4 Coefficients of Determination

Coefficients of determination are useful quantitative measures of the effectiveness of a regression model to account for the variability in a scatter plot. To see how and why that is true, it is only necessary to realize that the total (squared) variability in a set of calculated outputs Y_i, as measured from their mean value, that is, $(Y_i - \bar{Y})^2$, can be represented as the sum of two specific variabilities; namely, (i) the sum of the (squared) variabilities measured by the regression model compared to the mean and (ii) the sum of the (squared) variabilities due to the difference between the regression model and the data themselves, that is, the lack of fit by the regression model (squared), which is also known as the sum of squares of the "residuals". In equation form, that property may be written as:

$$\sum_{i=1}^{N}\left(Y_i-\overline{Y}\right)^2=\sum_{i=1}^{N}\left(\hat{Y}_i-\overline{Y}\right)^2+\sum_{i=1}^{N}\left(Y_i-\hat{Y}_i\right)^2 .$$

Here, the first term on the right-hand side represents item (i) above and the second term represents item (ii). This identity leads immediately to the coefficient of determination, which is a convenient one-parameter quantitative measure of the adequacy of a regression model to account for the variability on a scatter plot. The coefficient of determination is defined as:

$$R_y^2=\sum_{i=1}^{N}\left(\hat{Y}_i-\overline{Y}\right)^2 \Big/ \sum_{i=1}^{N}\left(Y_i-\overline{Y}\right)^2 .$$

R_y^2 varies between 0 and 1 and measures the fraction of the variation in Y attributable to regression on the X's. If R_y^2 is near one, the regression model is accounting for much of the output variability. If it is near zero, the regression model is failing to represent output variability well.

4.2.5 Stepwise Regression Modeling

In WIPP PAs, it is instructive, efficient, and usually simpler to perform regression analysis in a stepwise fashion, rather than treating all the independent variables in a single model. In the stepwise approach, a sequence of regression models is constructed starting with a single selected input parameter, and including one additional input variable at each successive step until all significant input variables have been included in the model, where "significant" remains to be defined quantitatively. One would like to retain all regression coefficients that are significantly different from zero, that is, for which variations in the input variable have a significant impact on results. Because contributions diminish at each step, it eventually becomes questionable whether the model is describing a real statistical dependence or natural variability. To decide, a quantitative statistical criterion is invoked to determine when additional contributions are small enough to be indistinguishable from natural variability. Normally, the t-distribution (Neter, et al., 1983) is used to determine the probability that a linear relationship is significantly different from zero, meaning that the regression analysis could be due to chance or natural variability. The probability of obtaining a chance relationship at least that strong is referred to as the α-value. In STEPWISE the default α-values are set at 0.05 for a variable to enter the model (sigin), and 0.05 to leave (sigout). The user may specify different α-values in the input control file, however the value allowing a variable to enter the model must be less than or equal to the value by which a variable is allowed to leave the model to avoid looping.

The order in which the input variables should be added to the STEPWISE regression model is not obvious. The largest sample correlation coefficient (see Section 4.2.6) identifies unambiguously the first input parameter in the process. Thereafter, one requires additional direction, which, in STEPWISE, is provided by the relative magnitudes of the remaining partial correlation coefficients (see Section 4.2.8).

4.2.6 Sample Correlation Coefficients

Correlation coefficients measure the strength of the linear relationship between two variables, and are closely related to standardized regression coefficients. If, for example, we wish to quantify the strength of the linear relationship between the input variable X_j and the calculated result Y, based on a sampled parameter set X_{ij}, i = 1,..., N, j =1,..., P and the corresponding calculated results Y_i, i = 1,...,N, we would normally compute the sample correlation coefficient (SCC), r_{yj}, based on that sample.

The sample correlation coefficient, defined below, is based on all the sampled values of the j^{th} input parameter and all the calculated results that correspond to those values, that is,

$$r_{yj} = \frac{\sum_{i=1}^{N} X_{ij} Y_i - \sum_{i=1}^{N} X_{ij} \sum_{i=1}^{N} Y_i / N}{\left\{ \left[\sum_{i=1}^{N} X_{ij}^2 - \left(\sum_{i=1}^{N} X_{ij} \right)^2 / N \right] \left[\sum_{i=1}^{N} Y_i^2 - \left(\sum_{i=1}^{N} Y_i \right)^2 / N \right] \right\}^{1/2}} .$$

The value of r_{yj} can vary between positive one and negative one. When r_{yj} is close to positive one this indicates a positive linear relationship exists between the input variable X_j and the calculated result Y. When r_{yj} is close to negative one, this indicates a negative linear relationship exists between the input variable X_j and the calculated result Y. Values of r_{yj} close to zero indicate that no linear relationship exists between the input variable X_j and the calculated result Y. However, that does not mean the variables are statistically independent as their relationship, if one exists, may be defined in some other manner.

The largest of the r_{yj} (in absolute value) over j = 1,..., P, serves to identify the input variable having the strongest linear relationship with Y. STEPWISE chooses that variable as a starting point on which to construct a stepwise regression. However, once the model contains one variable X_j the sample correlation coefficient is no longer an adequate indicator of the next variable to enter the model. Partial correlation coefficients (PCCs) are the quantitative measures used by STEPWISE to establish the next variable to enter. Before defining PCCs, however, it is helpful to define residuals, which are used in their calculation.

4.2.7 Residuals

A residual, ε_i, is defined as the difference between the observed value Y_i, i = 1,...,N, and the fitted value $\hat{Y}_i$, i = 1,..., N resulting from evaluating the final regression model with the i^{th} input variable vector. That is, $\varepsilon_i = Y_i - \hat{Y}_i$, i = 1,...,N. As such, the residual may be regarded as the observed error, an estimate of the unknown true error in the regression model, or as that part of the variability in a scatter plot that is not accounted for by the regression model.

For regression models, the ε_i are assumed to be normally distributed random variables, with mean zero, and constant variance σ^2, which are uncorrelated, or statistically independent. If the regression model is appropriate for the independent and dependent variables under consideration, the residuals obtained from the model will exhibit these statistical characteristics, and procedures used in hypothesis testing can be used to test the validity of the regression model. Although these methods are very helpful in developing regression models, they are not very useful in sensitivity analysis because the assumptions required for their use generally cannot be met.

However, our present interest in residuals is not for their inherent utility in analyzing the success of the regression method, but rather (1) to explore their role in partial correlation coefficients, and (2) to explain how and why partial correlation coefficients help direct STEPWISE in calculating a sequence of regression models.

4.2.8 Partial Correlation Coefficients

Partial correlation coefficients (PCCs) are suitable measures to account for the unique relationship between two variables, that cannot be measured in terms of the relationship between those two variables and any other variables. Therefore, partial correlation coefficients provide an effective means of identifying the additional variables that should be added to an existing stepwise regression model. The following example should serve to clarify this.

Assume that P = 4, so there are four independent variables $X_1, X_2, X_3, X_4,$ along with one dependent variable Y whose values are observed. Further, assume that X_1 is found to have the largest sample correlation coefficient (SCC) with the dependent variable Y, so X_1 will be the first variable to enter the stepwise regression model. Then regressing Y on X_1 results in the equation,

$$\hat{Y} = b_0 + b_1 X_1 + \varepsilon_1,$$

where the subscript on the error term ε_1 will correspond to the independent variable X_1. The next step is to find which one of the remaining variables $X_2, X_3,$ or X_4 should enter the model next. This can no longer be done by calculating the SCC between the dependent variable Y and the remaining independent variables, because a portion of the variablity seen in the data has been explained by the model based on X_1. The partial correlation coefficients (PCCs) do not suffer from this restriction, and are calculated according to the following.

As X_1 is already in the model, new variables $X_2^*, X_3^*,$ and X_4^* are found by regressing the variables $X_2, X_3,$ and X_4 against X_1 resulting in,

$$X_2^* = c_0 + c_1 X_1 + \varepsilon_2$$

$$X_3^* = d_0 + d_1 X_1 + \varepsilon_3$$

$$X_4^* = e_0 + e_1 X_1 + \varepsilon_4 .$$

The PCCs corresponding to each variable are found by calculating SCCs on the residuals between the models. That is, the PCC corresponding to X_2 is found by calculating the SCC between ε_1 and ε_2, the PCC corresponding to X_3 is found by calculating the SCC between between ε_1 and ε_3, and finally the PCC corresponding to X_4 by calculating the SCC between ε_1 and ε_4. The next variable to enter the model will be the variable with the largest PCC.

Once the second variable has entered the model, the same procedure will be repeated, only this time the independent variables remaining outside the model will be regressed against the two variables already contained in the model. The PCCs will be calculated on the residuals, and the third variable to enter the model will correspond to the variable with the largest PCC as before. As this example only has four independent variables, there is no reason to calculate the PCC for the fourth and final variable, as it will be the next to enter the model by default. The extension of this method to more than four independent variables will follow the same pattern as presented here.

In addition, this discussion has assumed that any variable that is chosen to enter the regression model based on the size of its PCC is sufficiently significant at the user-supplied α value (sigin) that it would be allowed to enter. If the variable were not, it would be passed over and the next variable would be considered.

4.2.9 Rank Regression Coefficients

When nonlinear relationships are involved in a physical model, as in WIPP PAs, it is often more revealing to calculate regression coefficients on the ranks of the numerical values corresponding to a variable, rather than on the actual numerical values of the variable itself. Such coefficients are known as rank regression coefficients. Specifically, the smallest sampled value of each variable is assigned the rank 1, the next smallest value is assigned rank 2, and so on up to the largest value, which is assigned the rank N, where N denotes the sample size. In the case of ties, the average value will be assigned to each of the tied values. For example, assume that the third, fourth, and fifth sampled values are tied. In this case each would be assigned the rank (3+4+5)/3 = 4. The regression coefficients are then calculated using the rank indices (i.e., the numbers 1, 2,...,N) in place of the sampled data X_{ij}, where j = 1,...,P and i = 1,...,N. Because LHS sampling has been used, there can be no ties amongst the independent variable values, but ties may occur between values associated with the dependent variable. A thorough discussion of rank regression analysis has been given by Iman and Conover (1979).

4.2.10 Overfitting the Data

For computer models requiring the use of large numbers of input variables and large sample sizes, it is possible, in principle, for a regression model to attempt to fit specific data points rather than fitting the general shape of the data distribution. This so-called "overfitting" can produce a spurious model that leads to poor predictions regarding model sensitivities. Consequently, a statistical parameter called the predicted error sum of squares (PRESS) has been defined to check for overfitting.

If a model treats K input variables and N samples for each variable, PRESS (a number that depends on K and N) is calculated by removing the m^{th} observation from the sample set and calculating the regression model on all K variables but only N - 1 samples. The value of $\hat{y}_m$ for the missing sample X_{mj}, j = 1,...,K, m fixed, is then calculated from the regression model according to,

$$\hat{y}_m = b_o + \sum_{j=1}^{K} X_{mj} b_j \quad .$$

The process is repeated for every value of m, from 1 to N. PRESS, which depends only on the number K and the specific choice of input parameters employed in the regression model, is then defined as follows:

$$\mathrm{PRESS} = \sum_{m=1}^{N} (y_m - \hat{y}_m)^2 \quad .$$

The regression model having the smallest value of PRESS is preferred when overfitting is a possibility. STEPWISE calculates and reports PRESS, if requested by the user in the input control file.

Additional details on stepwise regression can be found in the original user's guide for the regression program known as STEPWISE (Iman, et al., 1980) and in a WIPP-related sensitivity-analysis paper by Helton et al. (1991).

5.0 CAPABILITIES AND LIMITATIONS OF THE SOFTWARE

STEPWISE capabilities include the ability to: 1) perform stepwise regression; 2) perform backward regression; 3) perform forward regression, which is a special case of the stepwise regression capability; 4) force variables to be included in the regression model; 5) drop observations from the regression model; 6) perform regression on the ranks of the data; 7) calculate the predicted sum of squares (PRESS) values for each model fitted; 8) perform weighted regression; and 9) perform regression on standardized variables. The output can include: mean, variances, standard deviations, standard errors, and coefficients of variation for the independent or transformed variables; correlation coefficients, residuals, dependent variable predictions, PRESS values, analysis of variance tables, and statistics for the regression coefficients.

STEPWISE can produce several different types of scatter plots, such as plots of each dependent-variable / independent-variable pair, and plots of the residuals for each dependent variable. STEPWISE can also create an ASCII text file of the plot data that can be input into the SigmaPlot program. There is a STEPWISE limitation of 1000 columns of data on this file. The transfer and use of such a large file by SigmaPlot is beyond the scope of STEPWISE.

STEPWISE has some capabilities that have not been tested and cannot be used in regulatory calculations. These are identified as such in the descriptions of the commands which exercise these capabilities in Section 7.1. The only untested capabilities mentioned in this section are the abilities to perform weighted regression and regression on standardized variables.

STEPWISE makes extensive use of dynamic memory. Because of this, there are no specific limitations on the problem size parameters (e.g., the number of independent variables or the number of observations).

6.0 USER INTERACTIONS WITH THE SOFTWARE

STEPWISE executes on a DEC Alpha under OpenVMS. To execute STEPWISE, type STEPWISE at the OpenVMS system "$" prompt and press the carriage-return key. The names of two graphics devices and five files will be requested sequentially by STEPWISE. Alternately, the user may append the names of the files (in order) to the STEPWISE command line before pressing the carriage-return key. The required information is as follows:

1. A terminal device must be selected from a list of available devices. If the NULL option is selected, a plot will generate no terminal output. The plot capabilities are **not tested** on any terminal devices. Thus, NULL is the only valid option in regulatory calculations.

2. A hardcopy device must be selected from a list of available devices. If the NULL option is selected, a plot will generate no hardcopy output. The hardcopy output is written to a plot file (or a set of files). On the command line the file name may be appended to the device code with a "@" (e.g., `ADOBE@PCC_TEST.ADOBE`). ADOBE creates a separate file for each plot by adding an underscore, "_", followed by the plot number to the file extension (e.g., PCC_TEST.ADOBE_1). The plot file(s) may be directed to the hardcopy device after STEPWISE is exited. The plot capabilities are **tested only on the ADOBE device**. Thus, NULL and ADOBE are the only valid options in regulatory calculations.

3. The input control file. This file specifies STEPWISE's processing options. The contents of this file are explained in detail in Section 7.

4. The input independent-variables data file. This file is normally generated by and transferred directly from the WIPP PA code LHS2STEP. This file may also be specified with the **untested** IND_FILE command.

5. The input dependent-variables data file. This file is normally generated by and transferred directly from the WIPP PA code CCD2STEP. This file may also be specified with the **untested** DEP_FILE command.

6. The output text file. This file contains a summary of the input, including any errors found, and the results of the STEPWISE analysis.

7. The text plot file. This file contains the requested plot information in tabular form for input into the SigmaPlot plotting program. This file is optional. The file will not be produced if the file name is input as "CANCEL".

For example, the following DCL command would execute STEPWISE:

```
$ STEPWISE NULL ADOBE@STP_TEST.ADOBE STP_TEST.INP -
     STP_TEST_IND.DAT STP_TEST_DEP.DAT -
     STP_TEST.OUT CANCEL
```

The full file name would be included in the STEPWISE command for any file that was not in the current directory.

The preceding information assumes that STEPWISE is defined on the system to be run by the PLT_PROGRAM_DRIVER. This driver sets up the plot devices and then runs the program. If the STEPWISE symbol does not include the PLT_PROGRAM_DRIVER but runs STEPWISE directly, then the terminal device and the hardcopy device (items 1 and 2 above) should be omitted from the command line and the user should issue the command `SELECT_DEVICES` with items 1 and 2 as parameters before executing STEPWISE. The following DCL commands run STEPWISE in this case:

```
$ SELECT_DEVICES NULL ADOBE@STP_TEST.ADOBE
$ STEPWISE STP_TEST.INP -
     STP_TEST_IND.DAT STP_TEST_DEP.DAT -
     STP_TEST.OUT CANCEL
```

If the input command file requests that any plots be generated, one of the following must be specified: a terminal device, a hardcopy device, or a SigmaPlot file. If a terminal device is specified, the plots will be generated on the terminal and, if interactive input is possible, a response will be requested from the user after each plot. The user can direct the displayed plot to the hardcopy file, if a hardcopy device has been specified. If both a terminal device and a hardcopy device are specified, this is the only way to generate hardcopy plots. If a hardcopy device is specified and the terminal device is not, the hardcopy plots are generated automatically with no user intervention.

If a SigmaPlot file is specified, all plot data will be output to the SigmaPlot file regardless of whether any devices are specified. "Plotting a curve" to a SigmaPlot file consists of writing up to two columns of data (one for the X data and one for the Y data). STEPWISE will try not to write a column of data that it has previously written, but some duplication may occur.

7.0 DESCRIPTION OF THE INPUT FILES

7.1 Input Control File

The input control file specifies the processing options for a given run of STEPWISE. The generic format of STEPWISE's input control file is described in detail by Rechard (1992; Section 2.9.4). Each command begins with a keyword, followed by options that apply for the command. A command may be continued on the next line by inserting an ampersand, "&", as the last character on the previous line. Any characters following an exclamation mark, "!", are ignored and treated by the code as comments.

The commands are described in detail below. The following paragraphs contain a brief summary of the more important commands. An example of an input control file that uses many of STEPWISE's commands is given in Appendix A.

The STEPWISE and BACKWARD commands control the regression analysis selection method. The RANK command determines whether the coefficients are calculated using the ranks of the data or the raw data. The DEP_VARS and IND_VARS commands select the variables to be processed. The OUTPUT command controls the information to be written to the output file.

The PLOT command controls whether plots are generated for the analysis. The SCATTER command causes scatter plots to be generated. The plots will be directed to some combination of a terminal, a hardcopy device, and a SigmaPlot file, depending on the execution directive described in Section 6.0. Commands that control the appearance of the plot (such as YLOG and XSCALE) do not affect the output to the SigmaPlot file.

A single execution of STEPWISE may contain multiple "data sets" that use the same input data files, but have different processing options. The END command ends the current data set and the next command in the input control file is part of the next data set. Most processing options do not change between data sets. If that is not the case, it is noted under the command that sets the processing option. If a single command is issued twice in the same data set, the first command is usually ignored. Any exception to this rule (e.g., a cumulative command) is noted under the appropriate command description. The user should exercise care when using the multiple data-set capability. The processing options should be checked carefully.

The format of each command is shown in the command description. Many of the commands expect parameters. Unless otherwise noted, all command parameters are required and cannot be omitted. Numeric parameters must be in a valid numeric format, but are not checked for reasonableness unless otherwise noted. If a parameter expects a value from a set of options, the value must be in the set. This includes variable identifiers, which must refer to valid variables by their label or number. There are several cumulative commands identified below. All cumulative commands expect a series of values, such as the variable parameters in the IND_VARS command. If no values are given, the command is ignored. If any value in a series is invalid, that value is ignored and a warning message is printed, but the command still processes the other values. If a parameter violates any of the above rules (with the exception of a series value), an error message is printed and the command is ignored. The command verbs and most parameter options can be abbreviated, as long as enough characters are given to uniquely identify the word. Variable labels and ON/OFF cannot be abbreviated.

STEPWISE'S INPUT CONTROL FILE COMMANDS

`BACKWARD` *`sigin`*

BACKWARD selects the backward elimination solution method for "best" subset selection of the regression-analysis sequence. The optional parameter *sigin* specifies the significance level. This parameter is not checked for reasonableness. If the parameter is omitted, the significance level is set to 0.05.

Since commands STEPWISE and BACKWARD both specify the analysis method, both commands cannot be present in the same data set. If both are present, a warning is printed and the last command supersedes the earlier one. Either a STEPWISE or BACKWARD command must be present in the first data set (unless the SCATTER command is specified), otherwise the resulting processing is untested and is not documented in this manual.

`DEP_FILE` *`file`*

DEP_FILE sets the name of the data file that contains the dependent variables. The *file* parameter must be the name of a valid data file or an error occurs. In regulatory work, this name is set on the command line. If the file is specified in both ways, the DEP_FILE command will override the selection on the command line, but a warning message will be displayed.

The DEP_FILE command **has not been tested** and cannot be used in regulatory calculations.

`DEP_VARS` *`id_1, ..., id_n`*

DEP_VARS selects those dependent variables to be included in the analysis. The *id* may be the variable's label or its number. This command is cumulative. A variable will be selected if it appears in any DEP_VARS command within a single data set. If no DEP_VARS command is present in the current or any previous data set, all dependent variables on the input data file will be selected.

`DROP_OBS` *`obs_1, ..., obs_n`*

DROP_OBS drops the listed observations from the regression analysis. If no DROP_OBS command is present in the current or any previous data set, no observations will be dropped. This command is cumulative. An observation will be dropped if it appears in any DROP_OBS command within a single data set. If no DROP_OBS command is present in the current or any previous data set, all observations will be included in the analysis.

`DUMP` *`id_1, ..., id_n`*

DUMP requests that the data used in the regression analysis (raw data, rank data, or standard 0-1 values) be written to the output text file. The values are written for the specified variables only. The *id* may be the variable's label or its number. If there are no parameters given for this command, the values for all selected independent and dependent variables are written. The raw data is written first if the rank or standard 0-1 values are to be written. This command is a debugging aid and is not enabled unless the DUMP command is present in the current data set.

`END`

END ends the "data set" and starts processing. Another data set with different parameters may follow the END command. Processing options are not reset at the start of a data set unless otherwise noted in the command description.

`FORCE` *`id_1, ..., id_n`*

FORCE causes the specified variables to be included in the model regardless of their contributions to the model. The *id* may be the variable's label or its number. This command is cumulative. A variable will be forced to be included if it appears in any FORCE command within a single data set. If no FORCE command is present in the current or any previous data set, STEPWISE will not force any variables to be included.

`IND_FILE` *`file`*

IND_FILE sets the name of the data file that contains the independent variables. The *file* parameter must be the name of a valid data file or an error occurs. In regulatory work, this name would normally be set on the command line. If the file is specified in both ways, the IND_FILE command will override the selection on the command line, but a warning message will be displayed.

The IND_FILE command **has not been tested** and cannot be used in regulatory calculations.

`IND_VARS` *`id_1, ..., id_n`*

IND_VARS selects the independent variables to be included in the analysis. The *id* may be the variable's label or its number. This command is cumulative. A variable will be selected if it occurs in any IND_VARS command within a single data set. If no IND_VARS command is present in the current or any previous data set, all independent variables on the input data file will be selected.

`OUTPUT` *`opt_1, ..., opt_n`*

OUTPUT selects the information to be written to the output text file. The analysis-of-variance table and regression-coefficient estimates for the final model for each dependent variable will always be written. The following options are available:

MEAN	Write the mean and standard deviations for all variables in the model.
CORR	Write the simple correlations among all variables in the model.
SSCP	Write the corrected sum-of-squares and cross-products matrix for all variables in the model. This option **has not been tested** and cannot be used in regulatory calculations.
INVERSE	Write the inverse correlation matrix at each analysis step (if the STEPS option is specified) and for the final model for all independent variables. This option **has not been tested** and cannot be used in regulatory calculations.
STEPS	Write the analysis-of-variance table and the regression-coefficient estimates for each step of the variable selection.
RESIDUALS	Write the dependent variable, Y, the portion of the dependent variable accounted for by the regression model, $\hat{Y}$, and the residual (their difference) for each observation using the final regression model.
OTHER	Write miscellaneous regression-analysis information, such as the standard error and the coefficient of interpolation. This option **has not been tested** and cannot be used in regulatory calculations.
ALL	Write the information requested for all the options above. This option **has not been tested** and cannot be used in regulatory calculations.
OFF	No special output options are requested.

This command is cumulative. An option will be selected if it appears in any OUTPUT command within a single data set. If no OUTPUT command is present in the current or any previous data set, no special output options are requested (OUTPUT OFF).

`PRESS ON or OFF`

PRESS ON (or just PRESS) specifies that the predicted error sum of squares criterion be used to determine the adequacy of the prediction model. PRESS OFF disables this option. If the no PRESS command is present in the current or any previous data set, this option will be disabled.

`QAAID "`*`qaaid`*`"`

QAAID sets the QA aid that is displayed at the bottom of the plot to *qaaid*. If no QAAID command is present in the current or previous data set, the QA aid is the name of the input control file. This command **has not been tested** and cannot be used in regulatory calculations.

`PLOT` *`opt_1, ..., opt_n`*

PLOT selects the plots to be generated during the regression analysis. The following options are valid:

RAW	Plot the raw data (versus the rank or standard 0-1 data) on any data plot. This option does not generate plots by itself; but is used in conjunction with the SCATTER and RESIDUALS options below. This option is not set by the ALL option and it does not retain its setting between data sets; it must be explicitly included in the current data set.
PRESS	Plot the step number versus the predicted error sum of squares (PRESS) as a scatter plot. One plot will be generated for each dependent variable in the model.
SCATTER	Plot dependent-variable/independent-variable scatter plots for variables selected by the analysis. If the PLOT RAW command is issued in the current data set, the raw data will be plotted; otherwise the data used in the analysis will plotted.
RESIDUALS	Plot a set of scatter plots with the following data versus the residuals: the time, the $\hat{Y}$ value, and each independent variable. If the PLOT RAW command is issued in the current data set, the raw data will be plotted for the independent variable plots; otherwise the data used in the analysis will be plotted. One set of plots will be generated for each dependent variable in the model.
ALL	All the plots requested for the options above will be generated.
OFF	No plots will be generated by the regression analysis.

This command is cumulative. An option will be selected if it occurs in any PLOT command within a single data set. If no PLOT command is present in the current or any previous data set, no plots will be generated (except by the SCATTER command).

`RANK ON or OFF`

RANK ON (or just RANK) specifies that the regression analysis be performed using the ranks of the data. RANK OFF specifies that raw data is to be used.

Since commands RANK and STAND01 both specify the data type for the analysis, both commands cannot be present in the same data set. If no RANK and STAND01 command is present in the current or any previous data set, raw data values will be used.

`SCATTER ALL or PARTIAL`

SCATTER PARTIAL (or just SCATTER) specifies that scatter plots be generated for all selected dependent-variable/independent-variable pairs. SCATTER ALL specifies that scatter plots be generated for all selected variable pairs including dependent-variable/dependent-variable pairs and independent-variable/independent-variable pairs. The raw or ranked or standard 0-1 data may be plotted, depending on the RANK and STD01 settings.

The scatter plots are only generated if the SCATTER command is present in the current data set. If no STEPWISE or BACKWARD command is included in the data set, the scatter plots will be generated but no regression analysis will be done. The PLOT command (including PLOT RAW and PLOT SCATTER) has no effect on this command.

`SOFTCHAR ON or OFF, TERMINAL or HARDCOPY or ALL`

SOFTCHAR ON sets the character output mode for all plots to software characters; SOFTWARE OFF sets hardware characters. Hardware characters are drawn faster, but may produce inferior lettering on some devices. Some characters (such as the axis numbering) will not be affected by this command.

The user may set the output mode on the terminal, the hardcopy device, or both devices with the TERMINAL, HARDCOPY, and ALL parameter, respectively. If this parameter is omitted, both devices are set.

If no SOFTCHAR command is present in the current or any previous data set, hardware characters will be output on both the terminal and the hardcopy device. This command **has not been tested** and cannot be used in regulatory calculations.

`STAND01 ON or OFF`

STAND01 ON (or just STAND01) specifies that the regression analysis be performed using standard 0-1 values of the data. STAND01 OFF (the equivalent of RANK OFF) specifies that raw data be used. The standard 0-1 data is linearly rescaled to have a mean of zero and a standard deviation of one.

Since commands RANK and STAND01 both specify the data type for the analysis, both commands cannot be present in the same data set. If no RANK and STAND01 command is present in the current or any previous data set, raw data values will be used.

The STAND01 command **has not been tested** and cannot be used in regulatory calculations.

`STEPWISE` *`sigin`*`,` *`sigout`*

STEPWISE selects the stepwise procedure for "best" subset selection. The optional parameters *sigin* and *sigout* specify the significance levels. If *sigout* is set to 1.00, STEPWISE performs a forward regression procedure. If *sigin* is omitted, the value will be set to 0.05. If *sigout* is omitted, the value will be set to the *sigin* value. Neither value is checked for reasonableness, although *sigout* must be greater than or equal to *sigin*.

Since commands STEPWISE and BACKWARD both specify the analysis method, both commands cannot be present in the same data set. If both are present, a warning is printed and the last command supersedes the earlier one. Either a STEPWISE or BACKWARD command must be present in the first data set (unless the SCATTER command is specified), otherwise the resulting processing is untested and is not documented in this manual.

`TITLE "`*`title`*`"`

TITLE defines the title that appears as the heading of the output text file and as the title for all plots generated. If no TITLE command is present in the current or any previous data set, the title used will be the second comment line, if any, from the dependent-variable data file.

`TRANSFORM` *`label = equation`*

TRANSFORM allows the user to define new variables dependent on the values of the variables from the input data files. The *label* parameter is the label assigned to the new variable, which is called an assigned variable. This label is limited to 8 characters and is converted to upper case.

Equation defines the assigned variable in a form much like a FORTRAN assignment statement. The equation may contain other variables, numeric constants, operators, and functions. An equation variable may be either an input or assigned variable. The numeric constants specify real numbers in any valid FORTRAN format, except the double-precision D format. A negative number following an operator must be enclosed in parenthesis. For example, 20, 10.5, .783, 4.87E3, and (-3.7) are all valid constants. All calculations are performed on real numbers. The valid operators are addition (+), subtraction (-), multiplication (*), division (/), and exponentiation (^). FORTRAN precedence applies to these operators and may be changed with parentheses. Functions ABS, LOG, LOG10, EXP, and SQRT perform the standard FORTRAN functions. Function arguments are enclosed in parentheses. Special functions IFLT0, IFEQ0, IFGT0 give a limited IF capability as explained below.

The format of the IF functions is:

```
result = IFLT0 (cond, true_value, false_value)
result = IFEQ0 (cond, true_value, false_value)
result = IFGT0 (cond, true_value, false_value)
```

The function parameter *cond* is evaluated and compared to zero. If the expression is less than zero for IFLT0, equal to zero for IFEQ0, or greater than zero for IFGT0, the result is the *true_value*; otherwise it is *false_value*. For example, the equation

```
RESVAR = IFLT0 (INVAR-100, -999, INVAR)
```

would be implemented in FORTRAN with the following code fragment:

```
DO I = 1, NUMOBS
     TRES = -999.0
     FRES = INVAR(i)
     IF (INVAR(i)-100 .LT. 0) THEN
          RESVAR(i) = TRES
     ELSE
          RESVAR(i) = FRES
     END IF
END DO
```

Note that the equation is applied independently to each observation of the assigned variable and any variables in the equation.

The assigned variable is a dependent variable if any dependent variables are used in the equation. Otherwise, it is an independent variable. The assigned variable is treated exactly like an input variable.

`WEIGHT` *`id`*

WEIGHT specifies that a weighted regression analysis be performed. The weights are from the variable specified by *id*. The value of *id* may be the variable's label or its number. The WEIGHT OFF command disables this option. If no WEIGHT command is present in the current or any previous data set, the data will not be weighted.

The WEIGHT command **has not been tested** and cannot be used in regulatory calculations.

`XLOG ON or OFF`

XLOG ON (or just XLOG) specifies logarithmic scaling on the X axis for all plots except PRESS plots and rank-variable plots; XLOG OFF specifies a linear scale. If no XLOG command is specified in the current or any previous data set, the X axis will have a linear scale.

`XSCALE` *`xmin xmax xtick`*

XSCALE sets the minimum and maximum X axis value and the optional tick interval for all plots except PRESS plots and rank-variable plots. The minimum and maximum are expanded a little on a linear axis so that they do not fall directly on the axis. If no XSCALE command is present in the current or any previous data set, the X axis will be scaled automatically from the data. The axis will also be scaled automatically if both the *xmin* and *xmax* parameters are omitted or if *xmax* is not greater than *xmin*. If the tick interval is not given, it will be automatically calculated. The tick interval parameter **has not been tested** and cannot be specified in regulatory calculations.

`YLOG ON or OFF`

YLOG ON (or just YLOG) logarithmic scaling on the Y axis for all plots except PRESS and rank-variable plots. YLOG OFF specifies a linear scale. If no YLOG command is specified in the current or any previous data set, the Y axis will have a linear scale.

`YSCALE` *`ymin ymax ytick`*

YSCALE sets the minimum and maximum Y axis value and the optional tick interval for all plots except rank-variable plots. The minimum and maximum are expanded a little on a linear axis so that they do not fall directly on the axis. If no YSCALE command is present in the current or any previous data set, the Y axis will be scaled automatically from the data. The axis will also be scaled automatically if both the *ymin* and *ymax* parameters are omitted or if *ymax* is not greater than *ymin*. If the tick interval is not given, it will be automatically calculated. The tick interval parameter **has not been tested** and cannot be specified in regulatory calculations.

7.2 Input Data Files

The independent-variable data file and the dependent-variable data file are transferred directly from other WIPP PA codes. Note that the two data files must contain the same number of observations. The format of the two data files is identical. An abbreviated example of a STEPWISE data file is included in Appendix B.

8.0 ERROR MESSAGES

Error messages resulting from invalid file names are output to the terminal. These messages cause STEPWISE to abort. Other error messages, usually the result of an invalid control-file commands or invalid data files, are output to the output text file. If any errors or warnings are detected, a single message to that effect will be output to the terminal. Errors cause the program to abort at some point; warnings do not. Warnings signal the presence of an unusual but non-fatal condition about which the user should be aware. All error and warning messages are directly preceded by the characters "%%%".

The input control file that is read by STEPWISE is also echoed to the output text file. If an input command contains an error, an error message will appear on the output text file directly below the incorrect input command and the command will be ignored. A warning message may also appear. A command that causes a warning will be executed unless otherwise noted in the warning. The error or warning message should provide sufficient explanation to identify the problem.

The output text file also contains a listing of most input control-parameter settings. Errors may also be displayed at this stage. If errors are detected up to this point, the program will abort.

If no errors are detected on the input control file, STEPWISE will process the input data files as instructed. Any errors detected during processing will cause a descriptive error message to be printed to the output text file. Such errors may cause STEPWISE to abort.

If multiple data sets are included in the input control file, STEPWISE will read and process the current data set before proceeding to the next data set.

9.0 DESCRIPTION OF THE OUTPUT FILES

9.1 Output Text File

The output text file contains a summary of the information contained on STEPWISE's input control file. Following this summary is the information requested by the OUTPUT command in the input control file. The file will always contain the analysis-of-variance table and the regression-coefficient estimates for the final model of each dependent variable. Other output can be requested through the OUTPUT command.

An example of STEPWISE's output text file may be found in Appendix C.

9.2 Output Hardcopy Plot File(s)

The plot output for a hardcopy device is written to a file. The ADOBE hardcopy device creates a set of files, one for each plot. The plot file is not human-readable. The file can be transferred to the appropriate hardcopy device or imported into documents, depending on the device type.

9.3 Output SigmaPlot File

The output SigmaPlot text file contains up to two columns of data for each curve plotted. STEPWISE attempts to remove duplicate columns of data from the file. Each row corresponds to a point on the curve. Up to 1000 columns of data can be written to the file. This file is an ASCII file, but because of the possibly large line length, it may be difficult to edit. The file may be input into the SigmaPlot program and other programs which expect column data.

10.0 REFERENCES

Draper and Smith, 1981. Applied Regression Analysis, Second Edition. John Wiley & Sons, New York, New York.

Helton, J. C., J. W. Garner, R. D. McCurley, and D. K. Rudeen. (1991). *Sensitivity Analysis Techniques and Results for Performance Assessment at the Waste Isolation Pilot Plant, March 1991*. SAND90-7103. Sandia National Laboratories, Albuquerque, NM.

Iman, R. L., J. M. Davenport, E. L. Frost, and M. J. Shortencarier. (1980). *Stepwise Regression with PRESS and Rank Regression (Program User's Guide).* SAND79-1472. Sandia National Laboratories, Albuquerque NM.

Iman, R. L. and W. J. Conover. (1979). *Use of the Rank Transform in Regression.* **Technometrics, 21**, 499-509.

Neter, Wasserman, and Kutner, 1983. Applied Linear Regression Models. Irwin, Homewood, Illinois, 60430.

Rechard, R. P., ed. (1992). *User's Reference Manual for CAMCON: Compliance Assessment Methodolgy Controller; Version 3.0.* SAND90 - 1983. Sandia National Laboratories, Albuquerque NM.

WIPP Performance Division, 1991. Preliminary Comparison with 40 CFR Part 191, Subpart B for the Waste Isolation Pilot Plant, December 1991, Volume 1: Methodology and Results. SAND91-0893/1, Albuquerque, NM: Sandia National Laboratories.

WIPP PA Department (1992a). *Preliminary Performance Assessment for the Waste Isolation Pilot Plant, December 1992. Volume 1: Third Comparison with 40 CFR 191, Subpart B.* SAND92 - 0700/1. Sandia National Laboratories, Albuquerque NM.

WIPP PA Department (1992b). *Preliminary Performance Assessment for the Waste Isolation Pilot Plant, December 1992. Volume 2: Technical Basis.* SAND92 - 0700/2. Sandia National Laboratories, Albuquerque NM.

WIPP PA Department (1992c). *Preliminary Performance Assessment for the Waste Isolation Pilot Plant, December 1992. Volume 3: Model Parameters.* SAND92 - 0700/3. Sandia National Laboratories, Albuquerque NM.

WIPP PA Department (1992d). *Preliminary Performance Assessment for the Waste Isolation Pilot Plant, December 1992. Volume 4: Uncertainty and Sensitivity Analyses for 40 CFR 191, Subpart B.* SAND92 - 0700/4. Sandia National Laboratories, Albuquerque NM.

WIPP PA Department (1992e). *Preliminary Performance Assessment for the Waste Isolation Pilot Plant, December 1992. Volume 5: Uncertainty and Sensitivity Analyses of Gas and Brine Migration for Undisturbed Performance.* SAND92 - 0700/5. Sandia National Laboratories, Albuquerque NM.

APPENDIX A: EXAMPLE OF THE INPUT CONTROL FILE

The following is an example of a STEPWISE input control file.

```
!==================================================================
!
! STEPWISE input command file for Test Case #1
!
!==================================================================
!
TITLE "Stepwise Test Case Output: RANK "
!
STEPWISE .02, .05
RANK
PRESS
DEP_VARS E1NP237, E1TOTAL
IND_VARS CULTORT  CULMD0   CULFRCSP CULBCRCH CULPREC
IND_VARS BHLGAREA BHLGPRES TIMINTR1
IND_VARS RTRDMPU RTRDMAM  RTRDMNP  RTRDMU   RTRDFPU  RTRDFAM  RTRDFNP  RTRDFU
END
!
TITLE "Stepwise Test Case Output: RAW"
RANK OFF
END
!
!==================================================================
```

APPENDIX B: EXAMPLE OF INDEPENDENT-VARIABLE DATA FILE

The following is a portion of the LHS2STEP output transfer file that would be transferred to STEPWISE as its input independent-variable data file. The format of the file is explained below. Note that the format of the independent-variable data file is identical to the format of the dependent-variable data file.

All the numbers in the file are input using a FORTRAN free-format READ. Any valid FORTRAN number (integer or real, as appropriate) is acceptable. Numbers must be separated by one or more spaces or line breaks. The variable labels may be separated by one comma or any number of spaces or tabs or line breaks. Blank lines may be inserted at any point for readability. Any number of numbers or labels may appear on a line. Any number of release values may appear on one line, but releases for each observation must be on a separate line.

Explanatory Comment: The first line of the file is the file version number. The version is not used at this time, but it should be set to 5 to prevent problems if the version is revised in the future. One or more comment lines follow the version number. A comment line is denoted by an exclamation mark, "!", in column 1.

```
5
! STEPWISE independent data test file - LHS sampled output
```

Explanatory Comment: The next line specifies that there are 29 variables, 40 observations, and one step. There is always one only one step in the STEPWISE input data file. The cluster of six lines following lists the labels that STEPWISE will use to identify the 29 variables. The 1.0 on the last line is simply a placeholder.

```
29      40        1
   SALCAP         SALPERM        RMLGPRES       SOLALL         TIMINTR1
   BHLGPRES       BHHYCND        BHPOR          BPVOL          CULTORT
   CULMD0         CULFRCSP       CULBCRCH       CULPREC        BHLGAREA
   RTRDMPU        RTRDMAM        RTRDMNP        RTRDMU         RTRDFPU
   RTRDFAM        RTRDFNP        RTRDFU         CUL1HYCN       CUL2HYCN
   CUL3HYCN       CUL4HYCN       CUL5HYCN       CUL7HYCN
   1.000000E+00
```

Explanatory Comment: The clusters of lines that follow specify the release data. Each cluster contains 29 numbers, one for each of the variables in the order specified above. There are 40 clusters, one for each of the observations. Only three of the 40 clusters are shown below.

```
   5.711000E-11   2.933000E-21   1.090000E+07   1.476000E-05   1.026000E+11
   1.397000E+07   3.164000E-05   3.840000E-01   1.115000E+07   3.112000E-02
   1.752000E-10   6.330000E+00   1.620000E-01   5.220000E-01   3.142000E-02
   7.301000E+03   5.121000E+03   1.170000E+02   1.010000E+01   2.005000E+04
   2.780000E+02   1.700000E+01   9.320000E+00   1.042000E-05   2.816000E-08
   1.831000E-07   8.296000E-08   4.438000E-06   1.556000E-04

   1.939000E-11   4.500000E-21   1.496000E+07   2.943000E-07   6.249000E+09
   8.397000E+06   4.858000E-04   3.740000E-01   2.245000E+06   9.380000E-02
   3.767000E-10   2.320000E+00   3.590000E-01   5.783000E-02   3.142000E-02
   1.295000E+03   2.870000E+02   8.030000E+00   1.550000E+01   1.274000E+04
   9.120000E+01   1.930000E+00   1.240000E+01   2.320000E-05   1.338000E-08
   1.367000E-07   5.326000E-08   4.164000E-06   1.685000E-05

   1.205000E-11   1.459000E-20   8.317000E+06   4.963000E-09   1.569000E+11
   9.141000E+06   2.094000E-04   2.870000E-01   9.393000E+06   3.505000E-02
   3.204000E-10   1.260000E+00   6.810000E-01   1.960000E+00   7.604000E-02
   4.738000E+03   2.579000E+03   9.630000E+00   3.020000E+01   1.210000E+02
```

```
   6.060000E+02   1.670000E+00   1.950000E+01   2.046000E-05   3.747000E-08
   1.326000E-07   4.659000E-08   4.703000E-06   9.794000E-05
```

Explanatory Comment: The data for observations 5 to 40 would appear here. The observations are omitted for brevity.

APPENDIX C: EXAMPLE OF THE OUTPUT TEXT FILE

The following is an example of STEPWISE's output text file.

```
STEPWISE_PA96 2.20 PROD PA96 05/30/96           06/07/96 14:48:34   Page  1

                                  STEPWISE_PA96

  SSSSSS TTTTTT  EEEEEEE PPPPPP  WW   WW   IIII   SSSSSS EEEEEEE         PPPPPP
 SS        TT    EE      PP   PP WW   WW    II   SS      EE              PP   PP
 SS        TT    EE      PP   PP WW   WW    II   SS      EE              PP   PP
  SSSSS    TT    EEEEE   PPPPPP  WW   WW    II    SSSSS  EEEEE           PPPPPP
      SS   TT    EE      PP      WW W WW    II        SS EE              PP
      SS   TT    EE      PP      WWWWWWW    II        SS EE              PP
 SSSSSS    TT    EEEEEEE PP       WW WW    IIII  SSSSSS  EEEEEEE _______ PP

                            STEPWISE_PA96 Version 2.20
                              PROD PA96 Built 05/30/96
                  Written by Kansas State University, Amy Gilkey
                              Sponsored by Amy Gilkey

                            Run on 06/07/96 at 14:48:34
                        Run on ALPHA AXP BEATLE OpenVMS V6.1

           ************************************************************

                                    Disclaimer
                                    ----------

           This computer program was prepared as  an  account  of  work
           sponsored  by  an  agency  of  the United States Government.
           Neither the United States Government nor any agency thereof,
           nor  any  of  their employees, nor any of their contractors,
           subcontractors, or  their  employees,  makes  any  warranty,
           express  or  implied,  or  assumes  any  legal  liability or
           responsibility for the accuracy, completeness, or usefulness
           of any information, apparatus, product, or process disclosed
           or represents that its  use  would  not  infringe  privately
           owned  rights.   Reference herein to any specific commercial
           product, process,  or  service  by  trade  name,  trademark,
           manufacturer,  or otherwise, does not necessarily constitute
           or imply its endorsement, recommendation, or favoring by the
           United States Government, any agency thereof or any of their
           contractors  or  subcontractors.   The  views  and  opinions
           expressed  herein  do not necessarily state or reflect those
           of the United States Government, any agency thereof  or  any
           of their contractors or subcontractors.

           ************************************************************


 ********************************************************************************

 FILE ASSIGNMENTS:
 -----------------
 Input control file:
    U1:[APGILKE.0CMS.STEPWISE.TEST]STP_TEST.INP;5
       Written on 06/07/96 14:48:30
 Input independent data file:
    U1:[APGILKE.0CMS.STEPWISE.TEST]STP_TEST_IND.DAT;1
       Written on 03/26/96 08:02:05
 Input dependent data file:
    U1:[APGILKE.0CMS.STEPWISE.TEST]STP_TEST_DEP.DAT;1
       Written on 03/26/96 08:01:53
 Output diagnostics / text file:
    U1:[APGILKE.0CMS.STEPWISE.TEST]STP_TEST.OUT
 Output plot text file:
```

```
STEPWISE_PA96 2.20 PROD PA96 05/30/96              06/07/96 14:48:34   Page  2

>!=================================================================
>!
>! STEPWISE input command file for Test Case #1
>!
>!=================================================================
>!
>TITLE "Stepwise Test Case Output: RANK "
>!
>STEPWISE .02, .05
>RANK
>PRESS
>DEP_VARS E1NP237, E1TOTAL
>IND_VARS CULTORT  CULMD0   CULFRCSP CULBCRCH CULPREC
>IND_VARS BHLGAREA BHLGPRES TIMINTR1
>IND_VARS RTRDMPU RTRDMAM  RTRDMNP  RTRDMU   RTRDFPU  RTRDFAM  RTRDFNP  RTRDFU
>END

Stepwise Test Case Output: RANK

Number of input Independent Variables   29
Number of input Dependent Variables   48

Number of observations on data files =   40
Number of observations to DROP =    0
Number of observations to process =   40

Do regression analysis on RANKS of data
Use STEPWISE procedure for "best" subset selection
   Significance level: 20.00E-3 and 50.00E-3
Use Predicted Error Sum of Squares (PRESS) to protect against overfit

WRITE the following information:
 o Analysis of variance table and regression coefficient estimates
   for final model of each dependent variable
NO analysis plots are requested

INDEPENDENT VARIABLES       DEPENDENT VARIABLES
selected for analysis      selected for analysis
---------------------      ---------------------
    10  CULTORT                 31  E1NP237
    11  CULMD0                  45  E1TOTAL
    12  CULFRCSP
    13  CULBCRCH
    14  CULPREC
    15  BHLGAREA
     6  BHLGPRES
     5  TIMINTR1
    16  RTRDMPU
    17  RTRDMAM
    18  RTRDMNP
    19  RTRDMU
    20  RTRDFPU
    21  RTRDFAM
    22  RTRDFNP
    23  RTRDFU
```

```
STEPWISE_PA96 2.20 PROD PA96 05/30/96             06/07/96 14:48:34   Page  3

                        Stepwise Test Case Output: RANK

              REGRESSION ANALYSIS for Dependent Variable -E1NP237-

PRESS =  2.9573E+03

PRESS =  2.4876E+03

PRESS =  2.0222E+03

               ANALYSIS OF VARIANCE TABLE for Variable -E1NP237-

SOURCE       DofF      SS            MS             F        SIGNIF
----------   ----   ----------    ----------    ----------    ------
REGRESSION     3    3.5688E+03    1.1896E+03    2.5173E+01    0.0000
RESIDUAL      36    1.7012E+03    4.7256E+01
TOTAL         39    5.2700E+03

R-SQUARE =  0.67719      INTERCEPT =  3.6309E+01

VARIABLE    R-SQUARE    REGRESSION  STANDARDIZED      PARTIAL       T-TEST       R-SQUARE
           WHEN INCL    COEFFICIENT   REGR COEFF        SSQ          VALUES        DELETES
--------    --------    ----------    ----------    ----------    ----------    ----------
RTRDMNP      0.48404   -7.3647E-01   -7.3486E-01    2.8193E+03   -7.7240E+00    1.4221E-01
CULTORT      0.58520   -3.3703E-01   -3.3894E-01    5.9797E+02   -3.5572E+00    5.6372E-01
CULFRCSP     0.67719    3.0232E-01    3.0404E-01    4.8478E+02    3.2029E+00    5.8520E-01
   VARIABLE      ALPHA
                  HATS
   --------    ----------
   RTRDMNP     3.7746E-09
   CULTORT     1.0726E-03
   CULFRCSP    2.8450E-03
```

Note that the ALPHA HATS are the last column in the table above in the output file, but they are shown here separately to aid in readability.

```
STEPWISE_PA96 2.20 PROD PA96 05/30/96             06/07/96 14:48:34   Page  4

                        Stepwise Test Case Output: RANK

              REGRESSION ANALYSIS for Dependent Variable -E1TOTAL-

PRESS =  2.8768E+03

PRESS =  1.1720E+03

               ANALYSIS OF VARIANCE TABLE for Variable -E1TOTAL-

SOURCE       DofF      SS            MS             F        SIGNIF
----------   ----   ----------    ----------    ----------    ------
REGRESSION     2    4.3908E+03    2.1954E+03    8.6492E+01    0.0000
RESIDUAL      37    9.3916E+02    2.5383E+01
TOTAL         39    5.3300E+03

R-SQUARE =  0.82380      INTERCEPT =  1.8060E+01

VARIABLE    R-SQUARE    REGRESSION  STANDARDIZED      PARTIAL       T-TEST       R-SQUARE
           WHEN INCL    COEFFICIENT   REGR COEFF        SSQ          VALUES        DELETES
--------    --------    ----------    ----------    ----------    ----------    ----------
BHLGAREA     0.50523    6.8493E-01    6.6946E-01    2.3761E+03    9.6752E+00    3.7801E-01
TIMINTR1     0.82380   -5.6593E-01   -5.6593E-01    1.6979E+03   -8.1788E+00    5.0523E-01
   VARIABLE      ALPHA
                  HATS
   --------    ----------
   BHLGAREA     2.7839E-08
   TIMINTR1     2.8639E-08
```

Note that the ALPHA HATS are the last column in the table above in the output file, but they are shown here separately to aid in readability.

```
STEPWISE_PA96 2.20 PROD PA96 05/30/96              06/07/96 14:48:34   Page  5

>!
>TITLE "Stepwise Test Case Output: RAW"
>RANK OFF
>END

Stepwise Test Case Output: RAW

Number of input Independent Variables   29
Number of input Dependent Variables   48

Number of observations on data files =   40
Number of observations to DROP =    0
Number of observations to process =   40

Do regression analysis on RAW data
Use STEPWISE procedure for "best" subset selection
   Significance level: 20.00E-3 and 50.00E-3
Use Predicted Error Sum of Squares (PRESS) to protect against overfit

WRITE the following information:
 o Analysis of variance table and regression coefficient estimates
   for final model of each dependent variable
NO analysis plots are requested

INDEPENDENT VARIABLES       DEPENDENT VARIABLES
selected for analysis      selected for analysis
---------------------      ---------------------
    10  CULTORT                 31  E1NP237
    11  CULMD0                  45  E1TOTAL
    12  CULFRCSP
    13  CULBCRCH
    14  CULPREC
    15  BHLGAREA
     6  BHLGPRES
     5  TIMINTR1
    16  RTRDMPU
    17  RTRDMAM
    18  RTRDMNP
    19  RTRDMU
    20  RTRDFPU
    21  RTRDFAM
    22  RTRDFNP
    23  RTRDFU
```

```
STEPWISE_PA96 2.20 PROD PA96 05/30/96              06/07/96 14:48:34   Page  6

                         Stepwise Test Case Output: RAW

              REGRESSION ANALYSIS for Dependent Variable -E1NP237-

%%% ERROR - No selected independent variables qualify for entry into the model



STEPWISE_PA96 2.20 PROD PA96 05/30/96              06/07/96 14:48:34   Page  7

                         Stepwise Test Case Output: RAW

              REGRESSION ANALYSIS for Dependent Variable -E1TOTAL-

PRESS =  3.6830E-05

PRESS =  2.0941E-05

PRESS =  1.7622E-05

               ANALYSIS OF VARIANCE TABLE for Variable -E1TOTAL-

SOURCE       DofF       SS           MS            F       SIGNIF
----------   ----   ----------   ----------   ----------   ------
REGRESSION     3    4.2610E-05   1.4203E-05   3.7879E+01   0.0000
RESIDUAL      36    1.3499E-05   3.7497E-07
TOTAL         39    5.6109E-05

R-SQUARE =  0.75942      INTERCEPT =  1.7243E-03

VARIABLE    R-SQUARE    REGRESSION  STANDARDIZED      PARTIAL       T-TEST       R-SQUARE
           WHEN INCL   COEFFICIENT    REGR COEFF        SSQ         VALUES        DELETES
--------    --------   ----------   ----------    ----------    ----------    ----------
BHLGAREA     0.44401    2.7422E-02   7.5035E-01    3.0641E-05    9.0397E+00    2.1333E-01
TIMINTR1     0.69733   -8.5752E-15  -5.4098E-01    1.5711E-05   -6.4729E+00    4.7942E-01
RTRDMAM      0.75942   -2.1560E-07  -2.5110E-01    3.4838E-06   -3.0481E+00    6.9733E-01
   VARIABLE      ALPHA
                  HATS
   --------    ----------
   BHLGAREA     8.5802E-11
   TIMINTR1     1.6293E-07
   RTRDMAM      4.2982E-03
```

Note that the ALPHA HATS are the last column in the table above in the output file, but they are shown here separately to aid in readability.

```
STEPWISE_PA96 2.20 PROD PA96 05/30/96              06/07/96 14:48:34   Page  8

>!
>!==================================================================

Normal termination on end of input file

STEPWISE_PA96 CPU time is 0:01 (minute:second)



*** END OF STEPWISE_PA96 ***
STEPWISE_PA96 2.20 PROD PA96 05/30/96                        06/07/96 14:48:34
```

APPENDIX D: QA REVIEW FORMS

As a convenience to users, this section contains a complete history of all the Quality-Assurance Review Forms issued during the review of this User's Manual. Review forms are issued after the User's Manual has been completed and reviewed. Thus, they will be appended to the manual after it has been paginated and therefore they may not, themselves, be paginated.