Chapter 1 Regression Concepts

1.1 Statistical Background

Multiple linear regression is a means to express the idea that a response variable, *y*, varies with a set of independent variables $x_1, x_2, \ldots, x_m$. The variability that *y* exhibits has two components: a systematic part and a random part. The systematic variation of *y* can be modeled as a function of the *x* variables. The model relating *y* to $x_1, x_2, \ldots, x_m$ is called the *regression equation*. The random part takes into account the fact that the model does not exactly describe the behavior of the response.

Formally, multiple linear regression fits a response variable *y* to a function of regressor variables and parameters. The general linear regression model has the form

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_m x_m + \varepsilon$$

where

y is the response, or dependent, variable

$\beta_0, \beta_1, \ldots, \beta_m$ are unknown parameters

$x_1, x_2, \ldots, x_m$ are the regressor, or independent, variables

ε is a random error term.

Least squares is a technique used to estimate the parameters based on a set of observed values of these variables. The goal is to find estimates of the parameters $\beta_0, \beta_1, \ldots, \beta_m$ that minimize the sum of the squared differences between the actual *y* values and the values of *y* predicted by the equation. These estimates are called the *least-squares estimates,* and the quantity minimized is called the *error sum of squares*.

Typically, you use regression analysis to do the following:

- ❑ obtain the least-squares estimates of the parameters
- ❑ estimate the variance of the error term
- ❑ estimate the standard error of the parameter estimates
- ❑ test hypotheses about the parameters
- ❑ calculate predicted values using the estimated equation
- ❑ evaluate the fit or lack of fit of the model.

The classical linear model assumes that the responses, y, are sampled from several populations. These populations are determined by the corresponding values of $x_1, x_2, \ldots, x_m$. As the investigator, you select the values of the x's; they are not random. However, the response values are random. You select the values of the x's to meet your experimental needs, carry out the experiment with the set values of the x's, and measure the responses. Often, though, you cannot control the actual values of the independent variables. In these cases, you should at least be able to assume that they are fixed with respect to the response variable.

In addition, you must assume that

1. the form of the model is correct; that is, all important independent variables are included and the functional form is appropriate

2. the expected values of the errors are zero

3. the variances of the errors (and thus the response variable) are constant across observations

4. the errors are uncorrelated

5. for hypothesis testing, the errors are normally distributed.

Not all regression models are necessarily linear in the parameters. For example, the model

$$y = \beta_1 e^{\beta_2 x} + \varepsilon$$

is not linear in the parameter β_2. Specifically, the term $e^{\beta_2 x}$ is not a linear function of β_2. This particular nonlinear model, called the exponential growth or decay model, is used to represent increase (growth) or decrease (decay) over time (t) of many types of responses such as population size or radiation counts. Chapter 7, "Nonlinear Models," is devoted to analyses appropriate for this type of model.

Additionally, the random error may not be normally distributed. If this is the case, the least squares technique is not necessarily the appropriate method for estimating the parameters. One such model, the logistic regression model, is presented in Section 7.5.

1.1.1 Terminology and Notation

The principle of least squares is applied to a set of n observed values of y and the associated x_j to obtain estimates $\hat{\beta}_{0,} \hat{\beta}_{1,\ldots,} \hat{\beta}_m$ of the respective parameters $\beta_0, \beta_1, \ldots, \beta_m$. These estimates are then used to construct the fitted model, or estimating equation,

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \ldots + \hat{\beta}_m x_m$$

Many regression computations are illustrated conveniently in matrix notation. Let y_i, x_{ij}, and ε_i denote the values of y, x_j, and ε, respectively, in the ith observation. The **Y** vector, the **X** matrix, and the $\boldsymbol{\varepsilon}$ vector can be defined as follows:

$$\mathbf{Y} = \begin{bmatrix} y_1 \\ \cdot \\ \cdot \\ \cdot \\ y_n \end{bmatrix}, \mathbf{X} = \begin{bmatrix} 1 & x_{11} & \cdot & \cdot & \cdot & x_{1m} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & & & & \cdot \\ \cdot & \cdot & & & & \cdot \\ 1 & x_{n1} & \cdot & \cdot & \cdot & x_{nm} \end{bmatrix}, \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \bullet \\ \bullet \\ \bullet \\ \varepsilon_n \end{bmatrix}$$

Then the model in matrix notation is

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}.$$

where $\boldsymbol{\beta}' = (\beta_0 \, \beta_1 \ldots \beta_m)$ is the vector of parameters.

The vector of least-squares estimates is $\hat{\boldsymbol{\beta}}' = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_m)$ and is obtained by solving the set of normal equations (NE)

$$\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \ \mathbf{X}'\mathbf{Y}\,.$$

Assuming that **X′X** is of full rank (nonsingular), the unique solution to the normal equations is given by

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}\ .$$

The matrix = $(\mathbf{X}'\mathbf{X})^{-1}$ is very useful in regression analysis and is often denoted as follows:

$$(\mathbf{X}'\mathbf{X})^{-1} = \mathbf{C} = \begin{bmatrix} c_{00} & c_{01} & \cdot & \cdot & \cdot & c_{0m} \\ c_{10} & c_{11} & \cdot & \cdot & \cdot & c_{1m} \\ ..\cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ c_{m1} & c_{m2} & \cdot & \cdot & \cdot & c_{mm} \end{bmatrix}$$

1.1.2 Partitioning the Sums of Squares

A basic identity results from least squares, specifically,

$$\Sigma(y - \bar{y})^2 = \Sigma(\bar{y} - \hat{y})^2 + \Sigma(y - \hat{y})^2.$$

This identity shows that the total sum of squared deviations from the mean, $\Sigma(\mathrm{y} - \bar{\mathrm{y}})^2$, is equal to the sum of squared differences between the mean and the predicted values, $\Sigma(\bar{y} - \hat{y})^2$, plus the sum of squared deviations from the observed *y*'s to the predicted values, $\Sigma(y - \hat{y})^2$. These two parts are called the sum of squares due to regression (or model) and the residual (or error) sum of squares. Thus,

Corrected Total SS = Model SS + Residual SS.

Corrected Total SS always has the same value for a given set of data, regardless of the model that is used; however, partitioning into Model SS and Residual SS depends on the model. Generally, the addition of a new *x* variable to a model increases the Model SS and, correspondingly, reduces the Residual SS. The residual, or error, sum of squares is computed as follows:

$$\begin{aligned}\text{Residual}\quad \text{SS} &= \mathbf{Y}'\left(\mathbf{I} - \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\right)\mathbf{Y} \\ &= \mathbf{Y}'\mathbf{Y} - \mathbf{Y}'\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{Y} \\ &= \mathbf{Y}'\mathbf{Y} - \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{Y} \quad .\end{aligned}$$

The error, or residual, mean square

$$s^2 = \text{MSE} = (\text{Residual SS}) \,/\, (n - m - 1)$$

is an unbiased estimate of σ^2, the variance of the ε's . This is the so-called error variance generally used in hypothesis testing.

Sums of squares, including the different sums of squares computed by any regression procedure such as the REG and GLM procedures, can be expressed conceptually as the difference between the regression sums of squares for two models, called complete (unrestricted) and restricted models, respectively. This approach relates a given SS to the comparison of two regression models.

For example, denote as SS_1 the regression sum of squares for a complete model with m=5 variables:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \varepsilon \quad .$$

Denote as SS_2 the regression sum of squares for a restricted model not containing x_4 and x_5:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon \quad .$$

Reduction notation can be used to represent the difference between regression sums of squares for the two models:

$$\text{R}\left(\beta_4, \beta_5 \mid \beta_0, \beta_1, \beta_2, \beta_3\right) = \text{Model SS}_1 - \text{Model SS}_2 \quad .$$

The difference or reduction in error $\text{R}\left(\beta_4, \beta_5 \mid \beta_0, \beta_1, \beta_2, \beta_3\right)$ indicates the increase in regression sums of squares due to the addition of β_4 and β_5 to the restricted model. It follows that

$$\text{R}\left(\beta_4, \beta_5 \mid \beta_0, \beta_1, \beta_2, \beta_3\right) = \text{Residual SS}_2 - \text{Residual SS}_1$$

that is, the decrease in error sum of squares due to the addition of β_4 and β_5 to the restricted model. The expression

$$\text{R}\left(\beta_4, \beta_5 \mid \beta_0, \beta_1, \beta_2, \beta_3\right)$$

is also commonly referred to in the following ways:

- ❑ the sums of squares due to β_4 and β_5 (or x_4 and x_5) adjusted for $\beta_0, \beta_1, \beta_2, \beta_3$ (or the intercept and x_1, x_2, x_3)
- ❑ the sums of squares due to fitting x_4 and x_5 after fitting the intercept and x_1, x_2, x_3
- ❑ the effects of x_4 and x_5 above and beyond or partialing the effects of the intercept and x_1, x_2, x_3.

1.1.3 Hypothesis Testing

Inferences about model parameters are highly dependent on the other parameters in the model under consideration. Therefore, in hypothesis testing, it is important to emphasize the parameters for which inferences have been adjusted. For example, $\mathrm{R}(\beta_3 \mid \beta_0, \beta_1, \beta_2)$ and $\mathrm{R}(\beta_3 \mid \beta_0, \beta_1)$ may measure entirely different concepts. In other words, a test of $\mathrm{H}_0 : \beta_3 = 0$ may have one result for the model

$$y = \beta_0 + \beta_1 x_1 + \beta_3 x_3 + \varepsilon$$

and another for the model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon \quad .$$

Differences reflect actual dependencies among variables in the model rather than inconsistencies in statistical methodology.

Statistical inferences can also be made in terms of linear functions of the parameters of the form

$$\mathrm{H}_0 : \mathbf{L}\boldsymbol{\beta}_0 : \ \ell_0\beta_0 + \ell_1\beta_1 + \ldots + \ell_m\beta_m = 0$$

where the ℓ_i are arbitrary constants chosen to correspond to a specified hypothesis. Such functions are estimated by the corresponding linear function

$$\mathbf{L}\hat{\boldsymbol{\beta}} = \ell_0\hat{\beta}_0 + \ell_1\hat{\beta}_1 + \ldots + \ell_m\hat{\beta}_m$$

of the least-squares estimates $\hat{\boldsymbol{\beta}}$. The variance of $\mathbf{L}\hat{\boldsymbol{\beta}}$ is

$$\mathrm{V}\left(\mathbf{L}\hat{\boldsymbol{\beta}}\right) = \left(\mathbf{L}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{L}'\right)\sigma^2 \quad .$$

A *t* test or *F* test is used to test $\mathrm{H}_0 : \left(\mathbf{L}\boldsymbol{\beta}\right) = 0$. The denominator is usually the residual mean square (MSE). Because the variance of the estimated function is based on statistics computed for the entire model, the test of the hypothesis is made in the presence of all model parameters. Confidence intervals can be constructed to correspond to these tests, which can be generalized to simultaneous tests of several linear functions.

Simultaneous inference about a set of linear functions $\mathbf{L}_1\boldsymbol{\beta}, \ldots, \mathbf{L}_k\boldsymbol{\beta}$ is performed in a related manner. For notational convenience, let **L** denote the matrix whose rows are $\mathbf{L}_1, \ldots, \mathbf{L}_k$:

$$\mathbf{L} = \begin{bmatrix} \mathbf{L}_1 \\ . \\ . \\ . \\ \mathbf{L}_k \end{bmatrix}$$

Then the sum of squares

$$\mathrm{SS}(\mathbf{L}\boldsymbol{\beta} = 0) = \left(\mathbf{L}\hat{\boldsymbol{\beta}}\right)' \left(\mathbf{L}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{L}'\right)^{-1} \left(\mathbf{L}\hat{\boldsymbol{\beta}}\right)$$

is associated with the null hypothesis

$$\mathrm{H}_0 : \mathbf{L}_1\boldsymbol{\beta} = \ldots = \mathbf{L}_k\boldsymbol{\beta} = 0 \quad .$$

A test of H_0 is provided by the *F* statistic

$$F = \left(\mathrm{SS}(\mathbf{L}\boldsymbol{\beta} = 0) / k\right) / \mathrm{MSE} \quad .$$

Three common types of statistical inferences are

❑ a test that all parameters $(\beta_1, \beta_2, \ldots, \beta_m)$ are zero. The test compares the fit of the complete model to that using only the mean:

F = (Model SS / m) / MSE

where

$$\text{Model SS} = \mathrm{R}(\beta_1, \beta_2, \ldots, \beta_m \mid \beta_0)$$

The *F* statistic has (m, n–m–1) degrees of freedom.[1]

❑ a test that the parameters in a subset are zero. The problem is to compare the fit of the complete model

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_g x_g + \beta_{g+1} x_{g+1} + \ldots + \beta_m x_m + \varepsilon$$

to the fit of the restricted model

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_g x_g + \varepsilon \quad .$$

[1] $\mathrm{R}(\beta_0, \beta_1, \ldots, \beta_m)$ is rarely used. For more information, see the NOINT option in Section 2.4.5.

An F statistic is used to perform the test

$$F = (\mathrm{R}(\beta_{g+1},\ldots,\beta_m \mid \beta_0,\beta_1,\ldots,\beta_g)\,/\,(m-g))\ \mathrm{MSE} \quad .$$

Note that an arbitrary reordering of variables produces a test for any desired subset of parameters. If the subset contains only one parameter, β_m, the test is

$$\begin{aligned} F &= \left(\mathrm{R}(\beta_m \mid \beta_0, \beta_1, \ldots, \beta_{m-1}) / 1\right) / \ \mathrm{MSE} \\ &= \left(\text{partial SS due to } \beta_m\right) / \ \mathrm{MSE} \end{aligned}$$

which is equivalent to the t test

$$t = \hat{\beta}_m \ / \ s_{\hat{\beta}m} = \hat{\beta}_m \ / \ \sqrt{c_{mm}\mathrm{MSE}} \quad .$$

The corresponding $(1-\alpha)$ confidence interval about β_m is

$$\hat{\beta}_m \pm t_{\alpha/2}\sqrt{c_{mm}\mathrm{MSE}} \quad .$$

❑ estimation of a subpopulation mean corresponding to a specific *x*. For a given set of *x* values described by a vector **x**, denote the population mean by $\mu_{\mathbf{x}}$. The point estimate of that population mean is

$$\hat{\mu}_{\mathbf{x}} = \hat{\beta}_0 + \hat{\beta}_1\mathbf{x}_1 + \ldots + \hat{\beta}_m\mathbf{x}_m = \mathbf{x}\hat{\boldsymbol{\beta}} \ .$$

The vector **x** is constant; hence, the variance of the estimate, $\hat{\mu}_{\mathbf{x}}$, is

$$\mathrm{V}(\hat{\mu}_{\mathbf{x}}) = \mathbf{x}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}\sigma^2 \ .$$

This equation is useful for computing confidence intervals. A related inference concerns a future single value of *y* corresponding to a specified *x* whose estimate is denoted by *x*. The point estimate is the same as that for the mean, but its variance is

$$\mathrm{V}(\hat{y}_{\mathbf{x}}) = (1 + \mathbf{x}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x})\sigma^2 \ .$$

1.1.4 Using the Generalized Inverse

Many applications of regression procedures involve an **X′X** matrix that is not of full rank and has no unique inverse. PROC GLM and PROC REG compute a generalized inverse $(\mathbf{X}'\mathbf{X})^-$ and use it to compute a regression estimate

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^-\,\mathbf{X}'\mathbf{Y} \quad .$$

A generalized inverse of a matrix **A** is any matrix **G** such that **AGA=A**. Note that this also identifies the inverse of a full-rank matrix.

If $\mathbf{X'X}$ is not of full rank, then an infinite number of generalized inverses exist. Different generalized inverses lead to different solutions to the normal equations that have different expected values; that is, $\mathrm{E}(\mathbf{b}) = (\mathbf{X'X})^{-}\mathbf{X'Y\beta}$ depends on the particular generalized inverse used to obtain **b**. Therefore, it is important to understand what is being estimated by the solution.

Fortunately, not all computations in regression analysis depend on the particular solution obtained. For example, the error sum of squares is invariant with respect to $(\mathbf{X'X})^{-}$ and is given by

$$\mathrm{SSE} = \mathbf{Y}'\left(\mathbf{1} - \mathbf{X}(\mathbf{X'X})^{-}\mathbf{X}'\right)\mathbf{Y} \quad .$$

Hence, the model sum of squares also does not depend on the particular generalized inverse obtained.

The generalized inverse has played a major role in the presentation of the theory of linear statistical models, notably in the work of Graybill (1976) and Searle (1971). In a theoretical setting it is often possible, and even desirable, to avoid specifying a particular generalized inverse. To apply the generalized inverse to statistical data using computer programs, a generalized inverse must actually be calculated. Therefore, it is necessary to declare the specific generalized inverse being computed. For example, consider an $\mathbf{X'X}$ matrix of rank k that can be partitioned as

$$\mathbf{X'X} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix}$$

where $\mathbf{A}_{11}$ is k x k and of rank k. Then $\mathbf{A}_{11}^{-1}$ exists, and a generalized inverse of $\mathbf{X'X}$ is

$$(\mathbf{X'X})^{-} = \begin{bmatrix} \mathbf{A}_{11}^{-1} & \varphi_{12} \\ \varphi_{21} & \varphi_{22} \end{bmatrix}$$

where each φ_{ij} is a matrix of zeros of the appropriate dimension.

This approach to obtaining a generalized inverse, the method used by PROC GLM and PROC REG, can be extended indefinitely by partitioning a singular matrix into several sets of matrices as illustrated above. Note that the resulting solution to the normal equations, $\mathbf{b}=(\mathbf{X'X})^{-}\mathbf{X'Y}$, has zeros in the positions corresponding to the rows filled with zeros in $(\mathbf{X'X})^{-}$. This is the solution printed by these procedures, and it is regarded as providing a biased estimate of β .

However, because **b** is not unique, a linear function, **Lb**, and its variance are generally not unique either. However, a class of linear functions called *estimable functions* exists, and they have the following properties:

❑ The vector **L** is a linear combination of rows of **X**.

❑ **Lb** and its variance are invariant through all possible generalized inverses. In other words, **Lb** is unique and is an unbiased estimate of $\mathbf{L}\beta$.

Analogous to the full-rank case, the variance of an estimable function **Lb** is given by

$$\mathrm{V}(\mathbf{Lb}) = (\mathbf{L}\,(\mathbf{X'X})^{-}\,\mathbf{L}')\sigma^{2}$$

This expression is used for statistical inference. For example, a test of $\mathrm{H}_0 : \mathbf{L\beta} = 0$ is given by the *t* test

$$t = \mathbf{Lb} / \sqrt{\left(\mathbf{L}(\mathbf{X}'\mathbf{X})^{-}\ \mathbf{L}'\right)\mathrm{MSE}}\ \ .$$

Simultaneous inferences on a set of estimable functions are performed in an analogous manner.

1.2 Performing a Regression with the IML Procedure

As you can see in Section 1.3, “Regression with the SAS System,” and in greater detail in subsequent chapters, the SAS System provides a flexible array of procedures for performing regression analyses. You can also perform these analyses by direct application of the matrix formulas presented in the previous section using SAS/IML software. This software, which is implemented as PROC IML, is most frequently used for the custom programming of methods too specialized or too new to be packaged into the standard regression procedures. It is also useful as an instructional tool for illustrating linear model and other methodologies.

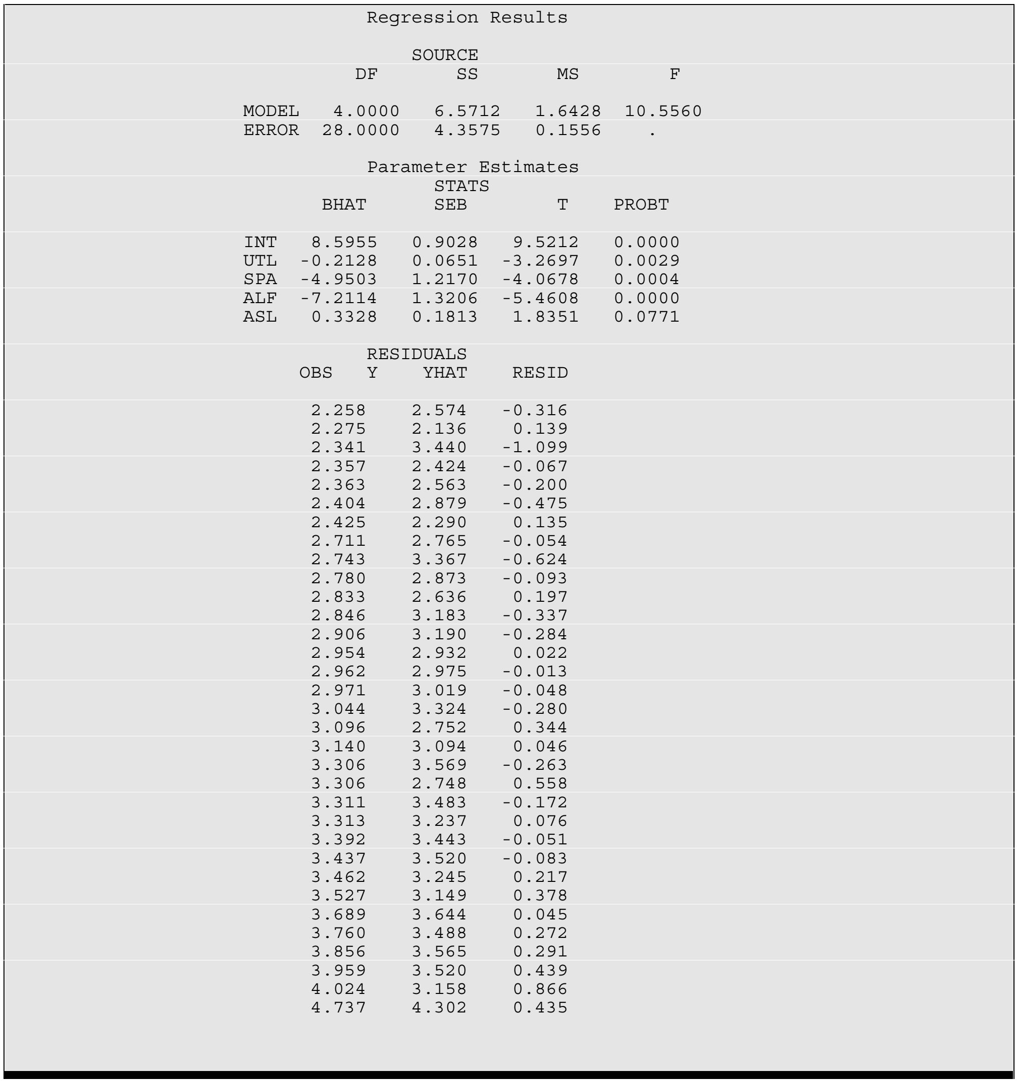
The following example represents a regression analysis performed by PROC IML. This example is not intended to serve as a tutorial in the use of PROC IML. If you need more information on PROC IML, see the *SAS/IML User’s Guide*. The example for this section is the one used in Chapter 2, “Using the REG Procedure,” to illustrate PROC REG. The data set is described, and the data are presented in Section 2.1, “Introduction.” For this presentation, the variable CPM is the dependent variable *y*, and the variables UTL, SPA, ALF, and ASL are the independent variables x_1, x_2, x_3, and x_4, respectively. Comment statements are used in the SAS program to explain the individual steps in the analysis.

```
/* Invoke PROC IML and create the x and y matrices using    */
/* the variables UTL, SPA, ALF, and CPM from the SAS data   */
/* set AIR.                                                 */

proc iml;
   use air;
   read all var {‘utl’ ‘spa’ ‘alf’ ‘asl’} into x;
   read all var {‘cpm’} into y;
/* Define the number of observations (N) and the number of  */
/* variables (M) as the number of rows and columns of X.    */
/* Add a column of ones for the intercept variable to the X */
/* matrix.                                                  */

   n=nrow(x);     /* number of observations   */
   m=ncol(x);     /* number of variables      */
   x=j(n,1,1)||x; /* add column of ones to X */

/* Compute C, the inverse of X’X and the vector of     */
/* coefficient estimates BHAT.                        */

   c=inv(x’*x);
   bhat=c*x’*y;
```

```
/* Compute SSE, the residual sum of squares, and MSE, the   */
/* residual mean square (variance estimate).                */

   sse= y’*y-bhat’*x’*y;
   dfe= n-m-1;
   mse=sse/dfe;

/* The test for the model can be restated as a test for     */
/* the linear function L where L is the matrix.             */

   l={0 1 0 0 0,
      0 0 1 0 0,
      0 0 0 1 0,
      0 0 0 0 1};

/* Compute SSMODEL and MSMODEL and the corresponding F      */
/* ratio.                                                   */

   ssmodel=(l*bhat)’*inv(l*c*l’)*(l*bhat);
   msmodel=ssmodel/m;
   f=(ssmodel/m)/mse;

/* Concatenate results into one matrix.                     */
   source=(m||ssmodel||msmodel||f)//(dfe||sse||mse||{.});
/* Compute                                                  */
/* SEB   vector of standard errors of the estimated         */
/*       coefficients                                       */
/* T     matrix containing the t statistic for testing that */
/*       each coefficient is zero                           */
/* PROBT significance level of test                         */
/* STATS matrix which contains as its columns the           */

/*       coefficient estimates, their standard errors,      */
/*       and the t statistics.                              */

   seb=sqrt(vecdiag(c)#mse);
   t=bhat/seb;
   probt=2*(1 - cdf(‘t’,abs(t),dfe));
   stats=bhat||seb||t||probt;

/* Compute                                                  */
/* YHAT  predicted values                                   */
/* RESID residual values                                    */
/* OBS   matrix containing as its columns the actual,       */
/*       predicted, and residual values, respectively.      */

   yhat=x*bhat;
   resid=y-yhat;
   obs=y||yhat||resid;

/* Print the matrices containing the desired results.       */
   print ‘Regression Results’,
   source (|colname={DF SS MS F} rowname={MODEL ERROR}
   format=8.4|),,
   ‘Parameter Estimates’,
   stats (|colname={BHAT SEB T PROBT} rowname={INT UTL SPA ALF
   ASL}
   format=8.4|) ,,,
   ‘RESIDUALS’, obs (| colname={Y YHAT RESID} format=8.3|) ;
```

The results of this sample program are shown in Output 1.1.

Output 1.1
Output Produced by PROC IML



```
                        Regression Results

                             SOURCE
                    DF        SS        MS         F

          MODEL   4.0000   6.5712   1.6428  10.5560
          ERROR  28.0000   4.3575   0.1556     .

                       Parameter Estimates
                              STATS
                  BHAT       SEB         T    PROBT

          INT   8.5955   0.9028   9.5212   0.0000
          UTL  -0.2128   0.0651  -3.2697   0.0029
          SPA  -4.9503   1.2170  -4.0678   0.0004
          ALF  -7.2114   1.3206  -5.4608   0.0000
          ASL   0.3328   0.1813   1.8351   0.0771

                       RESIDUALS
               OBS   Y    YHAT    RESID

                2.258    2.574   -0.316
                2.275    2.136    0.139
                2.341    3.440   -1.099
                2.357    2.424   -0.067
                2.363    2.563   -0.200
                2.404    2.879   -0.475
                2.425    2.290    0.135
                2.711    2.765   -0.054
                2.743    3.367   -0.624
                2.780    2.873   -0.093
                2.833    2.636    0.197
                2.846    3.183   -0.337
                2.906    3.190   -0.284
                2.954    2.932    0.022
                2.962    2.975   -0.013
                2.971    3.019   -0.048
                3.044    3.324   -0.280
                3.096    2.752    0.344
                3.140    3.094    0.046
                3.306    3.569   -0.263
                3.306    2.748    0.558
                3.311    3.483   -0.172
                3.313    3.237    0.076
                3.392    3.443   -0.051
                3.437    3.520   -0.083
                3.462    3.245    0.217
                3.527    3.149    0.378
                3.689    3.644    0.045
                3.760    3.488    0.272
                3.856    3.565    0.291
                3.959    3.520    0.439
                4.024    3.158    0.866
                4.737    4.302    0.435
```

When you use PROC IML, all results are in the form of matrices. Each matrix is identified by its name, and its elements are identified by row and column indices. You may find it necessary to refer to the program to identify specific elements.

The results of this analysis are discussed thoroughly in Chapter 2; therefore, in this section only the results that can be compared with those from PROC REG (shown in Output 2.5) are identified.

In Output 1.1, the first matrix corresponds to overall model statistics produced by PROC REG. Included here are the degrees of freedom, sums of squares, and mean square for the model and for the error. The *F* statistic tests the significance of the entire model, which includes the independent variables UTL, SPA, ALF, and ASL.

The matrix **STATS** contains the information on the parameter estimates. Rows correspond to parameters (intercept and independent variables UTL, SPA, ALF, ASL, respectively), and columns correspond to the different statistics. The first column contains the coefficient estimates (from matrix **BHAT**), the second contains the standard errors of the estimates (from matrix **SEB**), and the third contains the *t* statistics (from matrix **T**). The final column (from matrix **PROBT**) contains the probability associated with the *t* statistic.

The matrix **OBS** contains the information on observations. The rows correspond to the observations. The first column contains the original *y* values (matrix **Y**), the second contains the predicted values (from matrix **YHAT**), and the third contains the residuals (from matrix **RESID**).

The results achieved by using PROC IML agree with those from PROC REG, as shown in Output 2.5. Because PROC IML is most frequently used for the custom programming of new or specialized methods, the standard regression procedures are more efficient with respect to both programming time and computing time. For this reason, you should try to use these procedures whenever possible. In addition, the output produced with the standard regression procedures is designed to present analysis results more clearly than the printed matrices produced with PROC IML. See Section 1.3 for an overview of standard regression procedures.

1.3 Regression with the SAS System

This section reviews the following SAS/STAT software procedures that are used for regression analysis:

CALIS
CATMOD
GENMOD
GLM
LIFEREG
LOESS
LOGISTIC
NLIN
ORTHOREG
PLS
PROBIT
REG
RSREG
TPSPLINE
TRANSREG

PROC REG provides the most general analysis capabilities; the other procedures give more specialized analyses. This section also briefly mentions several procedures in SAS/ETS software.

Many SAS/STAT procedures, each with special features, perform regression analysis. The following procedures perform at least one type of regression analysis:

CALIS fits systems of linear structural equations with latent variables and path analysis.

CATMOD analyzes data that can be represented by a contingency table. PROC CATMOD fits linear models to functions of response frequencies and can be used for loglinear models and logistic regression.

GENMOD fits generalized linear models. PROC GENMOD is especially suited for responses with discrete outcomes, and it performs logistic regression and Poisson regression as well as fitting generalized estimating equations for repeated measures data.

GLM uses the method of least squares to fit general linear models. In addition to many other analyses, PROC GLM can perform simple, multiple, polynomial, and weighted regression, as well as analysis of variance and analysis of covariance. PROC GLM has many of the same input/output capabilities as PROC REG but does not provide as many diagnostic tools or allow interactive changes in the model or data.

LIFEREG fits parametric models to failure-time data that may be right-, left-, or interval-censored. These types of models are commonly used in survival analysis.

LOESS fits a response curve or plane to data without using a specified model.

LOGISTIC fits logistic regression models. PROC LOGISTIC can perform stepwise regressions as well as compute regression diagnostics.

NLIN fits nonlinear regression models. Several different iterative methods are available.

ORTHOREG performs regression using the Gentleman-Givens computational method. For ill-conditioned data, PROC ORTHOREG can produce more accurate parameter estimates than other procedures such as PROC GLM and PROC REG.

PLS performs partial least squares regression, principal components regression, and restricted rank regression, with cross validation for the number of components.

PROBIT performs probit regression as well as logistic regression and ordinal logistic regression. PROC PROBIT is useful when the dependent variable is either dichotomous or polychotomous and the independent variables are continuous.

REG performs linear regression with many diagnostic capabilities, selects models using one of nine selection methods, produces scatter plots of raw data and statistics, highlights scatter plots to identify particular observations, and allows interactive changes in both the regression model and the data used to fit the model.

PROC REG provides options for special estimates, outlier and specification error detection (row diagnostics), collinearity statistics (column diagnostics), and tests of linear functions of parameter estimates. It can perform restricted least-squares estimation and multivariate tests. It can also produce SAS data sets containing the parameter estimates and most of the statistics produced by the procedure.

RSREG builds quadratic response-surface regression models to determine the factor levels of optimum response, and it performs a ridge analysis to search for the region of optimum response.

TPSPLINE fits a response curve or plane to data without using a specified model.

TRANSREG obtains optimal linear and nonlinear transformations of variables using alternating least squares. PROC TRANSREG creates an output data set containing the transformed variables.

SAS/ETS software provides tools for economic analysis and modeling, time-series analysis, and forecasting. Since many of these tools are forms of regression, many procedures in this software also perform regression. These include the following:

AUTOREG implements regression models using time-series data where the errors are autocorrelated.

MODEL handles nonlinear simultaneous systems of equations, such as econometric models.

PDLREG performs regression analysis with polynomial distributed lags.

SYSLIN handles linear simultaneous systems of equations, such as econometric models.

TSCSREG handles regression models that use both time-series and cross-sectional data.

Finally, if a regression method cannot be performed by any of the SAS procedures above, SAS/IML software provides an interactive matrix language than can be used, as shown in Section 1.2.