CHAPTER 87
Regression and Correlation*

RAJA M. PARVEZ
Lucent Technologies

DONALD FUSARO
Lucent Technologies

*Parts of this chapter were originally presented by the late Douglas C. Crocker in his chapter in the second edition of this Handbook.

1. INTRODUCTION TO REGRESSION ANALYSIS

Regression analysis is:

A technique for measuring and explaining (reducing unexplained) variability in a system
An aid to understanding interrelationships in complex systems
A process for building a useful model of a system
A method for improving forecasting or prediction
A mechanism for focusing on important phenomena
A system for evaluating theories or beliefs
An aid in formulating new theory
A method for obtaining better control of variation
A technique for estimating equation parameters

Regression modeling involves practical problems, problems of judgment, and a good deal of art. This chapter is not intended to be a recipe book or a catalog of rules of thumb. It is intended to introduce the reader to some basic principles involved in statistical modeling while at once exposing the dangers. In this spirit, this chapter discusses many of the difficulties that may be encountered in attempting to model systems displaying statistical variation. It is intended to serve as a good blend of theoretical structure, philosophical outlook, and practical guidance.

1.1. General Linear Model

An equation of the form

$$Y_i = \sum_{j=0}^{P} b_j X_{ij} \tag{1}$$

is sometimes referred to as the general linear model. In this equation, Y is a variable whose behavior is of interest. It was once common to refer to Y as the dependent variable, taken from the mathematical concept of a function. In statistical modeling, most authors have come to call Y the response variable. This is the convention adopted here.

In Eq. (1), Y is a linear additive function of the X variables, which are P in number, $P \geq 1$. These X's were formerly often referred to as independent variables, again using the mathematical sense. They are now sometimes called regressors or explanatory variables but are more commonly called predictors (although prediction may not be the goal). The subscript j denotes which predictor. In this general form of Eq. (1), there is a dummy variable, $X_0 = 1$ (''dummy'' because it does not vary), which is not counted as a predictor but is included in the summation. Its coefficient, b_0 is the constant term or intercept. It is in units of Y. The other regression coefficients, b_j ($j = 1$ to P), are the slopes (multipliers of their respective predictors) and are expressed in units of Y/X_j. These b_j's are unknowns that are to be determined from the analysis. The values obtained are estimates of the ''true'' unknown coefficients, β_j. Geometrically, Eq. (1) represents a line, a plane, or a hyperplane in $P + 1$ dimensional space. This process is known as multiple linear regression (MLR) analysis. The subscript, i, denotes the series of observations in the sample going from 1 to n. Each observation provides a value for each of the variables—the predictors and the response—for each of the units or individuals in the sample. The unit of observation may be, for example, a day, a person, an automobile, a task, an event, or a batch.

The X's can, of course, represent quite complicated transformations of originally observed bits of information about each unit. Reciprocals, powers, and logs are examples, and so are ratios or products of two (or more) predictors. It is astounding to witness how often this linear additive equation form gives a very good representation of the underlying physics that relate the response to the predictors.* That the variability, that is, the behavior, of so many things in nature can be so well described (predicted) by this simple summation process is truly profound.

1.2. Utility and Dangers

Variation is the essence of statistical modeling. Variation is the problem. Information is contained in variation. In fact, without variation, there is no information. The activities of industrial engineers virtually always involve dealing with variation in multiple-variable systems. The goal may be to evaluate or explain previous events or to predict or control future events. Modeling a response variable in such systems is usually complex and difficult. Part of the difficulty arises in most cases because the data come from the existing system as it normally operates rather than being generated during a designed experiment. Such data might be called nonexperimental or clinical.

Some major difficulties found in dealing with nonexperimental data result from the interrelationships naturally present among the predictors. The unwanted intercorrelations are avoided in controlled experiments by keeping the predictors uncorrelated with (orthogonal to) each other. This difficulty in dealing with clinical data is shared with many other disciplines. In fact, the exception is the analyst who is able to operate with ''scientific'' laboratory technique.

The great power of MLR lies in its ability to relate simultaneously the many intercorrelated predictors to the response—to deal with nonexperimental data. Herein also lies the main source of danger. Successful modeling of nonexperimental data is a tricky business. But not all the dangers are associated with the natural intercorrelations of nonexperiments. The variety of ways in which the analyst can encounter trouble is nearly as great as the variety of problem situations. Perhaps no other technique suffers more misuse and abuse than regression analysis. Because of this, much criticism of the general technique is offered by those who apparently do not understand its power or proper use and who misrepresent it. The dangers can be avoided or treated if they are recognized and understood. Much of the balance of this chapter deals directly or indirectly with establishing appropriate safeguards.

1.3. Importance of Goals

Multiple linear regression should not be a process that follows a fixed, predetermined path or employs an established ritual for achieving a goal. That is because different goals require different analytic behavior. As illustrated by this chapter's opening list, regression goals are various. Before attempting to model a system, it is important to know what the model is supposed to do. What is the question the analysis is supposed to answer?

Because we are dealing with practice in industrial engineering, it is important first to make the distinction between science and decision making (see Healy 1978). The statistical requirements for establishing scientific truth are much more stringent than for decision making. The manager cannot wait for the discovery of ultimate truth but must decide today. Ordinarily, the industrial engineer operates in support of that process and will serve the manager best if the decision process is supported in a timely manner. This is not to suggest carelessness or disregard for theory. It is to suggest recognition of the basic fact that the manager will make the decision with or without the potential help. A responsibly derived, yet imperfect, model can be very much better than no model at all.

Very broadly, the various goals can be put into five categories. These represent a natural evolutionary sequence of four steps, any one of which may be the intended end use.

1. *Exploration:* fishing, hypothesis finding (see Finch 1979)
2. *Specification:* hypothesis testing, confirmation of the model form (rarely an end use)
3. *Estimation:* estimating model parameters with sufficient precision (estimating future events is referred to in this chapter as ''prediction'')
4. *Prediction:* use of the model for anticipation or for ''inverse estimation'' (calibration)
5. *Control:* use of the model to prescribe change or to direct or guide policy or the behavior of a system

*A known underlying causal relationship is not a requirement for useful statistical modeling.

TABLE 1 Simple Classification of Regression Models

Class	Kind	Basis
Associative	Concomitant, precursory	Observation
Physical	Empirical	
	Causal, mechanistic	Theory

1.4. Kinds of Models

Kinds of models seem to lie more along a continuum and are therefore less easily classified. The main continuum is closeness to causality. The scale slides from loose empiricism to exact causal representation (mechanism). How far along the scale the analyst moves may depend on either the maturity of the corresponding physical discipline or the needs imposed by the goals.

There is another subset where causality is not an issue. These models might be called ''associative.'' Here the response and the predictors may both be ''caused'' by some outside force. They behave concomitantly. An example is the use of leading indicators in economic models. Another example is the precursory use of animal characteristics or behavior to predict the severity of the winter. Presumably no one would claim that the extra hair on the woolly bear caterpillar causes snow to fall (causation might be suspected in the other direction in such cases if it were not for our belief that cause must precede in time its effect).

This simple classification scheme thus takes the form shown in Table 1.

1.5. Appropriate Use of Statistics

Statistical measures and diagnostics can and do serve an essential role in regression modeling, but they must be used appropriately. Their use must be related to goals. In general, any adequate MLR computer program system will list many statistics that may not be relevant in any given situation. For example, the multiple correlation coefficient, R, is universally printed. It may be of no interest. Further, even if it is of interest, its value must be judged in the context of the problem. It depends on the question. The analyst must know what questions need to be answered and must use relevant statistical measures accordingly.

1.6. Role of Assumptions

Assumptions are, in most regression articles and texts, listed as a sort of litany to precede the analysis as if they universally apply. Moreover, they are treated as if they describe the problem setting. They are really descriptions of the mathematical model whose behavioral properties are known and that is to be used as an analog to the system under study. Assumptions (model characteristics) relate to goals just as statistics do. Those that are relevant are rarely ever exactly met by the problem system. The severity of trouble the analyst may expect because of the remaining differences (''violations of assumptions'') is a matter for judgment and experience and cannot be removed from the problem context.

Table 2 offers a skeleton relationship of assumptions to goals in a hierarchical order (for a more complete discussion, see Eisenhart 1947). Random residual variation in Y is associated with a host of small, unimportant (in context) contributions. Notice that the usual assumptions of homoscedasticity and normality are not imposed for specification and estimation. The least-squares estimates of the regression coefficients provided by MLR are the most efficient, unbiased linear estimates among all linear estimates for uniform error variance and are still unbiased for nonuniform error variance. The central limit theorem will give very good protection—just as with ordinary averaging—allowing

TABLE 2 Cumulative Relationships of Assumptions to Goals

Goals	Desirable Data Characteristics	Model and Process Characteristics
Exploration	Random Y for given X	Least-squares fitting
Specification	''Complete'' X set	''Correct'' model form
Estimation	Spread, balanced X's	b's normal by central limit theorem
Prediction and Control	Typical X space	Specified error distribution

the normal model to be used with nonnormal data for establishing confidence intervals. Stated characteristics are cumulative descending the table.

2. RELATING TWO VARIABLES

The actual use of the simple (single-predictor) model is rare (real systems are rarely that simple.) However, for examining the principles involved in regression modeling, the simple model serves well.

For illustration, hypothetical data representing steam consumption (Y) for a particular building are modeled here. This response variable was chosen because (1) energy use has universal relevance and global importance, (2) such a wide variety of goals can be authentically represented in an energy system, and (3) this same problem setting can be expanded in following sections to represent more complex modeling ventures. The structure under study might be an office complex, a factory, a warehouse, a hospital, a hotel, or even a home. For demonstration, only 20 observations are contained in the sample. Each observation represents a four-week period. Weekly data would be preferred in most cases, but to cover extremes of weather in only 20 data points, four-week periods were chosen. The goal is to establish control of steam consumption for this building. ''Excessive'' use is now dismissed as being weather related.

At first it is assumed that comfort heating in this building is the major use of steam. Its use (measured in giga-British thermal units, gBtu), should be reasonably well related to degree-days (X). This is measured relative to 65°F (degree-days F/1.8 = degree-days C) and is also reported here on a per-period basis. The 20 observations are shown in Table 3, with the periods numbered within year from 1 to 13 and years numbered 1, 2, and 3. The method of least squares will be employed to relate steam use to degree-days.

Table 3 also shows the fitted values, called *YHAT,* and the residual fitting errors, called *Y-YHAT*. The names and their symbolic forms are discussed following Eq. (3). Diagnostic attention will be given later to the values in these two columns.

2.1. Least-Squares Method

In MLR, understanding variation is the basis for problem solving. Variation in the response is made up (theoretically) of two parts:

TABLE 3 Hypothetical Data for Modeling Steam Consumption

i	Period Number	X (k degree-days)[a]	Y (gBtu)[b]	Fitted *YHAT* (gBtu)	Residual *Y-YHAT* (gBtu)
1	10-1	0.156	7.991	6.990	1.001
2	11-1	0.419	8.589	8.095	0.494
3	12-1	0.658	9.145	9.100	0.045
4	13-1	1.009	11.212	10.575	0.637
5	1-2	1.380	11.754	12.134	−0.380
6	2-2	1.103	11.469	10.970	0.499
7	3-2	1.000	10.584	10.537	0.047
8	4-2	0.703	9.509	9.289	0.220
9	5-2	0.207	7.457	7.204	0.253
10	6-2	0.086	6.989	6.696	0.293
11	7-2	0.024	6.537	6.435	0.102
12	8-2	0.005	4.938	6.355	−1.417
13	9-2	0.026	5.275	6.444	−1.169
14	10-2	0.161	7.452	7.011	0.441
15	11-2	0.307	7.962	7.625	0.337
16	12-2	0.664	8.915	9.125	−0.210
17	13-2	1.039	9.758	10.701	−0.943
18	1-3	1.275	11.183	11.693	−0.510
19	2-3	1.193	11.523	11.348	0.175
20	3-3	0.953	10.426	10.340	0.086

$n = 20$; $\Sigma X_i = 12.369$; $\Sigma Y_i = 178.67$
$\Sigma X_i^2 = 11.915$; $\Sigma Y_i^2 = 1678.7$; $\Sigma X_i Y_i = 128.42$

[a] k degree-days = 10^3 degree-days.
[b] gBtu = giga-British thermal units = 10^9 Btu.

1. The systematic variation (signal), which is associated with or is in response to changes in the predictors
2. Leftover variation (noise), which is called "residual error" or "experimental error."

The distinction is really not so sharp. The leftover error is actually associated with a great many things that, in practice, might be measured (and included in the model) if analysts had sufficient time, wisdom, patience, and money. They simply choose not to try to identify all sources of variation. They will discontinue the search when there seems to be no regular pattern of errors left over and when either all the reasonable predictors have been adequately tested or the residual error variance is small enough—again depending on goals. In terms of the true coefficients and residual error of the theoretical model, the observed response variable may be expressed as

$$Y_i = \sum_{j=0}^{P} \beta_j X_{ij} + \varepsilon_i \tag{2}$$

where ε_i is the "residual error" associated with Y and (theoretically) has variance σ_ε^2. The fitted model containing the estimates of the β_j's, then, is

$$\hat{Y}_i = \sum_{j} b_j X_{ij} \tag{3}$$

where the circumflex or "hat" on Y denotes the predicted or estimated value of the response. It is like an average (where a "bar" is used). In fact, it is the conditional average, given the location in the space defined by the X_{ij}'s. It is an estimate of the expected or true value of the response for that location or set of conditions.*

The differences between the observed and fitted values of Y are the residual errors or, simply, "residuals,"

$$e_i = Y_i - \hat{Y}_i = \hat{\varepsilon}_i \tag{4}$$

where e_i is an estimate of the "true error" ε_i. In practice, e_i may contain anything the analyst chooses to omit from the model. It has sample variance

$$s_{Y\cdot X}^2 = s_e^2 = \frac{\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2}{n - P - 1} = \frac{\sum_{i=1}^{n} e_i^2}{n - P - 1} = \hat{\sigma}_\varepsilon^2 \tag{5}$$

which, for the theoretical case, is an estimate of the "experimental" error variance. The subscript $Y \cdot X$ ("Y dot X") means "for Y, given the model containing a particular set of X's." Thus $s_{Y\cdot X}^2$ is the sample estimate of the residual variance in Y, given the model.

The least-squares method chooses values for the b_j's of Eq. (3), which are unbiased estimates of the β_j's of Eq. (2). The least-squares estimates are universally minimum variance unbiased estimates for normally distributed residual errors and are minimum variance among all linear estimates (linear combinations of the observed Y's), regardless of the residual error distribution shape (see Eisenhart 1964). The b_j's (as well as the $\hat{Y}_i$'s) are linear combinations of the observed Y_i's. The least-squares method determines the weight given to each Y value. The derivations of the least-squares solution and/or associated equations used later in this chapter are shown in other sources (see Additional Reading). In essence, the b_j's are chosen to minimize the numerator of Eq. (5)—the sum of squares of e_i's of Eq. (4)—hence "least squares."

In returning to the example problem, a geometric interpretation is presented first. Figure 1 is a plot of steam consumption vs. degree-days from Table 3. The regression coefficient, b_1 is represented by the slope of the least-squares line. It is the tangent of the angle θ. The e_i's whose squares are to be summed to a minimum are distances measured in the Y direction from the points to the line. They are illustrated by typical distances, e_4 and e_{17}.

The least-squares solutions for the simple model are

*It is important to realize that, although the model may be used for predicting future $\hat{Y}$ values, Y does not predict their individual behavior but estimates the conditional average about which those individuals are expected to vary.

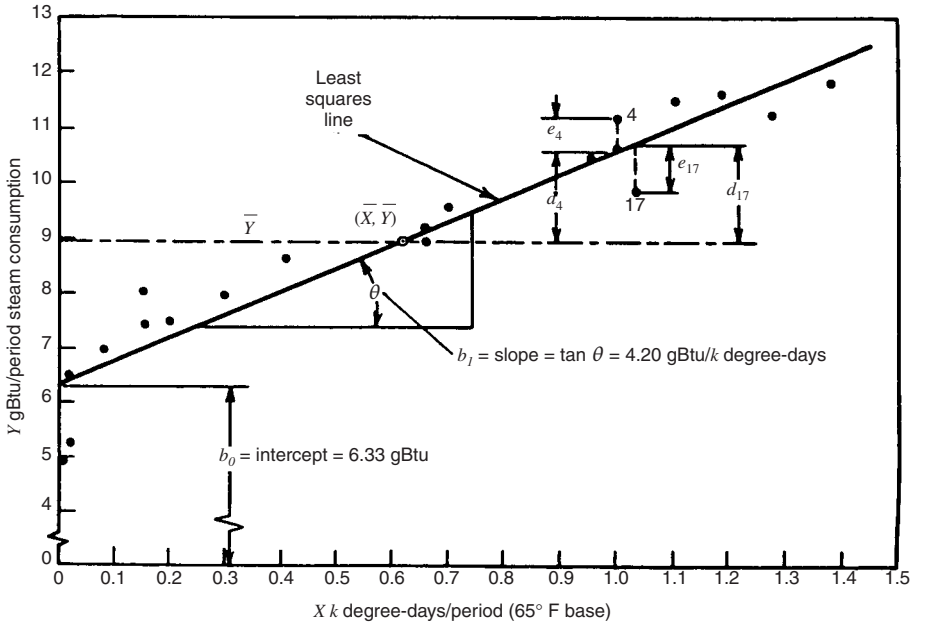


Figure 1 Relationship of Steam Use to Degree Days.

$$b_1 = \frac{\text{SPXY}}{\text{SSX}} \quad b_0 = \overline{Y} - b_1 \overline{X} \tag{6}$$

where SPXY = the (corrected) sum of products of the XY pairs
SSX = the (corrected) sum of squares of X's
$\overline{Y}$ and $\overline{X}$ = the arithmetic averages (which are also least-squares estimators) of the two variables

These averages and sums (with all sums taken for $i = 1$ to n) are

$$\begin{aligned} \overline{X} &= \sum X_i/n & \overline{Y} &= \sum Y_i/n \\ \text{SPXY}^* &= \sum (X_i - \overline{X})(Y_i - \overline{Y}) &&= \sum X_i Y_i - n\overline{XY} \\ \text{SSX} &= \sum (X_i - \overline{X})^2 &&= \sum X_i^2 - n\overline{X}^2 \end{aligned} \tag{7}$$

Equations (6) and (7) yield the following values for the example:

$$\begin{aligned} \overline{X} &= 12.369/20 = 0.618 & \overline{Y} &= 178.67/20 &&= 8.93 \\ \text{SPXY} &= 128.42 - 20(0.618)(8.93) &&&&= 17.93 \\ \text{SSX} &= 11.915 - 20(0.618)^2 &&&&= 4.23 \\ b_1 &= 17.93/4.27 = 4.20 & b_0 &= 8.93 - 4.20(0.618) &&= 6.33 \end{aligned}$$

Equation 3 then takes the form $\hat{Y}_i = 6.33 + 4.20X_i$.

2.2. Residual Variance

For $P = 1$, Eq. (5) reduces to $s^2_{Y \cdot X} = \text{SSRes}/(n - 1)$, where SSRes is the residual sum of squares given by

*These are not shown in the computationally easiest form. This form demonstrates meaning.

$$\text{SSRes} = \text{SSY} - \text{SSReg} \tag{8}$$

and where SSY is the (corrected) sum of squares of Y's and SSReg is the regression sum of squares. In Eq. 8, SSY is exactly parallel in form to SSX in Eq. (7), and when divided by its $n - 1$ DOF, it yields the Y mean square, which might be used to estimate the variance of Y. Regardless of the appropriateness of such an interpretation, the expression $\text{SSY}/(n - 1)$ is a measure of the raw variability in the response whose explanation is the goal. The contribution to SSY that is associated with X is SSReg. This is given by

$$\text{SSReg} = b_1\text{SPXY} \tag{9}$$

and is the sum of squared distances from $\overline{Y}$ to the regression line as shown by typical distances d_4 and d_{17} in Figure 1. For this example,

$$\text{SSY} = 82.55 \qquad s_Y = \left(\frac{\text{SSY}}{19}\right)^{1/2} = 2.08$$

$$\text{SSReg} = (4.20)(17.93) = 75.38$$

$$\text{SSRes} = 82.55 - 75.38 = 7.17$$

$$s_{Y\cdot X} = \left(\frac{\text{SSRes}}{18}\right)^{1/2} = 0.631$$

It can be seen that $s_{Y\cdot X}$ is only 30% of s_Y. That is, the regression equation provided a 70% reduction in variation in Y. Another way to evaluate this residual standard deviation or residual standard error is to compare it to the mean of Y. In this case it is $100\ (0.631/8.93) = 7.0\%$ of the mean and, depending on the goal, might represent a satisfactory reduction in variability. Still another way to measure the association between Y and X, and hence the residual lack of association, is to use the correlation coefficient that is developed next.

2.3. Correlation

The theoretical concept of correlation arises in conjunction with the bivariate normal distribution function. That function has five parameters. If the two variables are X and Y, the parameters are the means (μ_x, μ_y) and the variances (σ_x^2, σ_y^2) of each variate and a measure of covariation, the correlation coefficient, ρ (rho). This chapter does not deal with the theoretical bivariate (or multivariate) normal distribution. However, in practice, the sample correlation coefficient, r, is a useful measure of linear association. It is a dimensionless ratio ranging from -1.0 (perfect inverse linear agreement) through zero (orthogonal or linearly unrelated) to $+1.0$ (perfect direct linear agreement). The value can be obtained from Eq. (10) and used as an index without any assertion whatever being made about distribution form.

$$r_{XY} = \frac{\text{SPXY}}{[(\text{SSX})(\text{SSY})]^{1/2}} = \frac{s_{XY}}{s_X s_Y} \tag{10}$$

The first form of the expression for r_{XY} has the same numerator as b_1 in Eq. (6), which shows that it is just a rescaling of the same basic information. It is easily shown that $r_{XY} = b_1\, s_X/s_Y$.* In the second form in Eq. 10, s_{XY} is the sample covariance (not standard deviation). It has the same sign as SPXY and r and is $\text{SPXY}/(n - 1)$.

The square of r is called the coefficient of determination. It ranges from 0 to 1 and can be interpreted as the fraction of the variation in Y (with variation represented by SSY) that is accounted for or "explained" by variation in X. Thus and from Eq. (8):

$$r_{XY}^2 = \frac{\text{SSReg}}{\text{SSY}} = 1 - \frac{\text{SSRes}}{\text{SSY}} \tag{11}$$

Using Eq. (10) for the example data, $r_{XY} = 17.93/[(4.27)(82.55)]^{1/2} = 0.955$; $r_{XY}^2 = 0.912$. This is seen to be equal to the result of Eq. 11, where $r_{XY}^2 = 75.38/82.55 = 0.913$ (slight rounding error).

*The subscript order "XY" on r is arbitrary; $r_{XY} = r_{YX}$. But the ratio s_X/s_Y implies that b_1 is for "Y on X." With s_Y/s_{XY} b_1 would be "X on Y," with minimization of squared errors taken in the X direction, a different line except where $r = 1.0$.

So variation in X has accounted for 91% of SSY. This is approximately the same as claiming 91% reduction in the variance of Y (from s_Y^2 to $s_{Y\cdot X}^2$).

2.4. Model Specification

In other circumstances, where the physics and chemistry are not so well understood (e.g., in studying the ''cause'' of a disease), the question may focus on the statistical significance of the relationship. The analyst is attempting to decide whether the relationship seen in the sample is something real or just the result of chance association. This decision is appropriate along all goal sequences except where existing theory permits prior specification* of the model.

Model specification is the process of choosing an adequate representation of reality. To decide this question of reality, the analyst would want a test model for the behavior of the estimator, b_j, when the association is just chance. One way would be to use the t test model with the null hypothesis that $\beta_j = 0$ (or some other appropriate value). The alternative hypothesis might be $\beta_j > 0$. The t distribution is appropriate by the central limit theorem. Then

$$t_j = \frac{b_j}{s_{bj}} \tag{12}$$

with the critical value for t of $t_{n-2,\alpha}$, where α represents the specified degree of risk of rejecting a true null hypothesis (claiming a nonexistent association). The standard errors for b_0 and b_1 are given by

$$s_{b_0} = s_{Y\cdot X}\left(\frac{\Sigma X_i^2}{n\mathrm{SSX}}\right)^{1/2} \tag{13}$$

$$s_{b_1} = \frac{s_{Y\cdot X}}{(\mathrm{SSX})^{1/2}} \tag{14}$$

For the example data, $s_{b_0} = 0.236$ and $s_{b_1}, = 0.306$. The corresponding t ratios are $t_0 = 26.9$ and $t_1 = 13.8$, indicating, as was ''known'' in advance, that both constants are statistically well removed from zero (highly ''significant'' compared to a critical value of $t_{18,0.05} = 2.10$). This information is put to more appropriate use later in this section. Statgraphics (a statistical software package) output for this analysis is shown in Figure 2. An intermediate precaution should concern the analyst, that of model validation.

2.5. Model Validation

The least-squares method has permitted each of the data points to play a role in determining the constants b_0 and b_1. It is entirely possible (and nearly always true) that some observations in the data set contain errors (mistakes) in one or more of the variables or arise from unusual conditions that the model is not intended to represent. The ''back substitution'' (obtaining $Y_i - \hat{Y}_i$, values for the development data set, as shown in Table 3) may reveal suspicious points. Generally, residual errors in excess of $\pm 2s_{Y\cdot X}$ should be viewed with mild suspicion, although about 1 in 20 is expected to be in these regions. More sensitive measures will be developed in Section 5. There is an extensive literature associated with this problem of dealing with ''outliers''. One discussion that might serve as a starting point is Barnett (1978).

Perhaps the most useful graphic for examining residuals is a plot of *Y-YHAT* vs. *YHAT*. That will be illustrated later for the multiple regression model. In the simple case, one need not bother. Such a plot is equivalent to tipping Figure 1 so that the regression line is horizontal. No sophisticated techniques are required to see that the fit is poor. Ten of the first 11 points lie above the line. It may be tempting to suggest, say, a quadratic in X to achieve a better fit. However, there is no theoretical reason to expect curvature in the relationship. More properly, additional predictors might be sought, as will be shown presently. The development for the simple case is continued later, ignoring the lack of fit.

One possibility for testing the model's stability is to validate it on fresh data, that is, by back-substituting data points that were not used in developing the model.

This method of validation is called external validation. A method known as internal validation (which is arguably far superior) will be presented in Section 5.

*The distinction between specification and estimation is rarely made. See Hunter and Box (1965) for further discussion. Also see Healy (1978) regarding significance testing.

```
Simple Regression - Steam vs. Heat

Regression Analysis - Linear model: Y = a + b*X
------------------------------------------------------------------------------
Dependent variable: Steam
Independent variable: Heat
------------------------------------------------------------------------------
                              Standard          T
Parameter        Estimate       Error       Statistic        P-Value
------------------------------------------------------------------------------
Intercept         6.33429     0.235869        26.8551          0.0000
Slope             4.20295     0.305585        13.7538          0.0000
------------------------------------------------------------------------------

                            Analysis of Variance
------------------------------------------------------------------------------
Source              Sum of Squares     Df  Mean Square    F-Ratio       P-Value
------------------------------------------------------------------------------
Model                      75.3754      1      75.3754     189.17        0.0000
Residual                    7.1723     18     0.398461
------------------------------------------------------------------------------
Total (Corr.)              82.5477     19

Correlation Coefficient = 0.95557
R-squared = 91.3113 percent
Standard Error of Est. = 0.631238
```

Figure 2 Statgraphics Regression Analysis for Steam vs. Heat.

2.6. Coefficient Estimation

Suppose the model has been specified from existing theory, or by exploration, testing, and validation, and is judged adequate. Back-substitution residual errors are well behaved. Now, whether the values of the b_j's themselves are of interest or whether they are simply to be used in the equation for predicting future values of Y, the precision with which they estimate the β_j's is of concern.* Point estimates were obtained from Eq. (6), but coefficient estimation is not complete without obtaining interval estimates. It is not sufficient, where estimation is the goal (or a step on the path to the goal), just to have "significant" t ratios.

Use of the confidence interval (CI) concept helps to contrast these two steps, specification and estimation. With $\alpha = 0.05$ risk that the CI will not contain β as is claimed, the interval is

$$100(1 - \alpha)\ \text{CI} = 95\%\ \text{CI} = b_j \pm t_{n-2,0.05}\, s_{bj} \tag{15}$$

Notice that (with prescribed t) ts_b represents the maximum probable error associated with the estimate of β. This can be expressed as a percentage error (where engineers very often seek estimates that are within 5 or 10%). Using b as the base (in the absence of knowing β), let E represent the potential percentage error associated with a t value of 2, an approximate value that is never more than 2% in error for $df \geq 30$. (For $df < 30$, substitution of the correct value is advised.) Then

$$E = \frac{100ts_b}{b} = \frac{200s_b}{b} \tag{16}$$

But notice that s_b/b is just the inverse of the t ratio calculated from the sample using Eq. (12).

$$E = \frac{200}{t} \tag{17}$$

This implies the need for calculated t values of 20 or even 40 to meet our common expectation of a 10 or 5%, respectively, error of estimation!

*In general, the requirements for precision will be greatest for control, least for prediction (see Section 3.1.2).

From Eqs. (14) and (16) it can be seen that the error in estimating the slope is directly proportional to $s_{Y\cdot X}$ and inversely proportional to $(\mathrm{SSX})^{1/2}$. Thus, to achieve a prescribed value of E, either of two things must be done: (1) an improved (less noisy) model must be found to reduce residual error, or (2) a larger sample must be obtained to increase SSX (see Crocker 1985 for a more complete discussion of this issue). In general, precision will improve approximately as the square root of n.

2.7. Interval Estimation for a Point on the Line

The regression equation can be used to estimate the "true" value of the response for some specified value of the predictor. This is estimating a conditional population mean of Y and is analogous to estimating (unconditionally) the population mean in a univariate setting. The CI for this case is

$$100(1-\alpha)\ \mathrm{CI} = \hat{Y}_c \pm t_{n-2,\alpha}\, s_{\hat{Y}c}$$
$$s_{\hat{Y}c} = s_{Y\cdot X}\left[\frac{1}{n} + \frac{(X_c - \overline{X})^2}{\mathrm{SSX}}\right]^{1/2} \tag{18}$$

where the subscript c denotes the condition, the location in X, at which the estimate is to be made. Notice that the square root of n (again) determines the interval width at the mean of X and that the interval grows wider the greater distance X_c is from $\overline{X}$, the sample mean.

In most texts this CI is presented as a pair of curved lines, implying a confidence band for the entire line. Equation (18) is meant to be used for one specified location. To sustain α as the risk of not containing the true value, the entire procedure of selecting n observations, computing the coefficients, and so on would need to be followed for each X_c. Wider limits would be needed if the analyst desired limits for the entire true line. Acton (1959) gives a good discussion of this and many related concepts.

2.8. Predicting a Future Value

For predicting a future value at X_c, $\hat{Y}_c$ is obtained from the regression equation just as in the CI. Here it is the estimate of the mean about which individual values are expected to vary. The expression for prediction limits for a single future value of Y must recognize this extra source of variation associated with individuals. The interval for prediction is here abbreviated PI and called, for example, a "95% PI" for $\alpha = 0.05$.

$$100(1-\alpha)\mathrm{PI} = \hat{Y}_c \pm t_{n-2,\alpha}\, s_{Y\cdot X}\left[1 + \frac{1}{n} + \frac{(X_c - \overline{X})^2}{\mathrm{SSX}}\right]^{1/2} \tag{19}$$

The use of t in this expression implies the additional requirement that the individuals be normally distributed around the line. If this is not the case, some other constant (possibly with asymmetry) representing the actual distribution will be substituted for t. Statgraphics output for confidence intervals and predictions limits is shown in Table 4.

Again, this process applies for a single prediction. If some fraction of all future values is to be included within the limits, the limits would be called tolerance limits. (The reader is referred again to Crocker 1985 for more detailed discussion.) Table 5 offers selected values of K_c in Eq. (20) for obtaining tolerance intervals (TI) around the line at $\overline{X}(K_1)$ and at $\overline{X} \pm 2s_X(K_2)$. Linear interpolation may be employed to obtain straight-line approximations of the curved tolerance limits. The values of K were obtained by inverse interpolation of the normal distribution for 0.95 confidence of including at least 95% of all future values.

$$0.95/95\%\ \mathrm{TI} = \hat{Y}_c \pm K_c\, s_{Y\cdot X} \tag{20}$$

TABLE 4 95% Prediction and Confidence Limits

	Predicted	95.00% Prediction Limits		95.00% Confidence Limits	
X	Y	Lower	Upper	Lower	Upper
0.005	6.35531	4.94046	7.77015	5.86233	6.84828
0.250	7.38503	6.00567	8.76440	7.00572	7.76435
0.500	8.43577	7.07471	9.79683	8.12964	8.74190
0.750	9.48651	8.12495	10.84810	9.17816	9.79485
1.000	10.53720	9.15641	11.91810	10.15260	10.92190
1.250	11.58800	10.16980	13.00610	11.08560	12.09030
1.380	12.13440	10.69010	13.57860	11.56250	12.70620

TABLE 5 Coefficients for 0.95/95% Tolerance Limits[a] for $P = 1$

n	K_1 (at $\overline{X}$)	K_2 (at $\overline{X} \pm 2s_x$)	n	K_1 (at $\overline{X}$)	K_2 (at $\overline{X} \pm 2s_x$)
5	6.25	8.00	18	2.85	3.14
6	5.01	6.24	20	2.78	3.03
7	4.37	5.32	25	2.65	2.85
8	3.98	4.76	30	2.56	2.72
9	3.71	4.37	40	2.45	2.57
10	3.52	4.09	50	2.38	2.48
12	3.25	3.71	100	2.23	2.28
14	3.07	3.45	200	2.14	2.16
16	2.95	3.27	500	2.07	2.08

[a] Safe as approximate ''simultaneous'' limits within the $4s_x$ range.

3. MULTIPLE LINEAR REGRESSION

The ability of today's computers and software to handle large data sets has provided the analysts with many opportunities and dangers. Many of these dangers are associated with the intercorrelations found among the predictor variables in nonexperimental data sets. The predictor matrix is said to be ''ill conditioned'' or is carelessly referred to as ''multicollinear.'' (''Multicollinear'' really means the polar condition where some of the X's enter into linear combinations, resulting in an indeterminant system. ''Intercorrelation'' is used here to describe the general case of nonorthogonality among the predictors.)

The basic relationships and computational forms, represented in matrix notation, are shown here paralleling the equations of the simple case given earlier ($v = n - P - 1$).

(''true'' Y) $$E(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta} \tag{21}$$

(observed Y) $$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{22}$$

(b_j's) $$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \tag{23}$$

(predicted or fitted Y) $$\hat{\mathbf{Y}} = \mathbf{X}\mathbf{b} \tag{24}$$

$$\mathbf{SSRes} = \mathbf{Y}'\mathbf{Y} - \mathbf{b}'\mathbf{X}'\mathbf{Y} \tag{25}$$

$$\sum_{j=1}^{p} \mathrm{SSReg}_j = \mathbf{b}'\mathbf{X}'\mathbf{Y} = \sum_{j=1}^{P} b_j \mathbf{SPX}_j\mathbf{Y} \tag{26}$$

(variance–covariance) $$\hat{V}(\mathbf{b}) = [\mathbf{X}'\mathbf{X}]^{-1}\hat{\sigma}_\varepsilon^2 \tag{27}$$

$$s_{Y\cdot X}^2 = \hat{\sigma}_\varepsilon^2 = \frac{\mathrm{SSRes}}{n - P - 1} \tag{28}$$

(joint CI for b's) $$(\boldsymbol{\beta} - \mathbf{b})'\mathbf{X}'\mathbf{X}(\boldsymbol{\beta} - \mathbf{b}) \le (P + 1)s_{Y\cdot X}^2 F_{(p+1),v,\alpha} \tag{29}$$

(CI for $\hat{Y}$) $$100(1 - \alpha)\ \mathrm{CI} = \mathbf{X}_c'\mathbf{b} \pm t_{v,\alpha}s_{Y\cdot X}[\mathbf{X}_c'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_c]^{1/2} \tag{30}$$

(PI) $$100(1 - \alpha)\mathrm{PI} = \mathbf{X}_c'\mathbf{b} \pm t_{v,\alpha}s_{Y\cdot X}[\mathbf{X}_c'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_c + 1]^{1/2} \tag{31}$$

3.1. Intercorrelation Effects

In regression modeling, intercorrelation affects the process in three basic ways. These three have many secondary and corollary consequences, which will be easily perceived if the basic three are understood. They are:

1. Potentially enlarged variances of the b_j's
2. Intercorrelated estimates of the b_j's
3. Ambiguity in assessing the individual contributions to the regression sums of squares

3.1.1. Potentially Enlarged Variances

In the theoretical case—with ''correct'' model and fixed residual variance—the variances of the b_j's will grow larger as intercorrelated predictors are added to the model (see Snee 1973) as a consequence

of the inverse matrix in Eq. (27). (Notice that estimates of the variances of the b_j's are obtained because an estimate of residual error variance is used. The true theoretical variances result from using σ_ε^2). In many cases in practice also, the s_b^2's will grow larger with the addition of intercorrelated predictors. This is because the increase due to the inverse matrix will more than offset the decrease due to a smaller $s_{Y\cdot X}^2$, which results from additional regression sums of squares. However, in practice, $s_{Y\cdot X}^2$ is often reduced enough by the extra predictor(s) to offset the intercorrelation effect in the inverse matrix. These considerations are at the heart of the burgeoning variety of (predictor) variables selection schemes currently appearing in the literature. A brief discussion of this topic will provided later (see Hocking 1976).

3.1.2. Intercorrelated Estimates

In the left side of Eq. (27), in addition to the diagonal variances, there are off-diagonal covariances of pairs of b_j's. Just as with correlation between variables in Eq. (10), covariance of b_j's implies correlation of b_j's. Figure 3 depicts the joint sampling distribution for a pair of positively correlated b_j's. The distribution results from repeated samplings of n values of Y for a given X matrix. For the case of $P = 2$, the correlation of the b_j's is equal in magnitude but opposite in sign to the intercorrelation of the X_j's (they tend to be equal and opposite also for $P > 2$). Notice that the unconditional sampling range of, for example, b_2 (shown by distance A) is very large compared to the conditional range of b_2 (shown by distance B), given the particular estimate of β_1. The important consequence of these two considerations is that errors in estimating the β_j's tend to be compensating among intercorrelated predictors. So intercorrelations may adversely affect the precision of estimate of the β_j's but may have little adverse effect on the use of the model for prediction. This last conclusion depends, of course, on the intercorrelations among the predictors staying about the same in prediction as they were in the sample.

3.1.3. Ambiguity in Assessing Contributions

The underlying nature of the problem is easy to comprehend (for an introductory geometric interpretation of these phenomena, see Crocker 1967, 1969). Interpreting the specific consequences in a particular problem can be extremely complicated. This is true because the ambiguity can be of up to Pth order. The problem is further complicated by the existence of two basic classes of intercorrelated ambiguity, which, for $P \geq 3$, can simultaneously be present in all sorts of hierarchical combinations. Here, the surface will only be scratched with an illustration contrasting the two classes for $P = 2$, the least complex intercorrelation situation. (See also Sections 3.9 and 6.)

In most references, ''intercorrelated'' and ''confounded'' are regarded as synonymous. Actually, confounding is only one of the two classes just mentioned. The other has not been given a name by others but is here titled ''resolving.'' This name was chosen because the separate effects of the two or more (resolving) predictors are not ''resolved'' (clearly seen) until they appear in the model together. The contrast between confounding and resolving is shown in Table 6. The circles at the bottom left represent the two predictors. Area is proportional to regression sums of squares with values as shown. The shaded area of overlap represents intercorrelation. The table shows the allocation

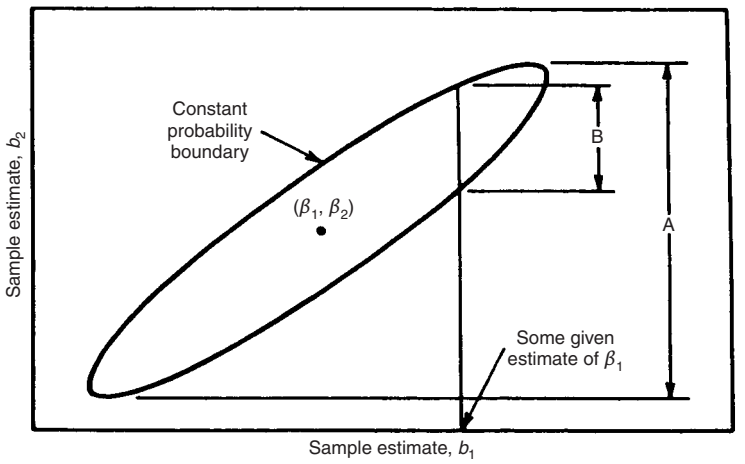
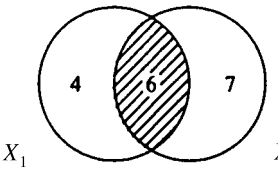


Figure 3 Correlation of Estimates of Slopes.

TABLE 6 Demonstration of Intercorrelation Effects[a]

	Class			
	Confounding		Resolving	
Definition	$R^2 < r_{1Y}^2 + r_{2Y}^2$		$R^2 > r_{1Y}^2 + r_{2Y}^2$	
Model order	X_1 first	X_2 first	X_1 first	X_2 first
$SSReg_1$	10	④	4	⑩
$SSReg_2$	⑦	13	⑬	7



[a]For assumed $n = 103$, SSY = 27, $F_j = 10$ $SSReg_j$.
For last predictor (circled entries), $F_j = t_j^2$.

of the 17 SSReg units to the two predictors for the two classes and for the two possible orderings of predictors in the model.

In the confounding case the ambiguous six units are allocated to the first predictor; the second predictor accounts for the balance. In resolving, the six units are available to the second predictor only after the first has clarified the picture. Notice that the total information, 17 units, is always the same. (other features of this table are discussed in subsequent sections as other diagnostic measures are presented). Relevant to this allocation process, care must be taken in interpreting Eq. (26). This equation says that the total regression sum of squares can be obtained from the sums of products of the b_j's with their corresponding SPX_jY's. It does not assert that the individual $SSReg_j$'s can be found this way. As can be seen from the foregoing discussion, the individual $SSReg_j$'s will depend on the order of appearance in the model. The total is order independent.

Extreme confounding is frequently encountered in nonexperimental data sets. It is important to recognize two quite different situations that may arise. Essentially, there is a duplication of information (i.e., a redundancy in the system). In one situation it may be that the same information is presented twice in slightly different forms (such as two different price indexes). This represents model redundancy and is dealt with by removing the redundant predictor. By contrast, it might be that two really different effects are present, but because in nature they are highly intercorrelated, their separate contributions cannot be discriminated statistically. This represents data redundancy and can clearly present a danger if one or the other predictor is arbitrarily excluded from the model, if estimation is the goal. An example is the use of R&D and capital expenditures to assess the number of technical staff needed in a business. Both effects are real, yet it would not be surprising to see them highly intercorrelated and thus inseparably confounded. This true dilemma motivates the current development of biased estimation techniques such as ridge regression (e.g., see Wichern and Churchill 1978) and will be discussed briefly later.

3.2. Detection of Intercorrelation

Several techniques have been proposed for detecting intercorrelation. These include examination of the off-diagonal elements of $\mathbf{X}'\mathbf{X}$, the examination of eigenvalues of $\mathbf{X}'\mathbf{X}$, the use of principal components, and the use of variance inflation factors (VIFs). Where VIFs are the diagonal elements of the $(\mathbf{X}'\mathbf{X})^{-1}$ matrix. Most current regression software will display these VIFs.

The VIF for each estimated coefficient b_j can be computed as $VIF_j = 1/(1 - R_j^2)$, where R_j^2 is the coefficient of determination obtained from regressing X_j on the other predictor variables. As R_j^2 approaches 1 (i.e., nearly linear dependent) the VIF for the estimated coefficient will tend to infinity. VIFs larger than 10 suggest problems with intercorrelation.

3.3. Meaning of Partial Correlation

For the two-predictor case, the (first-order) partial correlation is given by

$$r_{2Y\cdot 1} = \frac{r_{2Y} - r_{12}\, r_{1Y}}{[(1 - r_{12}^2)(1 - r_{1Y}^2)]^{1/2}} \tag{32}$$

This gives the correlation of X_2 with Y, given X_1 (or ''while holding X_1 constant '' or ''while first

removing the effect of X_1''). For a given pair of correlations of X_1 and X_2 with Y, r_{12} can influence this expression to be larger or smaller in absolute value than it would be in the orthogonal case ($r_{12} = 0$). When the partial is diminished compared to the orthogonal case, confounding exists. When the partial is increased, it is resolving. Partial correlations relating to the example in Table 6 would, in each case, be based on the circled (last-position) values. Ordinary correlations would be based on the uncircled (first-position) values. The coefficient of determination [Eq. (11)] can be used to represent these two views. The ordinary r^2 would use Eq. (11) as is. The partial coefficient of determination would place the circled value in the numerator and the net amount of SSY remaining, after removing the effect of the first predictor, in the denominator. Table 7 shows these ratios based on SSY = 27.

3.4. Multiple Correlation

The multiple correlation is represented by R. It is in fact the correlation of $\hat{Y}$ with Y, where $\hat{Y}$ a linear combination of the X's. Of course, the X's may individually have correlations with Y of either sign. Hence R is arbitrarily defined as being positive. Direct practical interpretation of R is difficult. Two transformations help to improve interpretation. One is R^2. As with the simple model, R^2 is the "coefficient of determination" and represents the fraction of SSY accounted for by the model (R^2 = SSReg/SSY). For orthogonal predictors, $R^2 = \sum_{j=1}^{P} r_{jY}^2$. For $P = 2$, if $R^2 > r_{1Y}^2$, + $r_{2Y}^2 X_1$ and X_2 represent a resolving pair, where $R^2 < r_{1Y}^2 + r_{2Y}^2$, X_1 and X_2 are confounded. These relationships were shown in Table 6 and evaluated in Table 7. A second transformation is "$\%s_Y$ removed." The percentage reduction in s_y is related to R as follows:

$$\%\ s_y \text{ removed} = 100 \left\{ 1 - \left[\frac{(1 - R^2)(n - 1)}{n - P - 1} \right]^{1/2} \right\} \tag{33}$$

For a more extensive discussion of R and a graph of Eq. 33, see Crocker 1972. Another related statistic is the "adjusted" R^2 and is used because the ordinary R^2 will never decrease when a new predictor is added to the model. The adjusted R^2, $\overline{R}^2$, is estimated by replacing SSRes and SSY with their mean squares (MS). The resulting equation is

$$\overline{R}^2 = 1 - \frac{(1 - R^2)(n - 1)}{n - P - 1} = \frac{s_Y^2 - s_{Y \cdot X}^2}{s_Y^2} \tag{34}$$

3.5. Relating t and F in Modeling

As shown in Section 2, the t ratio [Eq. (12)] gives the number of standard errors the estimated value of the coefficient is away from zero. That is still a correct interpretation in the multiple case. It is still useful in assessing the precision of the estimate as per Eq. (17). The t ratio does not, however, measure the contribution, the importance, the practical significance, or even the statistical significance of the associated term in the model! To use this statistic for assessing the contribution of a predictor, it must be carefully qualified. It answers the question "What is the impact of the unique contribution of this predictor?" "Unique" is taken here to mean "impact after resolving." Hence, it is the same as asking what the impact is for this predictor put last in the model.

For answering scientific questions about truth, this gives the t ratio a conservative interpretation. In terms of its influence in reducing $s_{Y \cdot X}^2$, $|t| = 1.0$ is the break-even value for any one predictor. With $|t| > 1.0$, $s_{Y \cdot X}^2$ is reduced by including this predictor. To have (unique) statistical significance, $|t|$ should exceed some appropriate critical value. For excellent precision in estimating β,$|t|$ should be near, say, 20 or 40 (see Section 2.6). The analyst must be wary not to exclude an important term with small t resulting from confounding. What action is appropriate depends heavily on goals (see Section 1) and upon intimate system knowledge.

The ordered F ratio for a single predictor is defined by the ratio of mean squares, regression/residual:

TABLE 7 Coefficients of Determination for Table 6 Values

	r_{1Y}^2	$r_{1Y \cdot 2}^2$	r_{2Y}^2	$r_{2Y \cdot 2}^2$	$R_{Y \cdot 12}^2$	$r_{1Y}^2 + r_{2Y}^2$
Confounding	0.370	0.286	0.481	0.412	0.630 <	0.851
Resolving	0.148	0.500	0.259	0.565	0.630 >	0.407

$$F_j = \frac{\text{MSReg}_j}{\text{MSRes}} = \frac{\text{SSReg}_j/1}{\text{SSRes}/(n - P - 1)} \tag{35}$$

It is called ''ordered'' because it contains the SSReg of the associated predictor, and this quantity is order dependent, as illustrated in Table 6. When $j = P$, $F_j = t_j^2$. Thus, the t ratios are all proportional to the square roots of the respective SSReg obtained for each predictor as if it were in last position.

For the example of Table 6, the denominator of Eq. (35) is $(27 - 17)/100 = 0.1$. Hence the ordered F values are the Table 6 entries multiplied by 10, and for the circled values these are t^2. So it is seen that t ratios are really ''partial'' t ratios and are best interpreted in terms of their relationship to last-position SSReg contributions.

3.6. Dealing with Interactions

Sometimes intercorrelation is carelessly referred to as ''interaction.'' Care should be taken to distinguish these two very different concepts. Intercorrelation is a data phenomenon and is not determined by the form of the regression equation, but rather by the particular set of observed values of the predictor variables. Interaction is a model characteristic. It is represented in the model by the product of two or more predictors. It is put there in an attempt to measure interactive behavior in the system represented by the model. Equation (36) shows an interactive model where $X_3 = X_1X_2$ represents a third predictor created from the first two (subscript i is omitted for simplicity).

$$\hat{Y} = b_0 + b_1X_1 + b_2X_2 + b_3X_3 \tag{36}$$

The meaning of ''interaction'' is this: The effect of one predictor depends on the value of another predictor. This is easily seen to be the case for Eq. (36) by factoring either X_1 or X_2. For illustration, X_1 is used.

$$\hat{Y} = b_0 + (b_1 + b_3X_2)\, X_1 + b_2X_2 \tag{37}$$

Here the coefficient of X_1 is $(b_1 + b_3X_2)$. Therefore, the effect of X_1 (its coefficient, $b_1 + b_3X_2$) depends on the value of X_2. By symmetry, the reverse is also true.

No special steps need to be taken to evaluate an interaction. Its t ratio will assess its additional contribution to SSReg, as was previously discussed. However, care is needed in interpreting the associated main effects. In general, where the X's are in their raw original forms, the interaction term will be highly confounded with the associated main effects—the predictors from which it is formed. This will tend to depress the t ratios of these main effects even where the interaction contributes a sizable SSReg (thereby reducing $s^2_{Y\cdot X}$). This should be of no concern. It is purely an arbitrary scaling problem. If desired, the interaction can be made approximately orthogonal to the main effects by subtracting their respective means before forming the product. This has no effect on the statistical assessment of the interaction.

3.7. Basics for Attribute Modeling

Regression modeling is not limited to using quantitative predictors. Any categorization, classification, or logical distinction can be represented. If there is a single class, no distinction is needed. If there are two classes, (e.g., male, female), an additional X is provided to give an attribute code to distinguish the two classes: $X = 0$ if the individual is in the first class (male), $X = 1$ for the second (female). The value chosen is arbitrary, but 0, 1 coding is easiest to interpret. This is ''differential coding,'' which means that the intercept, b_0, will represent the level in Y of the $X = 0$ group, and the coefficient for this code will estimate the difference in Y between the two classes.

To measure, for example, differences of each working day compared to Monday (arbitrarily chosen as the base of comparison), four extra predictors will be needed. Each will be given the value 1 only if the observation represents the associated day; otherwise it will be given the value 0. In general, the number of predictors added will be one less than the number of classes $(c - 1)$. In statistically evaluating the contribution of such a categorical coding scheme, a single test statistic should be used for the $c - 1$ DOF. This is because individual (single-DOF) SSReg contributions depend on the arbitrary choice of the base of comparison and the order. The total, however, is independent of the choice of base and order. The total can be evaluated using the F ratio as shown in Eq. (38), assuming that these terms are last in the model.

$$F_{c-1,n-P-1,\alpha} = \frac{\sum_{j=p-c+2}^{P} \text{SSReg}_j/(c-1)}{\text{MSRes}} \tag{38}$$

Variables selection programs that operate on individual DOF effects are clearly inappropriate for dealing with categorical structures.

Figure 4 illustrates a model with a single quantitative predictor, a two-class attribute shift, and an interaction of these two. Equation (36) applies here and implies that the slope in X_1 is different for the two classes.

3.8. Dealing with Covariates

A covariate is a source of variation contributing to SSY that may not be of particular interest but whose effect must be removed (1) in order to get unbiased estimates of other predictors of interest and (2) in order to reduce the noise level of the system so that predictors of interest can be more clearly seen. It may be that a covariate is confounded with a predictor of interest. The use of the t ratio in evaluating the reality of that predictor's contribution will then quite properly be conservative—discounting the information held in common with the covariate.

Where a categorical structure of three or more classes is involved in a covariate situation, special care must be taken. If the categorical group is the focus of interest, then it must be placed at the end of the model so that its apparent contribution, evaluated by Eq. (38), will have been reduced according to confounding with a covariate. If the categorical group is the covariate, then the term with which it is confounded will have a properly deflated t ratio independent of model position. Hence attribute code groups can always be safely placed at the end of the model.

3.9. Application to the Example

In the application of the simple model to the data of Table 3, a poor fit was obtained (see Section 2). This motivated an inquiry to find additional predictor variables. The plant engineer suggested testing steam used in production processes. He also later recalled a change in policy regarding comfort heating—a change that was coincident with the data set in hand. Table 8 extends the original data set of Table 3 by adding two more predictors. Production (X_2) is in units per period associated with a process which uses steam for heat in manufacturing. The change in policy is represented by the attribute code (variable 3), which starts at 0 when the heating level was 72°F (22°C) and goes through an adjustment (estimated by the engineer) over several periods to the new level of 65°F (18°C). The policy change should (only) affect the heating coefficient, β_1, and so is introduced as an interaction,

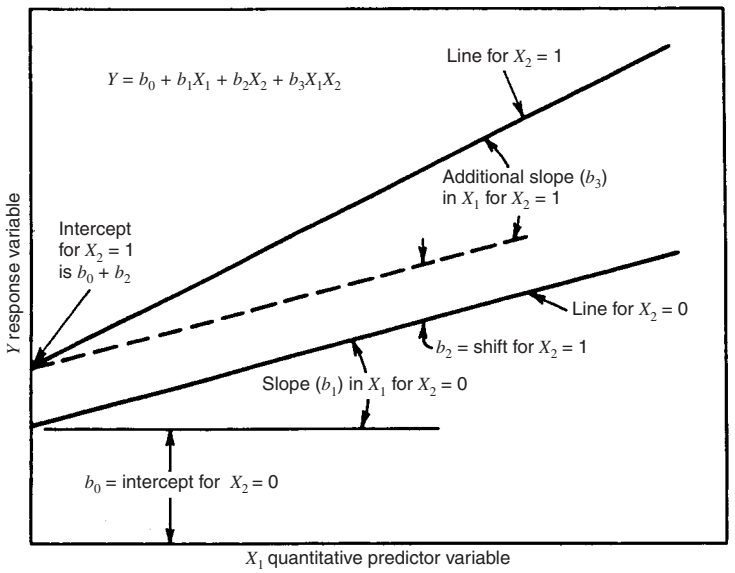


Figure 4 Illustration of an Attribute Code Shift and Slope Change (Interaction).

TABLE 8 Example Steam Consumption—Extended Data Set

i	Period Number	Heat (X_1) k degree-days	Production (X_2) Units	Policy (V_3) Attributes	$X_3 = X_1^* V_3$ k-degree-days	Steam (Y) (gBtu)
1	1-10	0.156	413	0.00	0.000	7.991
2	1-11	0.419	396	0.00	0.000	8.589
3	1-12	0.658	385	0.00	0.000	9.145
4	1-13	1.009	243	0.00	0.000	11.212
5	2-01	1.380	391	0.00	0.000	11.754
6	2-02	1.103	407	0.00	0.000	11.469
7	2-03	1.000	411	0.00	0.000	10.584
8	2-04	0.703	379	0.00	0.000	9.509
9	2-05	0.207	402	0.00	0.000	7.457
10	2-06	0.086	406	0.00	0.000	6.989
11	2-07	0.024	383	0.00	0.000	6.537
12	2-08	0.005	227	0.10	0.001	4.938
13	2-09	0.026	265	0.25	0.007	5.275
14	2-10	0.161	384	0.40	0.064	7.452
15	2-11	0.307	400	0.55	0.169	7.962
16	2-12	0.664	379	0.70	0.465	8.915
17	2-13	1.039	354	0.85	0.883	9.758
18	3-01	1.275	392	1.00	1.275	11.183
19	3-02	1.193	412	1.00	1.193	11.523
20	3-03	0.953	408	1.00	0.953	10.426

X_3 = degree-days × policy. This was suggested when the residuals from the two-predictor model displayed a slight downward trend over time. Essentially, this corrective third predictor is a covariate.

Analysis using Eq. (36) provided residual errors that were examined for pattern and excessive deviance. Figure 5 shows the residual plot (*Y-YHAT* vs. *YHAT*) produced by Statgraphics for the data of Table 8 using the $P = 3$ model and the full data set ($n = 20$). It may be noted that the residuals (*Y-YHAT*) form an arc over the range of *YHAT* and that one point (observation 4) plotted just below $P = 3$ in the title is distinctly away from the group. This fourth point was found to be at $+2.6s_{Y\cdot X}$ (Additional diagnostics relating to this are discussed in Section 6.) Further investigation revealed the malfunctioning of a steam trap in the production system. This would account for an indeterminate excess consumption of steam during the fourth period. Therefore the point was excluded.

Using the remaining 19 observations, the regression analysis was repeated for $P = 3$ [see Eq. (36)]. Statgraphics residual plot for this case is shown in Figure 6. The graph no longer displays any obvious lack of fit, outliers, or nonuniformity of variance. Additional diagnostics for this example will be displayed in Section 6 after some additional concepts have been introduced.

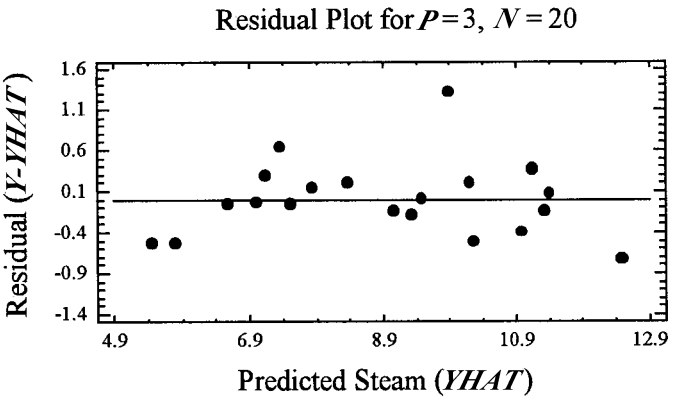


Figure 5 Representation of the Statgraphics Residual Plot.

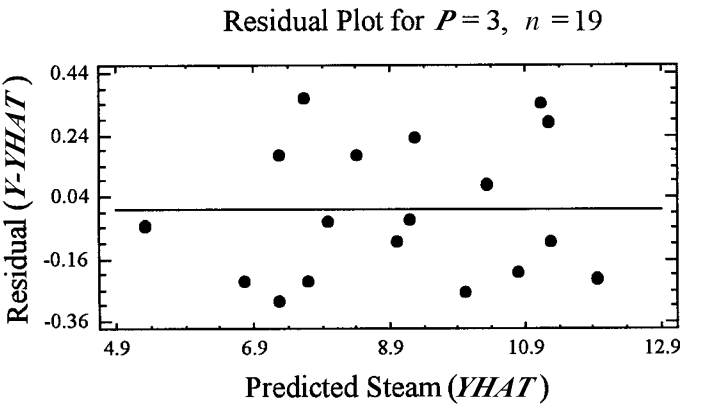


Figure 6 Representation of the Statgraphics Residual Plot.

4. RELATING DIAGNOSTIC QUESTIONS TO GOALS

4.1. Motivation

Regression analysis programs offer a great variety of analytic and diagnostic devices to assist the user. Some statistics that are automatically provided are not always relevant. Sometimes inexperienced users ask what a particular statistic is supposed to "do" or complain that it "isn't any good." Others try to discover some ritualized procedure that can be "followed" and wonder which are the key indicators belonging to such a procedure. Still others, seeking simplistic answers to complex questions, will attempt to impose rules of thumb or employ model selection algorithms. The analyst must learn to understand the complex relationships between diagnostic devices and analytic goals in order to make intelligent use of a program. Not all diagnostics need be examined just because they are there. There is no single, fixed path to follow in doing regression modeling, but rather a complicated cyclical, evolutionary behavioral process that requires simultaneous examination of many relevant diagnostics whose interpretation will frequently generate additional questions.

Space does not permit a complete display relating diagnostics and goals. It is hoped that the following summary of questions diagnostically related to goals will be helpful to the analyst in choosing (or choosing to examine) appropriate diagnostics. In addition to the five project goal-categories presented in Section 1, the analyst has two additional goals while performing regression analysis: data evaluation and model validation. They represent interconnecting analytic steps that motivate the generation of additional diagnostics.

4.2. Summary of Interrelated Diagnostic Questions

1. Are data typical, of adequate range, properly transformed, and error free (especially those of high influence)? Is the data sample large enough? Can a more balanced (less intercorrelated) sample be obtained?
2. Is there evidence of missing predictors or other lack of fit?
3. Is the residual standard deviation small enough to indicate probable model utility?
4. Are intercorrelations substantial? If so, are they reasonable and understood? What effects may they have on interpretation and decisions with respect to specific goals?
5. Do model coefficients have expected signs and reasonable magnitudes? Are they estimated with adequate precision?
6. Is there evidence of overfitting or instability?
7. Is the model to be used in regions of predictor space not seen in the development data set?

5. AN INTRODUCTION TO MODERN DIAGNOSTICS

Many modern regression diagnostics are magically intertwined—mathematically and computationally—and derive from a single germ. The germ idea can best be phrased as a question: What changes occur if a particular observation (one row of the data matrix) is deleted from the data set? Four aspects of change are of particular interest: change in the residual error of the deleted point and

changes in the regression surface, regression coefficients, and residual variance. Another closely related powerful diagnostic is the partial plot (also known as the partial-regression leverage plot). This involves deleting an X column from the data matrix. These row and column deletion constructs will be developed next after proposing some special notation.

5.1. Notation

In addition to the notation shown in Sections 2 and 3, the following conventions will be used in this development:

$\hat{Y}_i(i)$	The ith fitted response value using the regression equation derived with the ith observation row deleted ("Y − *YHAT* sub i not i")
$e_i(i)$	The deleted residual, $Y_i - \hat{Y}_i(i)$ ("e sub i not i")
$\mathbf{b}(i)$	$(P + 1) \times 1$ column vector of estimated regression coefficients derived with the ith row deleted ("**b** not i")
$\mathbf{X}_j$	$n \times 1$ column vector, jth column of **X**, one predictor
$\mathbf{X}_i$	$1 \times (P + 1)$ row vector, ith row of **X**, one observation's X's
$\mathbf{X}(i)$	$(n - 1) \times (P + 1)$ matrix of predictor variables excluding the ith observation row
$\mathbf{Y} \cdot (j)$	$n \times 1$ column vector of residuals in **Y** found when **Y** is fitted using all but the jth predictor
$\mathbf{X}_j \cdot (j)$	$n \times 1$ column vector of residuals in $\mathbf{X}_j$ found when $\mathbf{X}_j$ is fitted using all but the jth predictor
$s_{Y \cdot X}(i)$	Residual standard deviation for a model fitted to a data set excluding the ith observation row

5.2. Getting the Catcher and the Hat

Equation (23) can be rewritten in condensed notation as

$$\mathbf{b} = \mathbf{C}'\mathbf{Y} \tag{39}$$

where $\mathbf{C}'$ is called the matrix of catchers by Mosteller and Tukey (1977) and is defined as

$$\mathbf{C}' = [\mathbf{X}'\mathbf{X}]^{-1}\,\mathbf{X}' \tag{40}$$

So an individual regression coefficient can be found from

$$b_j = \sum_j c_{ij}\, Y_i = \mathbf{C}_j'\mathbf{Y} \tag{41}$$

which shows that the estimated coefficients are just linear combinations of the observed responses. Because each observation potentially contributes differently to this estimation process, each one also has a potentially different expectation for how well it will be fitted, depending on this leverage. This in turn gives an expected error variance for *Y-HAT* for each location in the sample:

$$s_{e_i}^2 = (1 - H_i)\, s_{Y \cdot X}^2 \tag{42}$$

where H_i, the leverage, is the ith diagonal of

$$\mathbf{H} = \mathbf{X}\mathbf{C}' \tag{43}$$

Substituting Eqs. (39) and (43) into Eq. (24) gives

$$\hat{\mathbf{Y}} = \mathbf{X}\mathbf{b} = \mathbf{X}\mathbf{C}'\mathbf{Y} = \mathbf{H}\mathbf{Y} \tag{44}$$

which gives **H** its name: the "hat" matrix (it puts the hat on Y). H_i sums to $P + 1$ and so represents the "consumption" of fractional DOF associated with individual observations. H_i ranges from $1/n$ at the centroid to a maximum of 1.0 for a shift parameter representing a single observation.

Equation (42) expresses the fraction of $s_{Y \cdot X}^2$ associated with the residual error around the surface. The error variance for the surface at the ith location is the complement,

$$s_{\hat{Y}_i}^2 = H_i s_{Y \cdot X}^2 \tag{45}$$

[Notice that for the simple case, H_i is just the familiar $1/n + (X_i - \overline{X}^2)/\mathrm{SSX}$ of Eq. (18)]. The estimated prediction-error variance for future observations recognizes the potential for influence on the model measured by the leverage. The consequence is the augmentation of the unit variance by H; just the reverse of Eq. (42).

$$s^2_{\text{PRED}i} = s^2_{Y\cdot X}\,(1 + H_i) \tag{46}$$

5.3. Row Deletion

One way to examine the effect of deleting the ith observation is to relate the deleted residual, $e_i(i)$, to the residual, e_i. It seems counterintuitive that this should be simply

$$e_i(i) = \frac{e_i}{1 - H_i} \tag{47}$$

because the leverage is dependent only on the X's: This relationship is unaffected by the observed value of Y. It is important to notice, however, that while the ratio of residuals, $e_i(\text{i})/e_i$, grows larger as H_i, the leverage, grows larger, both residuals may be very small. Leverage is a measure of potential influence; actual influence does indeed depend on observed Y. Unfortunately, some authors treat leverage and influence as being synonymous. By some, high-influence points are regarded as a great danger whereas correct high-influence points are to be cherished as the most informative points. (From very large data sets, leverage can be used for selection of a working subset—analogous to a designed experiment with extremes and center points.)

A proof of Eq. (47) was given by Allen (1971). Another (perhaps more accessible) path to Eq. (47) starts with the basic deletion formulas known as the Sherman–Morrison–Woodbury theorem, illustrated by Rao (1973) in an exercise.

5.4. Internal Validation

Conceptually, n regressions can be performed, each with one of the n observations deleted. From each regression the residual error for the deleted point can be calculated. This permits the validation process to be performed on all n observations while using the full available data set to estimate the regression function. How much more sensible this is than holding out some valuable data (information) for external validation! Fortunately, as seen from Eq. (47), only one regression computation need be performed to obtain all the desired information.

One useful statistic that can be derived from the $e_i(i)$ is called PRESS, an acronym for prediction sum of squares.

$$\text{PRESS} = \sum_i [Y_i - \hat{Y}_i(i)]^2 = \sum_i e_i^2(i) \tag{48}$$

It is axiomatic that the model that fits the development data set the best (minimum $s_{Y\cdot X}$) will not be the best prediction model. This is the consequence of the tendency to "overfit" to fit noise. The minimization of PRESS as a criterion for the choice of a predictive model, as suggested by Allen (1971), tends to counter this overfitting and is regarded by many as being superior to the C_P statistic (see, e.g., Mallow 1973), whose use is similarly motivated.

5.5. Examining Residual Errors and Influence

A variety of diagnostics for examining individual observations for extreme deviance and extreme influence can be derived using H_i. Four are chosen here. The following developments [except Eq. (57)] are given by or derived from the work of Belsley et al. (1980). The first diagnostic, the standardized residual for the ith observation, is

$$t_i = \frac{e_i}{s_{Y\cdot X}(1 - H_i)^{1/2}} \tag{49}$$

This statistic (sometimes called the "internalized t ratio") is often examined in screening for outliers (observations that may contain mistakes or may represent unusual conditions). Observations that generate values exceeding 2.0 in absolute value might be routinely examined.

Other analysts prefer the second of the four, the studentized residual (the "externalized t ratio"),

$$t_i(i) = \frac{e_i}{s_{Y\cdot X}(i)(1 - H_i)^{1/2}} \tag{50}$$

where

$$s_{Y\cdot X}(i) = \frac{[(n - P - 1)\, s^2_{Y\cdot X} - e_i^2/(1 - H_i)]^{1/2}}{(n - P - 2)^{1/2}} \tag{51}$$

[The original SSRes is reduced by the product of e_i and $e_i(i)$.] Here, one might usefully examine a

listing of the $s_{Y\cdot X}(i)$ as well. Here, one might also argue for the use of $e_i(i)$ in Eq. (50) with the prediction form (from Eq. (46) in the denominator. It turns out to be exactly the same thing! For the predicted point, the predictive location has $H_i(i)$ found by

$$H_i(i) = \mathbf{X}_i \, [\mathbf{X}'\mathbf{X}(i)]^{-1} \, \mathbf{X}_i' = \frac{H_i}{1 - H_i} \tag{52}$$

which provides the equivalence for the studentized residual. For Gaussian errors the distribution of $t_i(i)$ would closely follow the t distribution with $n - P - 2$ DOF.

One intuitively appealing measure of influence would be the deleted regression surface shift, $\hat{Y}_i - \hat{Y}_i(i)$. In developing a standardized form of this shift for assessment, it is discovered that the result is identical to t_i, the standardized residual of Eq. (49)! Hence, that statistic may be retained to provide this additional meaning. And in passing it is noted that the surface shift is

$$\hat{Y}_i - \hat{Y}_i(i) = \frac{e_i H_i}{1 - H_i} \tag{53}$$

An alternative scaling of this difference may be obtained by answering the question ''Compared to the uncertainty with which the position of the surface has been established at this location, how big is the shift created by this point's inclusion?'' This would use the standard error of the surface instead of the standard error of the shift in the surface. This scaling, the third measure of influence /deviance, has been labeled DFITS_i, defined as

$$\text{DFITS}_i = \frac{\hat{Y}_i - \hat{Y}_i(i)}{s_{\hat{Y}_i}(i)} \tag{54}$$

where

$$s_{\hat{Y}_i}(i) = s_{Y\cdot X}(i) \, H_i^{1/2} \tag{55}$$

Then substituting Eqs. (53) and (55) into (54), DFITS_i becomes

$$\text{DFITS}_i = \frac{H_i^{1/2}}{(1 - H_i)} \frac{e_i}{S_{Y\cdot X}(i)} \tag{56}$$

Cook (1977) developed his index of influence (the fourth examined here) in terms of the shift in the vector $\mathbf{b}$ associated with the deletion of the ith observation [see Eq. (59)]. It is structured so that the shift can be evaluated using $F(P + 1, n - P - 1, 1 - \alpha)$ with $1 - \alpha = 0.10$ giving an upper bound for ''an uncomplicated analysis,'' according to Cook. The index can be reduced to the following form:

$$\text{COOKD}_i = (t_i^2) \frac{1}{P + 1} \frac{H_i}{1 - H_i} \tag{57}$$

The third factor in this expression is the surface shift factor of Eq. (53) and is the ''deleted leverage'' of Eq. (52). It is also the ratio of the two partitioned parts of the residual variance [Eqs. (45) and (42)]:

$$\frac{s_{Y_i}^2}{s_{e_i}^2} = \frac{H_i}{1 - H_i} \tag{58}$$

This says that the larger the leverage, the larger the uncertainty of the location of the surface, the more the shrinkage of the residual. In Eq. (57), this ratio amplifies the measure of deviance given by t_i^2. Squaring DFITS and dividing by $P + 1$ gives the same measure as Eq. except that $t_i^2(i)$ is used instead of t_i^2.

It is noted in passing—for examining the deletion impact on the regression coefficients—that

$$\mathbf{b} - \mathbf{b}(i) = \frac{\mathbf{C}_i' e_i}{1 - H_i} \tag{59}$$

This result is also derived from the work of Beckman and Trussell (1974) and the basic deletion formulas discussed after Eq. (47).

5.6. Partial Regression Leverage Plots

A partial plot reveals the underlying relationship between the response and the jth predictor with the influence of all other predictors removed. That is, it plots $\mathbf{Y} \cdot (j)$ vs. $\mathbf{X} \cdot (j)$. Such a plot may reveal curvature, discontinuities, extreme influence, or other aberrations often not readily detected by plotting *Y-YHAT* or $\mathbf{Y} \cdot (j)$ vs. $\mathbf{X}_j$. This is especially true for resolving predictors. The simple regression of these residual variables gives the partial regression slope for the jth predictor in the full model—the usual multiple regression slope not always qualified as being partial. Deviations around this regression line are the full-model residuals. The correlation of these two residual variables is the $(P - 1)$st-order partial correlation of Y with X_j. One need not perform $2P$ multiple regressions to obtain the required vectors. Mosteller and Tukey (1977) and Velleman and Welsch (1981) discuss details leading to the following results. The starting point is the identity

$$\begin{aligned} \mathbf{Y} &= \sum_{(j)} b_k \mathbf{X}_k + b_j \mathbf{x}_j \cdot (j) + \mathbf{e} \\ &= \hat{\mathbf{Y}}(j) + b_j \mathbf{X}_j \cdot (j) + \mathbf{e} \end{aligned} \tag{60}$$

Then

$$\mathbf{Y} - \hat{\mathbf{Y}}(j) = \mathbf{Y} \cdot (j) = b_j \mathbf{X}_j \cdot (j) + \mathbf{e} \tag{61}$$

From Eq. (61) it is evident that the ordinate of the partial plot is simply $b_j \mathbf{X}_j \cdot (j) + \mathbf{e}$. The abscissa is $\mathbf{X}_j \cdot (j)$. Since the b_j and $\mathbf{e}$ are available from the complete multiple regression, all that is needed is $\mathbf{X}_j \cdot (j)$. This is obtained from

$$[\mathrm{X}_j \cdot (j)]_i = \frac{c_{ij}}{\sum_k c_{kj}^2} \tag{62}$$

where the denominator sums of squares of c_{kj} are just the diagonal elements of $[\mathbf{X}'\mathbf{X}]^{-1}$.

6. DIAGNOSTICS FOR THE EXAMPLE

6.1. Leverage and Influence

Available space is insufficient to permit demonstrating the partial plots for the example. However, Table 9 lists the leverages and the four influence diagnostics associated with individual observations

TABLE 9 Printing of Diagnostics for $P = 3$ $n = 20$

Row	Steam	Predicted	Residual	Studentized Residual	DFITS	COOKD	Leverage
1	7.991	7.357	0.634	1.406	0.592	0.083	0.151
2	8.589	8.376	0.213	0.430	0.131	0.005	0.086
3	9.145	9.333	−0.188	−0.378	−0.114	0.003	0.083
4	11.212	9.900	1.312	7.530	7.013	2.744	0.465
5	11.754	12.482	−0.728	−1.921	−1.375	0.405	0.339
6	11.469	11.396	0.073	0.159	0.083	0.002	0.213
7	10.584	10.979	−0.395	−0.858	−0.402	0.041	0.180
8	9.509	9.487	0.022	0.045	0.014	0.000	0.086
9	7.457	7.503	−0.046	−0.095	−0.035	0.000	0.121
10	6.989	7.009	−0.020	−0.041	−0.018	0.000	0.157
11	6.537	6.589	−0.052	−0.107	−0.044	0.001	0.146
12	4.938	5.469	−0.531	−1.440	−1.248	0.365	0.429
13	5.275	5.809	−0.534	−1.271	−0.791	0.151	0.279
14	7.452	7.149	0.303	0.625	0.219	0.123	0.110
15	7.962	7.825	0.137	0.277	0.090	0.002	0.097
16	8.915	9.056	−0.141	−0.282	−0.074	0.001	0.065
17	9.758	10.269	−0.511	−1.131	−0.531	0.069	0.181
18	11.183	11.318	−0.135	−0.318	−0.225	0.013	0.334
19	11.523	11.144	0.379	0.888	0.570	0.082	0.292
20	10.426	10.219	0.207	0.445	0.216	0.122	0.190

for the example data of Table 8. Observation 4 has the highest leverage (at nearly a half a *df* "consumed" by this one point) and is also shown by all four diagnostics to have high influence. As noted earlier, the steam consumption for this observation was found to be anomalous and so the point was removed from the data set. For contrast, it is noted that observation 12 has nearly as high a leverage (at the low end of the ranges of both X_1 and X_2) but fits rather well and so has much smaller measures of influence (although COOKD_{12} at 0.365 is above the 0.10 value of F, which is 0.259). The regression analysis from which these diagnostics were obtained ($n = 20$, $P = 3$) is shown in Table 10 as produced by Statgraphics.

6.2. Final Results

With the fourth observation removed, the $P = 3$ model was repeated for the $n = 19$ remaining periods. The final results are summarized in Tables 11 and 12. The following comments deal with the main conclusions demonstrated in the table and with some diagnostics not mentioned earlier.

1. The coefficients were judged reasonable in sign and size, and the two main predictors are estimated with reasonable precision [see Eq. (17)].
2. The covariate third predictor does not show strong influence but was retained in the model to avoid biasing the estimates of the other coefficients. (The intercept represents the average consumption of all other steam uses uncorrelated with the model predictors.)
3. The final value for $s_{Y \cdot X}$ of 0.238 is the residual standard deviation for this model and data set.

TABLE 10 Statgraphics Regression Analysis Results for $P = 3$ $n = 20$

Multiple Regression Analysis

Dependent variable: Steam

Parameter	Estimate	Standard Error	t Statistic	P-Value
CONSTANT	3.93783000	0.77098700	5.10752	0.0001
Heat	4.30712000	0.29227900	14.73630	0.0000
Production	0.00665141	0.00209879	3.16917	0.0060
Heat × Policy	−0.56407600	0.31144900	−1.81114	0.0889

Analysis of Variance

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
Model	78.48270	3	26.160900	102.97	0.0000
Residual	4.06502	16	0.254064		
Total (Corr.)	82.5477	19			

R-squared = 95.0756 percent
R-squared (adjusted for d.f.) = 94.1522 percent
Standard Error of Est. = 0.504047
Mean absolute error = 0.328052
Durbin-Watson statistic = 2.31226

Unusual Residuals

Row	Y	Predicted Y	Residual	Studentized Residual
4	11.212	9.9	1.312	7.53

Influential Points

Row	Leverage	Mahalanobis Distance	DFITS
4	0.464520	14.66730	7.01308
5	0.338859	8.27828	−1.37505
12	0.428938	12.57280	−1.24796

Average leverage of single data point = 0.2

TABLE 11 Statgraphics Regression Analysis Results for $P = 3$ $n = 19$

Multiple Regression Analysis
Dependent variable: Steam

Parameter	Estimate	Standard Error	t Statistics	P-Value
CONSTANT	1.9782800	0.44763600	4.41940	0.0005
Heat	3.7680100	0.15553200	24.22660	0.0000
Production	0.0122702	0.00124091	9.88810	0.0000
Heat × Policy	−0.2403990	0.15328000	−1.56837	0.1376

Analysis of Variance

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
Model	76.232000	3	25.4107000	448.18	0.0000
Residual	0.850463	15	0.0566975		
Total (Corr.)	77.0824	18			

R-square = 98.8967 percent
R-squared (adjusted for d.f.) = 98.676 percent
Standard Error of Est. = 0.238112
Mean absolute error = 0.188343
Durbin-Watson statistic = 1.96748

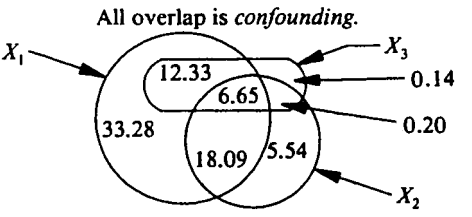
95.0% confidence intervals for coefficient estimates (Steam)

Parameter	Estimate	Standard Error	Lower Limits	Upper Limit	V.I.F.
CONSTANT	3.93783000	0.77098700	2.30341000	5.5722500	
Heat	4.30712000	0.29227900	3.68751000	4.9267200	1.43474
Production	0.00665141	0.00209879	0.00220216	0.0111006	1.06577
Heat × Policy	−0.56407600	0.31144900	−1.22432000	0.0961673	1.42454

4. The first predictor accounts for about 91% of the variability in Y (as measured by ordered SSReg/SSY). Therefore, it might be tempting to dismiss the use of steam for production (X_2) as unimportant. But X_2 accounts for 53% of the average steam use. That use simply does not vary as severely as weather.
5. The drop from $s_Y = 2.069$ to $s_{Y \cdot X} = 0.238$ represents an 88.5% reduction, and with $s_{Y \cdot X}$ at less than 3% of the mean, this model promises to be effective in monitoring consumption.
6. The VIF for each X is shown to help assess the intercorrelation effects.
7. The SSReg total of 76.23 is allocated among the three predictors according to model position. These SSReg_j can also be subdivided according to the effects of intercorrelations of the predictors. In this uncomplicated system the confounded portions are easily determined to be as shown in Table 12 by both the diagram and the listing to its left.

TABLE 12 SSReg Allocation and Diagram for $P = 3$, $n = 19$

SSReg Class	X_1	X_2	X_3
Unique	33.28	5.54	0.14
Two-Way confounding	18.09(2)	0.20(3)	—
Two-Way confounding	12.33(3)	—	—
Three-Way confounding	6.65	—	—
SSReg_j Totals	70.35	5.74	0.14



7. OTHER REGRESSION TOPICS

7.1. Variable Selection

Variable selection in regression arises when the set of variables to include into the model is not predetermined. The problem to be addressed is from the list of potential candidates to include in the model which ones should be included and in what form. The objectives here are to include as many predictors that can influence the prediction while in addition including as few as possible because the variance of the prediction increases as the number of predictors increase. Hence the goal of variable selection is to find an ''appropriate subset'' regression model.

Several criteria have been proposed to compare and evaluate the adequacy of the subset regression models. These include using R^2, adjusted R^2, MSE, Mallows' C_p, and PRESS. A brief description of the adjusted R^2 method will be provided here. Adjusted R^2 was previously defined as

$$\overline{R}^2 = 1 - \frac{(1 - R^2)(n - 1)}{n - P - 1} \tag{34}$$

The adjusted R^2 is used because the ordinary R^2 defined earlier will always increase when new terms are added to the regression model. The adjusted R^2 will not necessarily increase as new terms are added. This helps prevent over fitting the model and determining the ''appropriate subset'' regression model. Therefore one criterion associated with determining the appropriate subset model is to maximize the adjusted R^2. Note that this is equivalent to selecting the model with a minimum MSE.

7.1.1. All Possible Regressions

This procedure involves fitting all possible subset of regression models and choosing the ''best'' model based on suitable criteria. If we include the intercept in each model and there are q predictors, then there will be 2^q total equations to be fitted. Thus, if the number of predictors is 5, the total number of equations to be fitted is 32. As can be seen, the number of equations to fit grows rapidly as the number of predictors increase. In the steam example there are 3 predictor variables and thus 8 possible equations to fit. The results of fitting these 8 equations are shown in Table 13 and Figure 7. The equations have been listed by order of maximum adjusted R^2.

For cases where the number of possible equations is large, there are several procedures developed to evaluate only a small subset of these equations by adding or deleting predictors one at a time. These procedures can be classified into three groups (1) forward selection, (2) backward elimination, and (3) stepwise and are briefly described below.

7.1.2. Forward Selection

The procedure begins with the assumption that there are no predictors in the model other than the intercept. An optimal subset is determined by adding predictors into the model one at time with the first one to enter being the predictor with the largest simple correlation with the response variable. This predictor will be entered if it exceeds a predetermined F value (*Fin*). The second predictor to enter the model is then determined by the one that has the largest correlation with the response after adjusting for the effect of the first predictor (i.e., largest partial correlation). This predictor will enter the model if it exceeds *Fin*. This process continues until a predictor does not exceed the *Fin* or when all predictors have been added to the model.

TABLE 13 Statgraphics Output for Regression Model Selection for Steam

Models with Largest Adjusted R-Squared Results

MSE	R-Squared	Adjusted R-Squared	Cp	Included Variables
0.254064	95.0756	94.15220	4.00000	X_1 X_2 X_3
0.288141	94.0660	93.36790	5.28021	X_1 X_2
0.389220	91.9843	91.04130	12.04360	X_1 X_3
0.398461	91.3113	90.82860	12.23030	X_1
3.484540	28.2389	19.79640	219.15900	X_2 X_3
3.676710	19.8271	15.37310	244.48900	X_3
3.934500	14.2060	9.43971	262.75300	X_2
4.344620	0.0000	0.00000	306.90900	

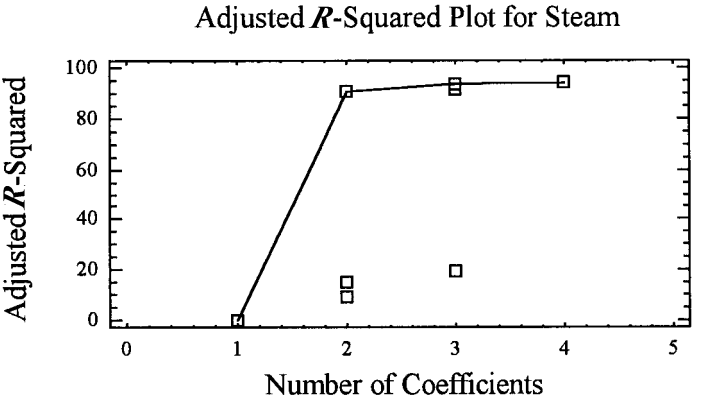


Figure 7 Statgraphics Adjusted R^2 Plot for Steam.

7.1.3. Backward Elimination

This procedure begins with the full equation fitted and successively drops one predictor out of the equation at a time. The predictor that is the first candidate to be eliminated is the one with the smallest contribution to the reduction of the error sum of squares. Based on the full model, a partial F statistic is computed for each predictor as if it were the last variable to enter the model. The predictor with the smallest partial F is eliminated if it is less then a predetermined critical F value (*Fout*). This process continues until no predictor has a partial F less than *Fout*.

7.1.4. Stepwise

The stepwise method is basically a combination of the forward selection and backward elimination methods. Thus, at each stage of the forward selection the possibility of deleting a predictor is also considered. Therefore, a variable that enters at an earlier stage may later be removed.

For all three methods the final model selected like any regression model should be evaluated to the regression diagnostics described earlier.

As an illustrative example of the method, the data for the steam example was analyzed using the forward selection method. The results are shown in Table 14. In this example the first predictor to enter was heat. Next the predictor production was added to the model. The last predictor, heat × policy, did not exceed the *Fin* value and was not included in the final model. However, as discussed earlier, the covariate third predictor was retained in the model to avoid biasing the estimates of the other coefficients.

7.2. Ridge Regression

Ridge regression is employed to combat intercorrelation between the regressors. A set of variables is exactly collinear if one of them is a linear combination of the others. The presence of intercorrelation is given by the variance inflection factors (VIF).

As discussed, least squares provides unbiased estimates with minimum variance of all linear unbiased estimators without upper limit on the variance of the estimators, and if intercorrelation exists, this may produce large variance. Therefore, in the presence of intercorrelation, a penalty is paid for the unbiasness property that is usually attained via least squares. Biased estimation procedures attempt to find a biased estimator of a regression coefficient that has smaller variance than the unbiased coefficient. Ridge regression is a biased estimation procedure to address this. In ridge regression, the analyst would like to select a bias, k, such that the reduction in variance is greater than the increase in the squared bias introduced. The ridge regression estimator, b_r is given by

$$b_r = (\mathbf{X}'\mathbf{X} + kI)^{-1}\,\mathbf{X}'y \tag{63}$$

The choice of k belongs to the analyst and should be chosen where strong evidence shows more stable estimates or improved prediction. One method suggested by Hoerl and Kennard (1970) is the use of a ridge trace. The ridge trace is a plot of the b_r vs. k, usually in the interval (0, 1). For values close to $k = 0$, intercorrelation will cause rapid changes in b_r. The objective is to select a small value of k where the b_r's stabilize.

TABLE 14 Statgraphics Output of Forward Selection Method for Steam

Multiple Regression Analysis

Dependent variable: Steam

Parameter	Estimate	Standard Error	t Statistic	P-Value
CONSTANT	4.11739000	0.81425000	5.05667	0.0001
Heat	4.03422000	0.26671300	15.12570	0.0000
Production	0.00624244	0.00222214	2.80920	0.0121

Analysis of Variance

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
Model	77.6493	2	38.824700	134.74	0.0000
Residual	4.8984	17	0.288141		
Total (Corr.)	82.5477	19			

R-squared = 94.066 percent
R-squared (adjusted for d.f.) = 93.3679 percent
Standard Error of Est. = 0.536788
Mean absolute error = 0.34515
Durbin-Watson statistic = 1.82542

Stepwise Regression
Method: forward selection
F-to-enter: 4.0
F-to-remove: 4.0
Step 0:
0 variables in the model. 19 d.f. for error.
R-squared = 0.00% Adjusted R-squared = 0.00% MSE = 4.34462

Step 1:
Adding variable Heat with F-to-enter = 189.166
1 variables in the model. 18 d.f. for error.
R-squared = 91.31% Adjusted R-squared = 90.83% MSE = 0.398461

Step 2:
Adding variable Production with F-to-enter = 7.89159
2 variables in the model. 17 d.f. for error.
R-squared = 94.07% Adjusted R-squared = 93.37% MSE = 0.288141
Final model selected.

8. SOME PRACTICAL CONCERNS

The analyst faces a variety of dangers in practice in addition to those discussed earlier. For example, there is often pressure to ''keep it simple.'' The danger is that, by avoiding complexity, the analyst may be seriously misled or fail to develop an adequate model. Simplicity is not necessarily a virtue.

It is also easy to acquire more faith in a complex regression model than it deserves. Even a good model is at best a crude approximation of reality. Yet, by being computer born, it takes on a special aura that may encourage undeserved faith in its utility.

Another danger is the predictive use of a regression model in regions of the joint predictor space not observed in the development sample, even though within all the observed predictor ranges. Equation (52) provides one index of the presence of this condition. ''Interpolation'' seems to fit, as the result is compared to the distribution of sample leverages.

From initial questions of which variables to collect to the final checking of a suspicious residual, data management is a major constituent of any modeling venture. The collection and ''scrubbing'' process very often consumes 70% or more of project funds and elapsed time. It bears repeating that a data set will rarely ever be free of mistakes. Some process for auditing the data must be devised at the outset or much effort will be wasted on false starts.

8.1. Model Use and Maintenance

The model is a waste if it is not used. The person who will have responsibility for its use must be involved early enough to gain understanding and develop faith. The model's use must serve an ongoing function that is desired and expected by the user's superiors or it will not survive.

Provision must be made for the timely reporting of the predictors. It is of no use to develop a prediction or control model if the necessary data cannot be obtained in a timely fashion. Results must then be reported to those who can take action. Good predictions kept in a desk serve no one.

Invariably, in practice, the β's estimated are not in fact constant but are creeping and shifting overtime. Additionally, there will inevitably be other systems changes, which, for example, may require the inclusion of additional predictors. So, if the model is to continue in use, provision must be made for updating it. Failing this, the model will begin to miss until it loses credibility and its use is discontinued.

8.2. Helpful Hints in Practice

The following is a summary list of prerequisites for successful use of regression modeling techniques that an analyst should have and/or use:

1. Reasonably specific goals
2. An understanding of statistical procedures
3. Reasonable familiarity with the system modeled
4. Restraint in transforming variables
5. Facility for adequate diagnostic analysis and data scrubbing
6. A cyclical approach with documentation of decisions and choices made
7. Good judgment instead of model selection algorithms
8. Great care when excluding important predictors that were not permitted to vary
9. Great care when including ''discovered'' relationships
10. A willingness to validate the model and/or anticipate model instability
11. Recognition of the need for maintenance of the model

Computer Software

Statgraphics, Manugistics, Rockville, MD.

REFERENCES

Acton, F. S. (1959), *Analysis of Straight Line Data,* John Wiley & Sons, New York.

Allen, D. M. (1971), ''Mean-Square Error of Prediction as a Criterion for Selecting Variables,'' *Technometrics,* Vol. 13, pp. 469–475.

Barnett, V. (1978), ''The Study of Outliers: Purpose and Model,'' *Applied Statistics,* Vol. 27, No. 3, pp. 242–250.

Beckman, R. J., and Trussell, H. J. (1974), ''The Distribution of an Arbitrary Studentized Residual and the Effects of Updating in Multiple-Regressions,'' *Journal of the American Statistical Association,* Vol. 66, pp. 199–201.

Belsley, D. A., Kuh, E., and Welsch, R. E. (1980), *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity,* John Wiley & Sons, New York.

Cook, R. D. (1977), ''Detection of Influential Observations in Linear Regression,'' *Technometrics,* Vol. 19, pp. 15–18.

Crocker, D. C. (1967), ''Intercorrelation and the Utility of Multiple Regression in Industrial Engineering,'' *Journal of Industrial Engineering,* Vol. 18, No. 1, pp. 79–85.

Crocker, D. C. (1969), ''Linear Programming Techniques in Regression Analysis: The Hidden Danger,'' *AIIE Transactions,* Vol. 1, No. 2, pp. 112–126.

Crocker, D. C. (1972), ''Some Interpretations of the Multiple Correlation Coefficient,'' The American Statistician, Vol. 26, No. 2, pp. 31–33.

Crocker, D. C. (1985), *Volume 9: How to Use Regression Analysis in Quality Control,* American Society for Quality Control, Milwaukee.

Eisenhart, C. (1947), ''The Assumptions Underlying the Analysis of Variance,'' *Biometrics,* Vol. 3, No. 1, pp. 1–21.

Eisenhart, C., ''The Meaning of 'Least' in Least Squares,'' *Journal of the Washington Academy of Sciences,* Vol. 54, February, pp. 24–32.

Finch, P. D. (1979), "Description and Analogy in the Practice of Statistics," *Biometrika,* Vol. 66, No. 2, pp. 195–208.

Healy, M. J. R., "Is Statistics a Science?" *Journal of the Royal Statistical Society,* Vol. 141, A1978, Part 3, pp. 385–393.

Hocking, R. R. (1976), "The Analysis and Selection of Variables in Linear Regression," *Biometrics,* Vol. 32, No. 1, pp. 1–50.

Hoerl, A. E., and Kennard, R. W. (1970), "Ridge Regression: Applications to Nonorthogonal Problems," *Technometrics,* Vol. 12, pp. 69–82.

Hunter, W. G., and Box, G. E. P. (1965), "Experimental Studies of Physical Systems," *Technometrics,* Vol. 7, No. 1, pp. 23–42.

Mallows, C. L., "Some Comments on C_p," Technometrics, Vol. 15, pp. 661–675.

Mosteller, F., and Tukey, J. W. (1977), *Data Analysis and Regression,* Addison-Wesley, Reading, MA.

Rao, C. R. (1973), *Linear Statistical Inference and Its Applications,* 2nd Ed., John Wiley & Sons, New York, p. 33.

Snee, R. R. (1973), "Some Aspects of Nonorthogonal Data Analysis," *Journal of Quality Technology,* Vol. 5, No. 2, pp. 67–79.

Velleman, P. F., and Welsch, R. E. (1981), "Efficient Computing of Regression Diagnostics," *American Statistician,* Vol. 35, No. 4, November, pp. 234–242.

Wichern, D. W., and Churchill, G. A. (1978), "A Comparison of Ridge Estimators," *Technometrics,* Vol. 20, No. 3, 1978, pp. 301–311.

ADDITIONAL READING

Allen, D. M., and Cady, F. B., *Analyzing Experimental Data by Regression,* Lifetime Learning, Belmont, CA, 1982.

Chatterjee, S., and Price, B., *Regression Analysis by Example,* John Wiley & Sons, New York, 1977.

Dobson, A. J., *An Introduction to Statistical Modelling,* Chapman & Hall, New York, 1986.

Draper, N. R., and Smith, H., *Applied Regression Analysis,* 2nd Ed., John Wiley & Sons, New York, 1981.

Farebrother, R. W., *Linear Least Squares Computations,* Marcel Dekker, New York, 1988.

Freund, R. J., and Minton, P. D., *Regression Methods,* Marcel Dekker, New York, 1979.

Guttman, I., *Linear Models: An Introduction,* John Wiley & Sons, New York, 1982.

Horton, R. L., *The General Linear Model: Data Analysis in the Social and Behavioral Sciences,* McGraw-Hill, New York, 1978.

Kleinbaum, D. G., and Kupper, L. L., *Applied Regression Analysis and Other Multivariate Methods,* Duxbury Press, North Scituate, MA, 1978.

Montgomery, D. C., and Peck, E. A. (1992), *Introduction to Linear Regression Analysis,* John Wiley & Sons, New York.

Myers, R. H., *Classical and Modern Regression with Applications,* PWS-Kent, Boston, 1990.

Neter, J., and Wasserman, W., *Applied Linear Statistical Models,* Richard D. Irwin, Homewood, IL, 1974.

Rice, J. R., *Matrix Computations and Mathematical Software, McGraw-Hill, Tokyo, 1983.*

Wesolowsky, G. O., *Multiple Regression and Analysis of Variance,* John Wiley & Sons, New York, 1976.

Younger, M. S., *A Handbook for Linear Regression,* Duxbury Press, North Scituate, MA, 1979.