Division of Research
Graduate School of Business Administration
The University of Michigan

May 1976

EXPLORATORY ANALYSIS OF MULTIVARIATE DATA FOR BUSINESS RESEARCH

Working Paper No. 130

by
Roger L. Wright
and
W. Allen Spivey

The University of Michigan

FOR DISCUSSION PURPOSES ONLY

None of this material is to be
quoted or reproduced without
the express permission of the
Division of Research.

TABLE OF CONTENTS

Introduction

In an enormous variety of business research problems one is naturally interested in analyzing relationships between two or more variables. If data on such variables are available, one has a multivariate data set, and statistical analysis provides many tools for analyzing such data. Perhaps the most important analytic tool is multiple regression analysis, but it is often difficult to comprehend the full power of multiple regression in a conventional discussion of the subject. One is usually presented with a situation in which a regression problem has already been formulated--the dependent and explanatory variables having been selected--and one merely uses the techniques that are introduced to perform the necessary final calculations. Such simplifications obscure the important fact that much of statistical analysis is a multistage process of trial and error, that a good deal of exploratory work must be done to select appropriate variables for study and to determine relationships between them, and that a variety of statistical tests and other procedures must be performed and sound judgments made before one arrives at a satisfactory choice of dependent and explanatory variables.

In these notes we place much emphasis on using multiple regression analysis in conjunction with graphical techniques, on methods of selecting variables and constructing new variables by means of transformations, and on the use of statistical tests as a guide in model building. And although each of these methods makes a unique contribution to the analysis, the combination of them, used in an interactive mode, provides a powerful means of exploring, analyzing, and summarizing useful relationships in the data.

Access to a computer is, of course, essential for effective application of these concepts. If appropriate programs are readily available, the computer can execute the necessary graphical and statistical analyses quickly and inexpensively. With its storage devices the computer can provide convenient access to the large data sets often encountered in business research and can also retain new variables constructed in the course of an investigation. Moreover, the computer relieves the researcher of the need to involve himself personally in the internal algebraic complexities of the methods used, so that he can concentrate on the applied aspects of his investigation.

How can one learn to use these methods effectively? As in the use of many other statistical concepts, one develops an understanding of important technical aspects of each method and one practices and gains experience by applying these methods in realistic and challenging situations, using data of the kind encountered on the job and, of course, using appropriate computer programs.

Experience has shown that each of these methods has some characteristics which the user must keep in mind and a multitude of additional properties which, although interesting to the statistician, may not be particularly useful to the business researcher. In studying these methods it will therefore be appropriate for us to focus our attention on their especially useful characteristics and, whenever possible, to minimize discussion of mathematically interesting but inessential details.

Several of the sections that follow describe the basic methods of statistical data analysis that will be especially useful to us: multiple regression, graphical analysis of residuals or errors, methods of selecting variables, transformation of variables, and testing of hypotheses. Our discussion, although concise, comprehends many important features of these concepts and we utilize throughout an illustrative example which is typical of a variety of problems encountered in business research.

Multiple Regression Analysis

Nearly everyone interested in business research has seen an application of multiple regression analysis and knows some of its properties. Nevertheless, a brief review in the context of a problem may be useful. Suppose we are interested in studying absenteeism among employees of the ABX Company. The data in Table 1(a) provide three characteristics of a sample of 77 production employees in the company. The first column shows the number of occasions of absenteeism during 1975 for each of these employees. We will regard the dependent variable in this analysis to be absenteeism, as seems natural, and denote it as Y. Job complexity, denoted as as X_1, and employee seniority, denoted as X_2, are regarded as explanatory variables. X_1 is an index ranging from 0 to 100 and measures the complexity of the activities making up the job; X_2 is the number of complete years that the employee has been with the company (see Table 1(b)).

The regression equation relating absenteeism of employees to level of job complexity and seniority for the data in Table 1(a) is found by an appropriate computer program to be

Table 1(a). Absenteeism, Job Complexity, and Seniority of 77 Employees

Case Number	Absenteeism, Y	Job Complexity, X_1	Seniority, X_2	Case Number	Absenteeism, Y	Job Complexity, X_1	Seniority, X_2	Case Number	Absenteeism, Y	Job Complexity, X_1	Seniority, X_2
1	0	45	3	26	3	89	18	51	0	8	3
2	1	76	10	27	3	21	2	52	1	45	2
3	0	56	9	28	0	34	4	53	3	43	5
4	2	76	7	29	2	12	6	54	6	23	1
5	0	70	14	30	3	70	2	55	3	1	7
6	1	69	9	31	1	69	11	56	2	82	1
7	1	56	3	32	4	13	1	57	2	1	1
8	1	56	1	33	2	30	13	58	4	1	1
9	2	43	9	34	1	43	1	59	3	70	4
10	1	76	1	35	3	8	2	60	0	76	6
11	3	30	1	36	2	69	2	61	0	82	7
12	2	50	9	37	4	30	1	62	1	50	9
13	1	10	1	38	4	23	1	63	1	70	8
14	3	69	4	39	4	16	1	64	1	81	5
15	2	67	3	40	3	11	1	65	2	70	9
16	0	69	4	41	2	16	1	66	3	1	2
17	4	70	8	42	6	50	2	67	2	8	1
18	7	13	1	43	3	50	2	68	2	23	2
19	3	16	3	44	1	69	4	69	2	21	12
20	2	52	5	45	2	10	2	70	2	82	7
21	2	52	16	46	1	43	26	71	1	67	28
22	4	3	2	47	1	12	1	72	0	81	18
23	2	6	4	48	3	76	5	73	1	43	6
24	0	67	6	49	2	56	2	74	4	6	3
25	3	10	1	50	0	6	8	75	3	13	8
								76	2	52	7
								77	3	52	1

SOURCE: Computer simulation by one of the authors.

Table 1(b). Description of Variables for Which Data Are Shown in Table 1(a)

Variable Name	Description
Absenteeism	The number of distinct occasions that the worker was absent during 1975. Each occasion consists of one or more consecutive days of absence.
Job Complexity	An index ranging from zero to one hundred, measured according to procedures developed by Turner and Lawrence.*
Seniority	Number of complete years with the company on December 31, 1975.

*Arthur N. Turner and Paul R. Lawrence, Industrial Jobs and the Worker (Boston: Harvard University Press, 1965).

(1) $\hat{Y} = 3.07 - .015X_1 - .063X_2.$

This equation can be interpreted as providing an estimate of mean absenteeism for a given level of job complexity and seniority. Moreover, if seniority is held fixed, the equation shows that mean absenteeism tends to fall by .015 for each unit increase in job complexity. Also, if job complexity is held fixed, it shows that mean absenteeism tends to fall by 0.63 for each unit increase in seniority. It is clear that such information provides a useful summary of the data.

How does the computer determine a regression equation of the form (1)? First, let us consider a more general equation of which (1) is a special case,

(2) $$\hat{Y} = b_0 + b_1X_1 + b_2X_2.$$

The i^{th} of the n observations (n = 77 in the example) can be denoted as Y_i, X_{i1}, X_{i2}, i = 1,...,n. The <u>residual</u> e_i of the i^{th} observation is defined as

(3a) $$e_i = Y_i - \hat{Y}_i$$

(3b) $$= Y_i - (b_0 + b_1X_{i1} + b_2X_{i2}).$$

Table 2 illustrates residuals determined by equation (1).

Table 2. Construction of Selected Residuals from the Regression Equation $\hat{Y} = 3.07 - .015X_1 - .063X_2$

Case	X_1	X_2	Y	$\hat{Y}$	e
1	45	3	0	2.21	-2.21
2	76	10	1	1.30	-0.30
3	56	9	0	1.66	-1.66
4	76	7	2	1.49	0.51
5	70	14	0	1.14	-1.14
6	69	9	1	1.47	-0.47
7	56	3	1	2.04	-1.04
8	56	1	1	2.17	-1.17
9	43	9	2	1.86	0.14
10	76	1	1	1.87	-0.87
⋮	⋮	⋮	⋮	⋮	⋮
76	52	7	2	1.85	0.15
77	52	3	1	2.23	-1.23

The regression coefficients b_0, b_1, and b_2 are chosen so as to minimize the sum of squares of these residuals, denoted SSE,

$$(4) \qquad SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2.$$

The residuals (3a) and the sum of their squares (4) are extremely important entities in regression analysis. They are the basis for the well-known least squares criterion and in this context they are considered to be functions of the regression coefficients b_0, b_1, and b_2. When the values of these coefficients have been determined from the data, both the residuals e_i and SSE become fixed numbers and can be used to supplement the regression equation in summarizing the data. Various steps will be introduced below to make both the residuals and their sums of squares more meaningful for this purpose.

The sum of squared residuals is the basis of two important statistical measures of the discrepancy between the observed values of Y and the corresponding values of $\hat{Y}$ determined by the fitted regression equation: the standard error of the regression equation and the multiple correlation coefficient.

The standard error of the linear regression equation, which we denote by s, is the square root of the sum of squared residuals adjusted appropriately (or divided) by the number of degrees of freedom. The latter is calculated by subtracting the number of regression coefficients in (1) from n, the total number of observations. We thus write

$$(5) \qquad s = \sqrt{\frac{SSE}{n-3}} = \sqrt{\frac{\sum e_i^2}{n-3}} = \sqrt{\frac{\sum (Y_i - \hat{Y}_i)^2}{n-3}}$$

and s equals 1.36 in the absenteeism example. Under appropriate conditions, about 95 percent of the residuals will be smaller than 2s in absolute value, which means that for the corresponding observations Y will be within 2s of $\hat{Y}$.

The multiple correlation coefficient R can be thought of as the square root of the "fraction of the variation in Y (from its mean) explained by the linear regression equation." It is obtained by comparing the sum of squared residuals SSE to another quantity, the total sum of squares, SST. The total sum of squares is the sum of the squared deviations of Y about the mean $\overline{Y}$,

$$(6) \qquad SST = \sum_{i=1}^{n} (Y_i - \overline{Y})^2 .$$

The multiple correlation coefficient can be defined by the equation

$$(7) \qquad R = \sqrt{\frac{SST - SSE}{SST}} .$$

When there is only one explanatory variable in the regression analysis, we call this simple regression and the corresponding correlation coefficient is called a simple correlation. In this case the correlation coefficient is given the sign of the regression coefficient of the explanatory variable.

It is useful to observe that the multiple correlation coefficient of Y with X_1 and X_2 is equal to the simple correlation coefficient of Y with $\hat{Y}$.

These ideas can be extended easily to the more general case of regression analysis involving p variables, $X_1, X_2, \ldots, X_{p-1}$, and Y. The regression equation then has the form

$$(8) \qquad \hat{Y} = b_0 + b_1X_1 + \ldots + b_{p-1}X_{p-1}.$$

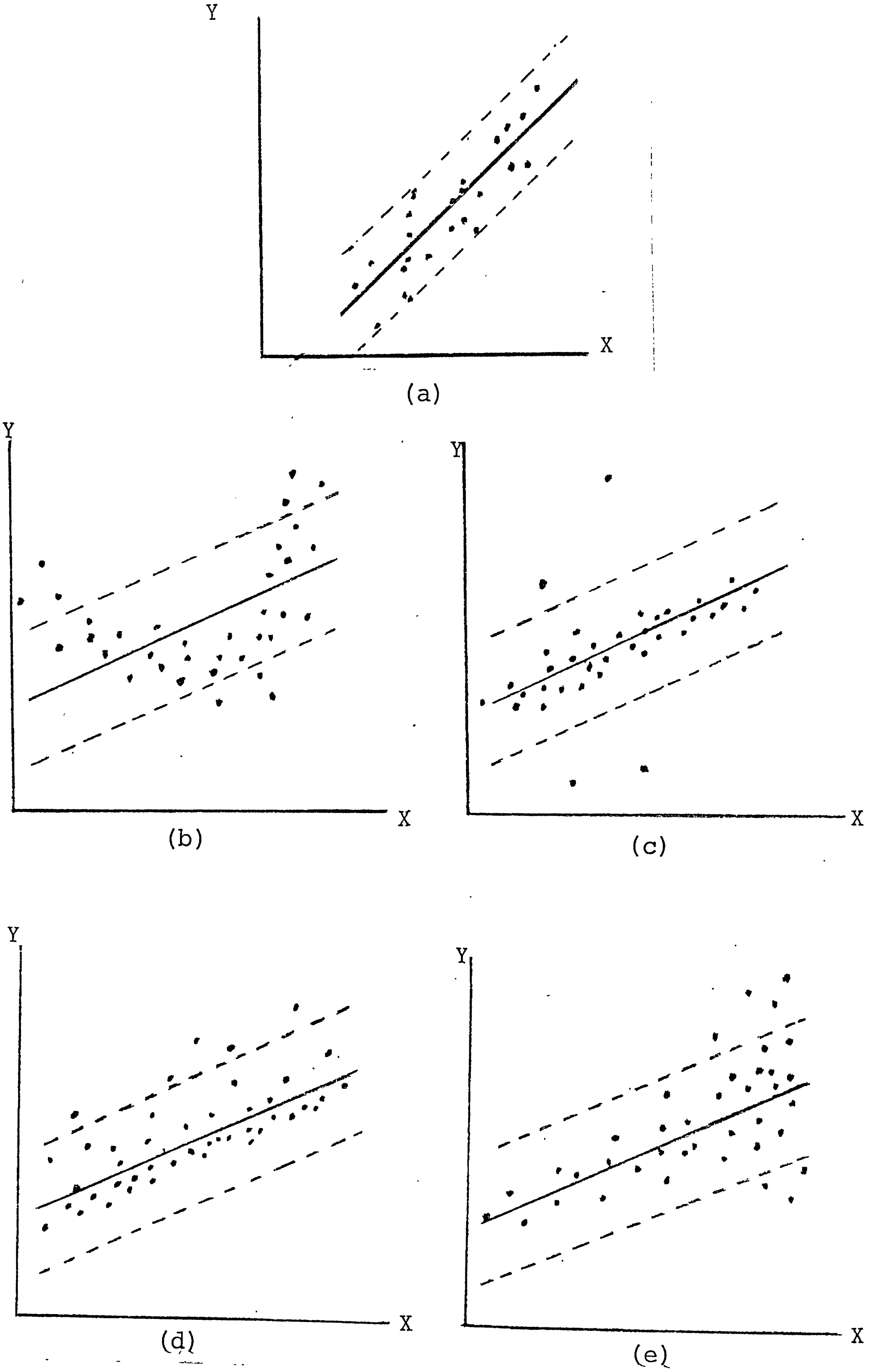
The residuals are defined by (3a) and form the basis for the least square criterion as well as for definitions of the standard error of the regression equation and the multiple correlation coefficient. The only modification required in equations (4) through (7) is appropriate adjustment of the number of degrees of freedom in the definition of s. Because there are p coefficients in the linear regression equation (8), the number of degrees of freedom is n - p so that

$$s = \sqrt{\frac{SSE}{n-p}} \quad . \tag{9}$$

Graphical Analysis of Residuals

Graphical examination of the data provides an effective means of assuring that the regression equation and its standard error and correlation coefficient summarize the data adequately. Without the use of graphs one runs a risk of being seriously misled by a linear regression analysis. Indeed, graphical examination of the relationship between the data and the regression equation should be regarded as essential and integral to regression analysis.

In the case of simple regression involving two variables X and Y, the scatterplot is the central tool. In addition to plotting the sample data points (X,Y), it is helpful to graph the regression line $\hat{Y} = a + bX$ as well as the error bands determined by the two lines $\hat{Y} = a + bX \pm 2s$ as is shown in Figure 1(a), where the linear regression seems appropriate for the data.



(a)

(b)

(c)

(d)

(e)

Fig. 1(a) - 1(e). Various Illustrative Scatterplots

Recall that in linear regression we assume that for any given value of X, the observed errors are a sample from a normally distributed (conditional) population with mean or expected value 0 and with variance σ^2, which does not vary with the given value of X. This implies that the error band $\hat{Y} \pm 2s$ should contain approximately 95 percent of all the data points and that for any given X approximately 95 percent of the corresponding points are also within the band.

With this background in mind, it is easy to see how the situations illustrated in Figures 1(b) - 1(e), which are of the kind commonly found in practice, render the linear regression equation or its standard error misleading. Figure 1(b) shows a situation in which it appears that the variables X and Y have a nonlinear relationship. Moreover, the mean of the errors does not appear to be zero for this sample; the mean (of the conditional distribution) appears to be positive for X values at the extremes of the data and the mean appears to be negative in the middle of the data. Figure 1(c) indicates the existence of outliers and these suggest that the population is not normal for the corresponding values of X. In Figure 1(d) the scatterplot suggests that the distribution of errors is skewed, and Figure 1(e), it seems reasonable to believe, shows that the variances of Y for larger values of X are not equal to those for smaller values of X. Any of these situations calls for caution in interpreting the regression analysis. In serious cases appropriate corrective steps (some of which will be discussed below) should be taken.

In the case of multiple regression it is more difficult to obtain effective graphical displays of the data. When three or more variables are involved in the analysis, several two-dimensional scatterplots are required to depict adequately the relationship between the data and the regression equation. One useful scatterplot is obtained by plotting the predicted values $\hat{Y}$ and the actual values Y, i.e., the points $(\hat{Y},Y)$ as in Figure 2.* One should include on the scatterplot a graph of the line $\hat{Y} = Y$ and the error bands $Y = \hat{Y} \pm 2s$. Any point on the line corresponds to a situation in which $\hat{Y} = Y$; the vertical distance from the line to any point represents an error $e = Y - \hat{Y}$.

In addition to plots of Y against $\hat{Y}$, it is important to examine scatterplots describing the

*On examination, Figure 2 indicates a possible problem in using linear regression with the absenteeism data. The errors are not a sample from a normal distribution. This occurs because the dependent variable Y is constrained to be integer valued. This situation, not uncommon in practice, causes little difficulty in the present example.

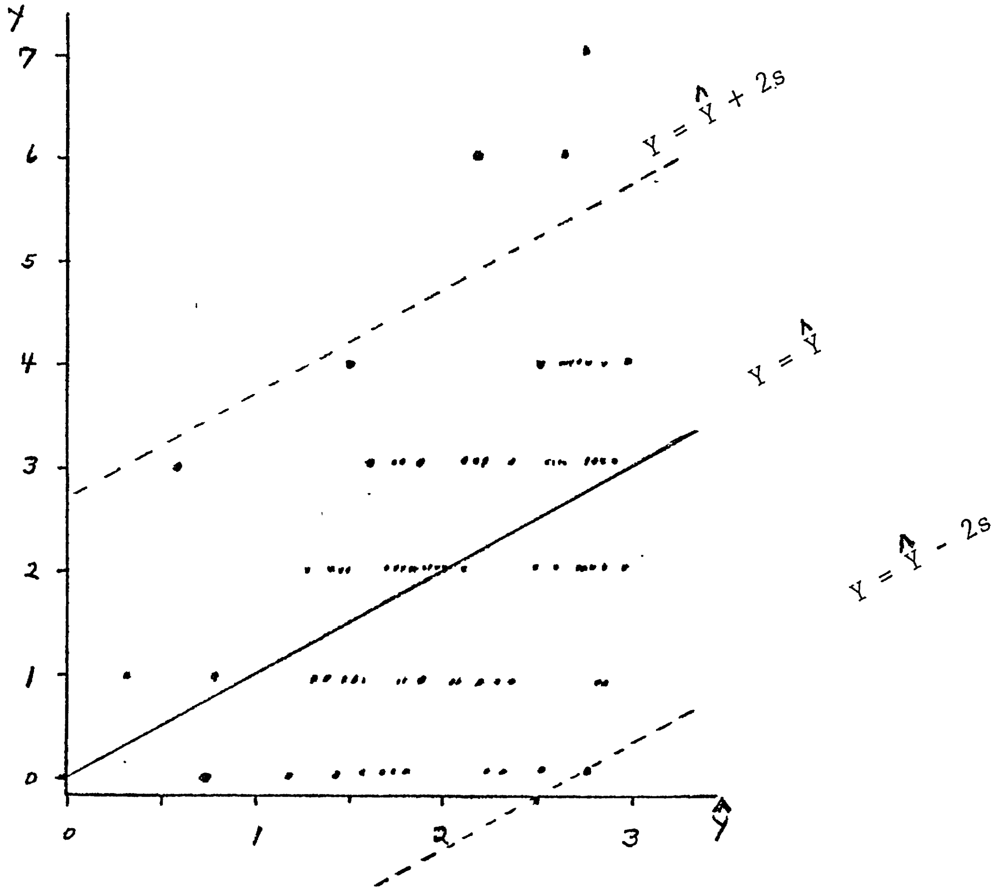


Fig. 2. Scatterplot of $(\hat{Y},Y)$ Using Regression Equation (1) and the Data of Table 1(a)

relationship between the residuals e and each of the explanatory variables X_j as illustrated in Figure 3. Here the points (X_j, e) are plotted as well as the line $e = 0$ and the band $e = \pm 2s$.

Sometimes these scatterplots are difficult to interpret and one can be assisted by a simple but infrequently used device called a <u>component scatterplot</u>. For any explanatory variable X_j included in the multiple regression equation, let $\hat{Y}_j$ denote $b_j X_j$; then the regression can be rewritten as

$$\hat{Y} = b_0 + \hat{Y}_1 + \ldots + \hat{Y}_{p-1} \tag{10}$$

and the actual values of the dependent variable can be expressed with an error term as

$$Y = b_0 + \hat{Y}_1 + \ldots + \hat{Y}_{p-1} + e. \tag{11}$$

The variables $\hat{Y}_1, \ldots, \hat{Y}_{p-1}$ are called the <u>systematic components</u> of the dependent variable Y. The component scatterplot associated with an explanatory variable X_j is then obtained by plotting the points $(X_j, \hat{Y}_j + e)$, and it is also helpful to graph the line $\hat{Y}_j = b_j X_j$ and the error band $\hat{Y}_j = b_j X_j \pm 2s$. These are illustrated in Figure 4 for the explanatory variable X_1.

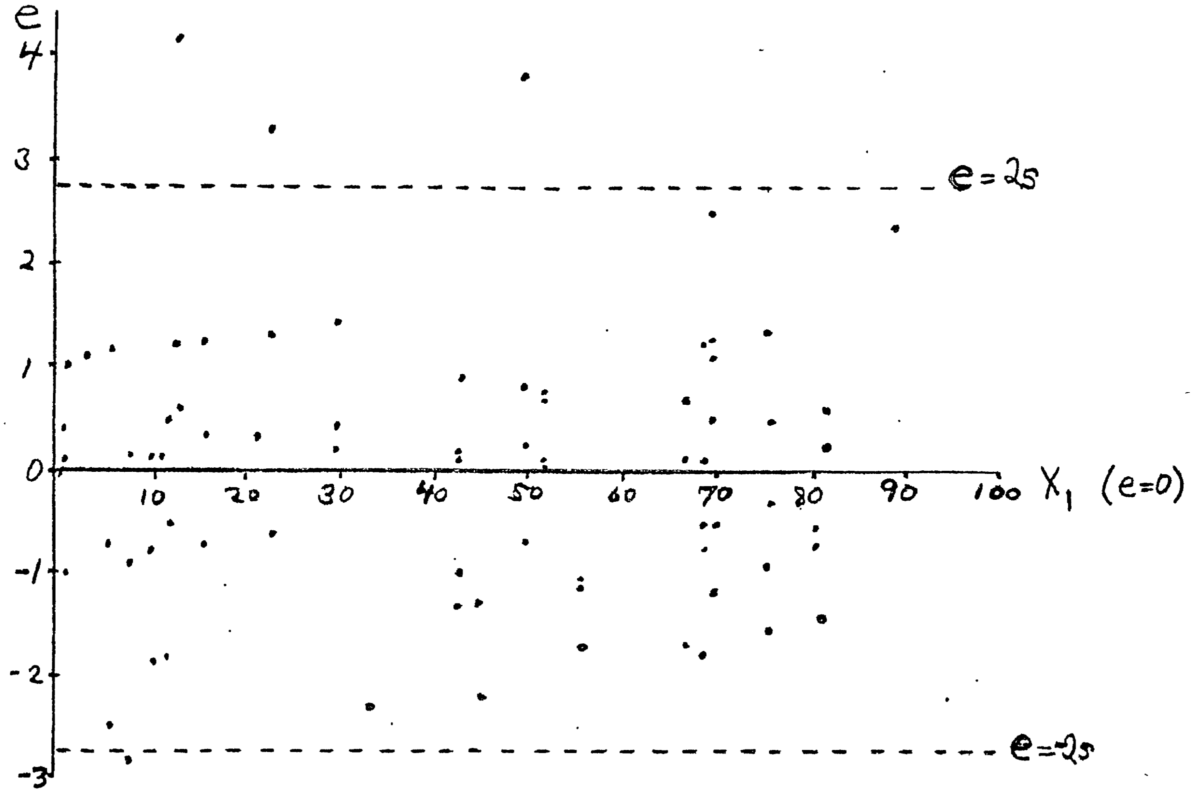


Fig. 3. Scatterplot of the Regression Error e With Job Complexity, X_1

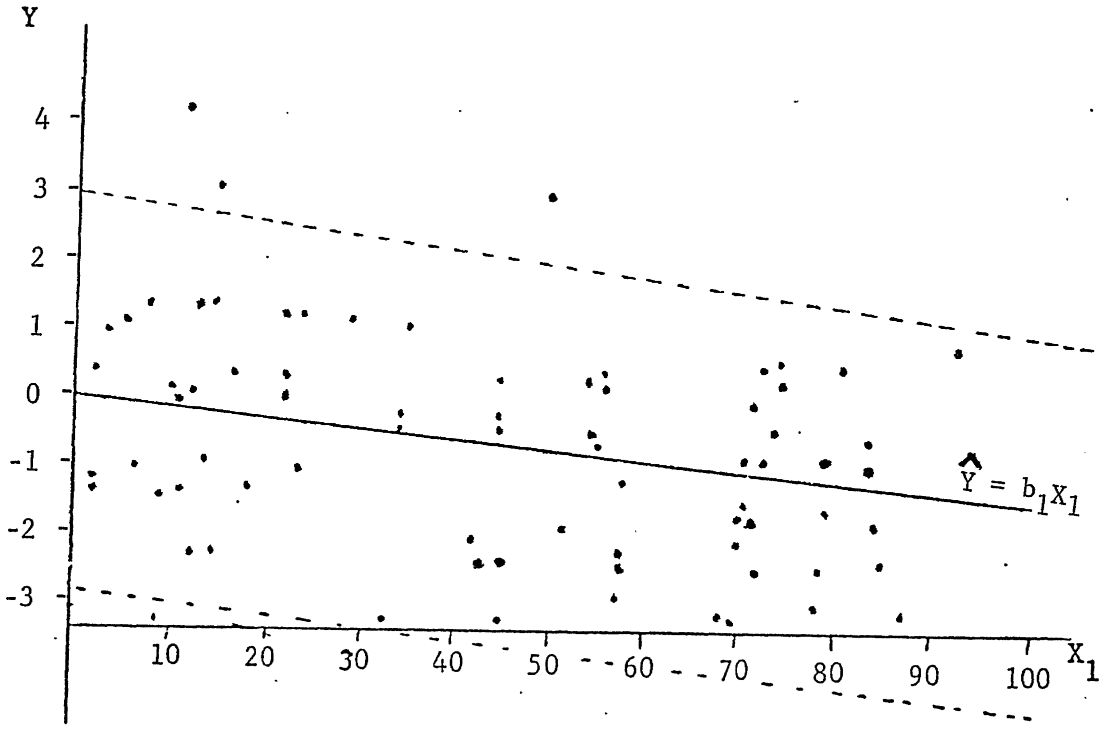


Fig. 4. Component Scatterplot of Y, Absenteeism With Job Complexity, X_1, for the Regression Equation (1)

In a component scatterplot we are assuming that $b_0 = 0$ and that $\hat{Y}_i = 0$ for $i \neq j$ in (11); this enables us to examine the relationship between the dependent variable $Y = \hat{Y}_j + e$ and X_j in temporary isolation. One can then use these scatterplots to detect problems such as those discussed in relation to Figure 1(b) through 1(e). For example, a component scatterplot which has an appearance like that of Figure 1(b) may indicate a non-linear association between the variable Y and the explanatory variable X_j, and a component scatterplot like that of Figure 1(e) would suggest that the error variances for various values of X_j are not equal.

When the existence of such problems is detected, it may be possible to correct for them by means of appropriate transformations of variables, as discussed below, or by using weighted regression. It is important to remember that none of the statistics provided by the usual tabular regression output are indicators for the presence of these kinds of problems and that graphical examination of the sort discussed above is essential.

Selecting Explanatory Variables

In developing a regression analysis of cross-section data, the dependent variable is often not difficult to choose. In many cases it is identified early in the study and usually remains fixed throughout the analysis. Although in a complex investigation there may be several dependent variables of interest, each is usually analyzed separately and serves as the center of attention in its own part of the project.

The problem of selecting explanatory variables is more difficult and typically occupies the researcher throughout most of his project. In the formulation stage of the typical project, the researcher builds up his personal understanding of the problem under study by reviewing relevant literature, discussing the problem with experienced people, extending his own direct observation of it, and assimilating underlying theory. Even at this early stage he is seeking factors that may help explain important features of the problem. Out of this work he chooses the dependent variable and a set of candidates for explanatory variables. An important part of this process is the investigation and development of appropriate measurement techniques. At the conclusion of this stage a sample of individual entities is often selected and a measurement of the dependent variable and each of the candidates for explanatory variables is obtained for each individual in the sample. These data are then transcribed and stored in the computer files which comprise the data base for the study. Appendix A presents a small data base of this type for the absenteeism study in which data on other candidate explanatory variables besides job complexity and seniority are contained.

Working with such data, the researcher uses statistical methods, especially regression analysis, to investigate relationships between the dependent variable and the candidate explanatory variables. He may begin by inspecting a table of simple correlation coefficients which measure the association between each pair of the variables (Table 3). He may then wish to obtain and evaluate a large number of multiple regression equations and scatterplots involving the dependent variable and various subsets of the candidate explanatory variables. In most cases it will be impractical to examine all possible multiple regression equations, so his selection must utilize his understanding of the problem at hand in addition to various statistical aids.

Typically, one of the researcher's goals is to select a regression equation which yields a high multiple correlation coefficient but which utilizes only a few, carefully chosen, and well-understood explanatory variables. He wants a high multiple correlation coefficient, because this measures the association between the dependent variable Y and the corresponding variable

Table 3. Correlation Matrix for Absenteeism Study

	Absenteeism	Job Complexity	Base Pay	Foreman Satisfaction	Seniority	Age	Number of Dependents
Absenteeism	1.00	-.36	-.23	-.19	-.34	-.31	-.05
Job Complexity	-.36	1.00	.50	-.25	.37	.28	-.08
Base Pay	-.23	.50	1.00	-.02	.49	.33	.06
Foreman Satisfaction	-.19	-.25	-.02	1.00	-.01	.20	.16
Seniority	-.34	.37	.49	-.01	1.00	.75	.15
Age	-.31	.28	.33	.20	.75	1.00	.15
Number of Dependents	-.05	-.08	.06	.16	.15	.15	1.00

$\hat{Y}$ determined by the regression (or between the dependent variable and the explanatory variables jointly). Roughly speaking, the higher the multiple regression coefficient, the more useful is the regression equation. On the other hand, the researcher needs to limit the number of explanatory variables chosen from among the candidates, because including too many variables in the regression equation complicates interpretation and application of the regression equation, reduces its statistical reliability, and, of course, increases the cost of data collection and manipulation.

A useful guide in selecting explanatory variables from the candidates is their partial correlation coefficients with the dependent variable. To define this statistic, suppose that the regression equation relating Y to the explanatory variables $X_1, X_2, \ldots, X_{p-1}$ has a multiple correlation coefficient R_1 and that a second regression equation relating Y to these same variables together with the additional variable X_p has a multiple correlation coefficient R_2. The partial correlation coefficient of X_p with Y, adjusted for the variables $X_1, X_2, \ldots, X_{p-1}$ is defined to be

$$\pm \sqrt{\frac{R_2^2 - R_1^2}{1 - R_1^2}} \,,$$

with the sign taken to be the same as the sign of b_p in the second regression equation. In selecting variables the absolute value of the partial correlation coefficient is used--the greater is the partial correlation coefficient in absolute value, the greater is the increase in the multiple correlation coefficient obtained by appending X_p to the explanatory variables already included in the regression equation.

Partial correlation coefficients can also provide insights into the effect of removing one of the explanatory variables already in a multiple regression equation. Most regression computer programs include as part of their output the partial correlation coefficient of each individual variable with the dependent variable, adjusted for all other explanatory variables included in the regression equation. The variable whose removal from the regression equation will cause the smallest decrease in the multiple regression coefficient is the one having the smallest (absolute) partial correlation coefficient.

There are several computer programs for automatically selecting explanatory variables from among specified candidates using partial correlation coefficients calculated from the data base of the study. Three that are commonly used are forward selection, stepwise regression, and backward elimination.

The forward selection program proceeds in an iterative fashion by first selecting a single explanatory variable from those specified by the researcher, then appending a second explanatory variable, and so on. The first variable that is selected is the candidate having the greatest (absolute) simple correlation coefficient with the dependent variable. In each subsequent step the forward selection program chooses for inclusion the candidate variable having the greatest (absolute) partial correlation coefficient with the dependent variable, adjusted for the other explanatory variables already selected. Selection continues as long as a variable can be found having a sufficiently large partial correlation coefficient.

Stepwise regression proceeds in a manner similar to that of the forward selection procedure but with one important difference. At each step beyond the first, after the new variable is appended, the program reexamines all explanatory variables currently in the regression equation to determine if any can be removed without unduly decreasing the value of the multiple regression coefficient. This is accomplished by evaluating the partial correlation coefficient of each included explanatory variable with the dependent variable, adjusted for all other explanatory variables included in the equation. The explanatory variable with the smallest (absolute) partial correlation coefficient is removed, provided its partial correlation coefficient is sufficiently close to zero. Consequently, the final result of stepwise regression is less dependent upon the early steps of the process than is the case with the forward selection program.

The other principal selection process is backward elimination. This procedure begins with a multiple regression equation which includes all of the candidate explanatory variables and then removes explanatory variables one at a time using the same criterion as that applied in the deletion stage in stepwise regression. Although there seems to be wide agreement that stepwise regression is preferred over forward selection, there is no general rule to determine whether stepwise regression or backward elimination is preferable. Many researchers often try both, compare the results, and engage in further experimentation when the differ.

There is general agreement that no automatic selection procedure should be used uncritically. None of these procedures will always arrive at the best selection from among candidate variables in terms of the highest possible multiple correlation coefficient for a fixed number of candidate variables. More importantly, they may fail to identify the regression equation most consistent with the researcher's understanding of the problem. But, used with care, scepticism, and willingness to experiment further, they can be effective tools. In any case they are no replacement for the graphical methods of examination described in the preceeding section, which should be used in conjunction with them.

Transforming Variables

One way of greatly extending the capabilities of linear regression analysis is to make use of nonlinear transformations of variables. A strategy for choosing such transformations is evident from a simple example. Consider the data in the scatterplot of Figure 5; no straight line given by the equation

$$\hat{Y} = b_0' + b_1'X \tag{12}$$

can summarize these data adequately. The scatterplot suggests that as X increases, Y tends to increase but at a diminishing rate. Among the many equations which

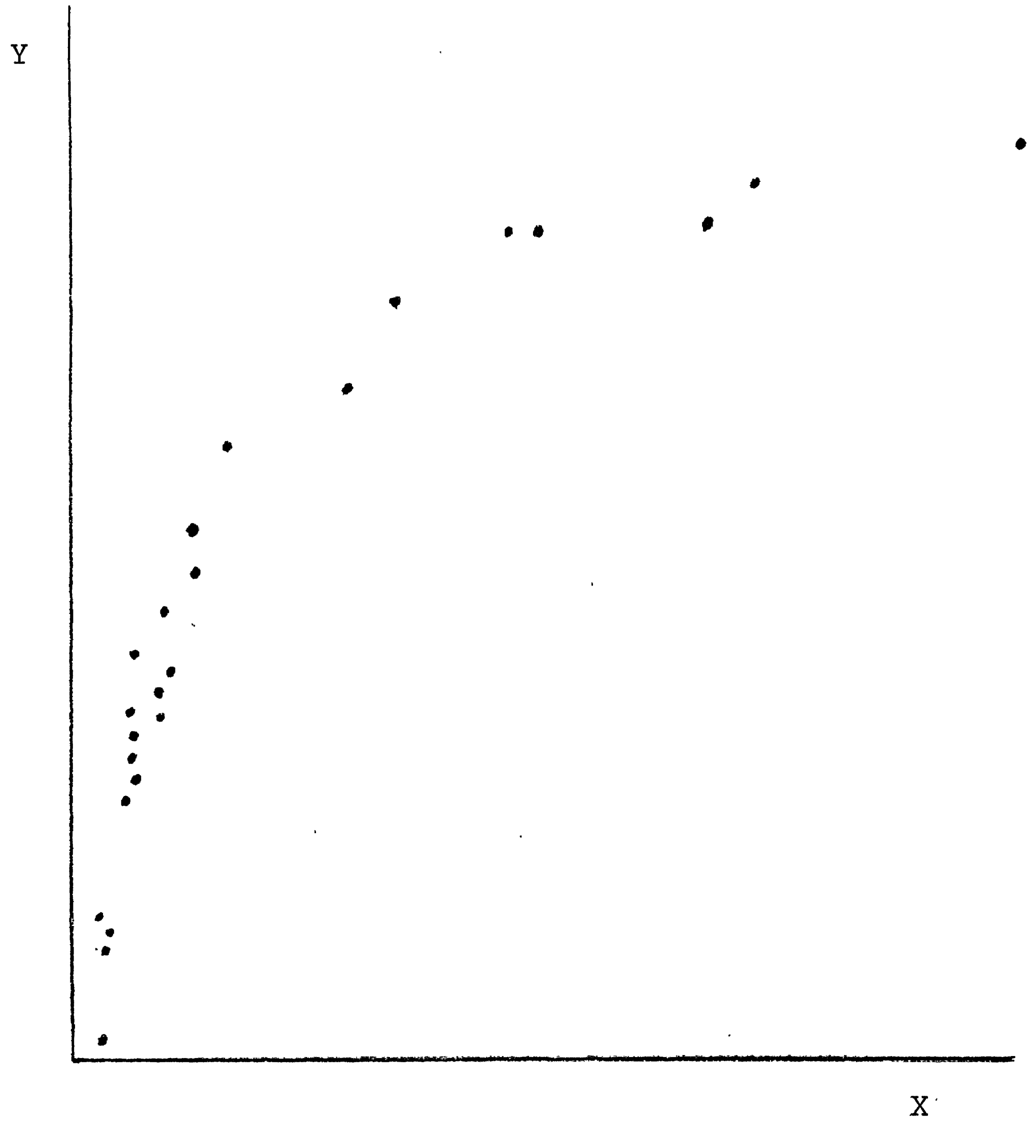


Fig. 5. A Scatterplot of an Apparently Nonlinear Relationship

summarize data having this property, some of the simplest are of the form

$$(13) \qquad \hat{Y} = b_0 + b_1(1/X).$$

How can one choose b_0 and b_1 in (13)? The approach developed earlier for handling equation (12) can be applied to (13) with very little modification. It is natural to define the sum of squared errors of (13) to be

$$\text{SSE} = \sum_{i=1}^{n} (Y_i - b_0 - b_1/X_i)^2$$

and to require that b_0 and b_1 minimize SSE as before.

It might appear that a new computational procedure is needed to find the values of b_0 and b_1, but it turns out that this is not so. All that is necessary is to proceed with a new variable X* given by

$$(14) \qquad X^* = 1/X.$$

If SSE is rewritten in terms of X* as

$$\text{SSE} = \sum_{i=1}^{n} (Y_i - b_0 - b_1X_i^*)^2,$$

it is clear that the desired values of b_0 and b_1 can be found by obtaining the linear regression equation relating Y to X*,

$$\hat{Y} = b_0 + b_1X^*.$$

One can therefore fit a nonlinear equation to the data following the least squares criterion by first transforming the variable X to the new variable X* using (14) and then developing the ordinary linear regression equation relating Y to X*. Operationally, the actual values of X* can be either computed and stored as values of a new variable in the data base or they can be temporarily computed when needed by the regression program. In the first case, the researcher utilizes a separate program for computing the transformed variables required and then uses a standard multiple regression program with these new variables. In the second case, the researcher uses a single program incorporating transformations and regression analysis; he specifies the variables to be included in the regression equation and any preliminary transformation of these variables that is required. In either mode of operation, many transformations in addition to the reciprocal transformation (14) can be employed.

Three simple classes of transformations are commonly used in business research:

(a) reciprocals: $X^* = 1/X$,

(b) powers: $X_1^* = X$, $X_2^* = X^2$, $X_3^* = X^3$, etc.

(c) logarithms: $X^* = \log(X)$.

Figure 6 shows some of the equations that can be fitted by combining linear regression with these transformations. The reciprocal transformation has already been discussed. A polynomial

$$\hat{Y} = b_0 + b_1X + b_2X^2 + \ldots + b_{p-1}X^{p-1}$$

of any degree p-1 can be fitted by the use of power transformations together with multiple regression, although polynomials of degree two or three suffice in many cases. The logarithm transformation is often useful in dealing with a variable whose values are all greater than zero.

These transformations can also be applied to several different explanatory variables. For example, one could compute a regression equation of the form

$$(15) \qquad \hat{Y} = b_0 + b_1(1/X) + b_2X_2 + b_3X_2^2 \ .$$

The main difficulty is that graphs cannot be so easily prepared in these multivariate cases. Despite this difficulty, appropriate graphs are extremely

	Equation	Alternate Form
(1)	$\hat{Y} = b_0 + b_1(1/X)$ $(X \neq 0;\ \text{usually } X > 0)$	None
(2)	$\hat{Y} = b_0 + b_1X + b_2X^2$	None
(3)	$\hat{Y} = b_0 + b_1\log(X)$ $(X > 0)$	None
(4)	$\hat{Y} = a_0(a_1)^X$ $(a_0 > 0,\ a_1 > 0)$	$\log \hat{Y} = b_0 + b_1X$ $(b_j = \log a_j,\ j = 1,2)$
(5)	$\hat{Y} = aX^{b_1}$ $(X > 0,\ a > 0)$	$\log \hat{Y} = b_0 + b_1\log X$ $(b_0 = \log a)$

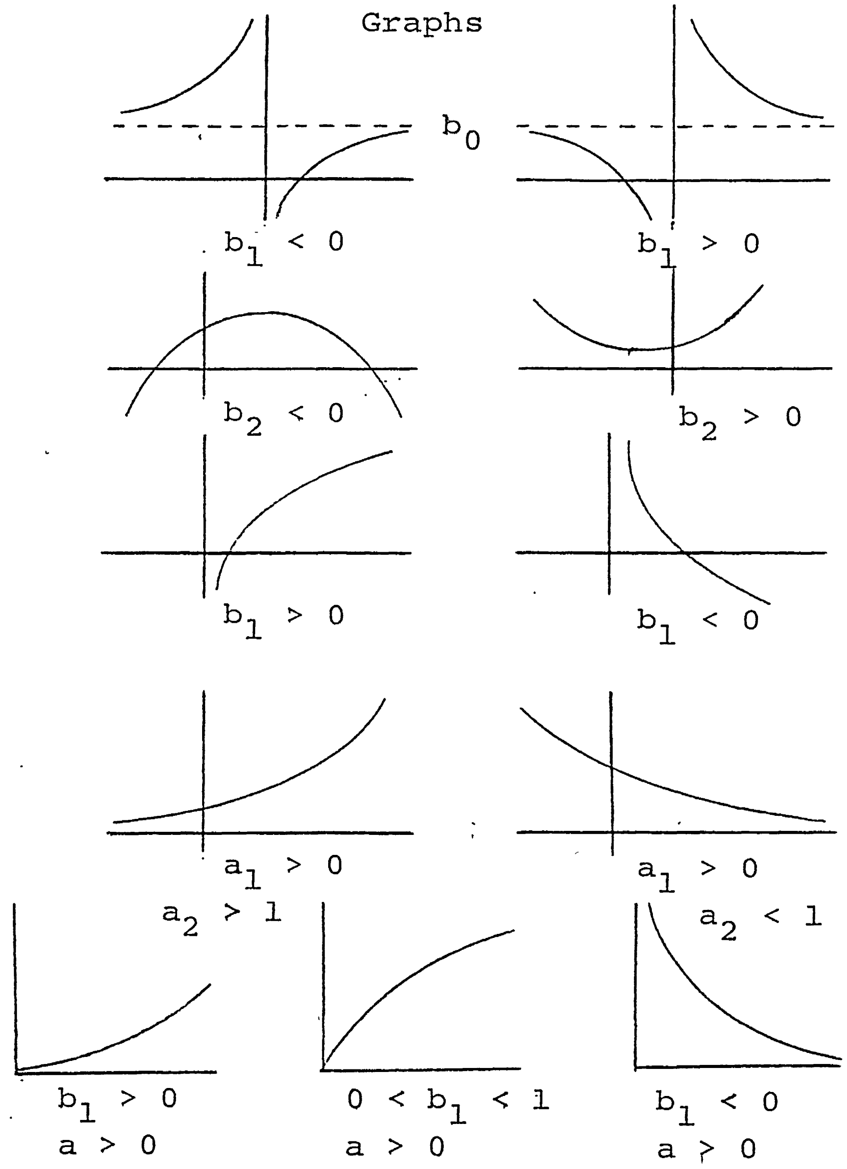


Fig. 6. Some Nonlinear Equations Obtainable Through Transformations

important in visualizing these multivariate, nonlinear regression equations and their relationships to data. Most of the techniques described earlier can be used with appropriate modifications, and component scatterplots are especially useful. For example, the component scatterplot associated with X_2 in (15) can be obtained by graphing the equation

$$\hat{Y}_2 = b_2X_2 + b_3X_2^2$$

as well as the points $(X_2, \hat{Y}_2 + e)$ where, as usual, the residual e is determined by $Y - \hat{Y}$.

Some of the equations shown in Figure 6 involve transformations of the dependent variables as well as of the explanatory variables. For example, the equation

$$\hat{Y} = a_0(a_1)^X$$

can be reformulated by taking natural logarithms of both sides of the equation to get

$$\log(\hat{Y}) = b_0 + b_1X,$$

where $b_j = \log(a_j)$ for $j = 0,1$. Thus b_0 and b_1 can be computed by finding the ordinary regression equation relating the transformed dependent variable log(Y) to X. Then a_0 and a_1 in the original equation are found by means of the antilogarithms of b_0 and b_1.

Care must be taken when a transformation of the dependent variable is utilized. In interpreting such an analysis it is best to return to the original equation rather than using the one containing the transformed dependent variable. In particular, the correlation coefficient associated with the transformed equation can be misleading, and it is better to compute directly the correlation coefficient of Y with $\hat{Y}$. It should also be recognized that the least squares criterion itself is altered by a nonlinear transformation of the dependent variable. One should carefully inspect the residuals of the transformed equation to determine the suitability of the regression analysis, as discussed previously in conjunction with Figure 1.

A particularly useful type of variable is an indicator variable, sometimes called a dummy variable. Any variable having exactly two values, zero and one, is called an indicator variable. Such a variable is used to record the presence or absence of a particular characteristic or condition of each observation. A simple example is the use of an indicator variable in the answer to a yes/no question, in which the integer one indicates yes and zero indicates no.

It is frequently useful in regression analysis to create an indicator variable by means of a transformation of some variable already included in the data base. For example, an indicator variable S_1 can be established from the variable Foreman Satisfaction, denoted S and indicative of a worker's satisfaction with his foreman, which is included in the absenteeism data base shown in Appendix A. The Foreman Satisfaction variable takes on values according to the following coding:

1 = very dissatisfied,

2 = somewhat dissatisfied,

3 = neither satisfied nor dissatisfied,

4 = fairly well satisfied,

5 = very satisfied.

An indicator variable S_1 can be developed which indicates whether or not an employee is very dissatisfied with his foreman. S_1 is made to take on the value one whenever S takes on the value one--the employee is very dissatisfied--and S_1 is given the value zero for all other values of S. Table 4 shows the variables S and S_1.

An indicator variable can, of course, be included among the explanatory variables in a multiple regression

Table 4. Illustration of Indicator Variables Based on the Explanatory Variable S, Foreman Satisfaction

Case	Values of Foreman Satisfaction	Value of Indicator Variables				
		S_1	S_2	S_3	S_4	S_5
1	4	0	0	0	1	0
2	4	0	0	0	1	0
3	1	1	0	0	0	0
4	3	0	0	1	0	0
5	3	0	0	1	0	0
6	3	0	0	1	0	0
7	4	0	0	0	1	0
8	4	0	0	0	1	0
9	1	1	0	0	0	0
10	3	0	0	1	0	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮
76	1	1	0	0	0	0
77	3	0	0	1	0	0

SOURCE: Appendix A

equation. Standard regression computer programs can handle an indicator variable in the same manner as any other kind of explanatory variable. For example, in using the absenteeism data one can find the regression equation relating absenteeism (A) to job complexity (C), seniority (SE), and the indicator variable S_1 introduced above, which indicates whether the employee is very dissatisfied with his foreman.

$$\hat{A} = 3.06 - .015C - .064SE + .220S_1.$$

Interpretation of the coefficient of the indicator variable is simple to explain. If we recall that S_1 takes on the values zero and one, and note that the coefficient of S_1 is .220, then whenever $S_1 = 1$, $\hat{A}$ from the equation above is larger by .220 than it is when $S_1 = 0$. Thus we can say that employees having the same level of job complexity and seniority and who are very dissatisfied with their foreman have an absenteeism that is .220 higher, on the average, than workers who are not very dissatisfied. It should be clear that in general the regression coefficient of an indicator variable represents the increment to $\hat{Y}$ associated with the characteristic or condition that the indicator variable represents.

It is often helpful to use several indicator variables to represent several mutually exclusive conditions or characteristics. For example, the variable foreman satisfaction, S, records the five mutually exclusive conditions: very dissatisfied, somewhat dissatisfied, etc. The indicator variable S_2 in Table 4 indicates whether or not an employee is somewhat dissatisfied; S_2 takes on the value one if the employee indicates that he is somewhat dissatisfied and the value zero for any other condition. The variables S_3, S_4, and S_5 are defined analogously. Note that only four indicator variables need be used, because the fifth condition, very satisfied, is implied whenever the four other indicator variables are each equal to zero.

When the first four indicator variables are included with job complexity and seniority as explanatory variables for absenteeism, we obtain the regression equation

$$(16)\quad \hat{A} = 3.048 - .017C - .042SE + .175S_1 + 1.063S_2 - .181S_3 - .462S_4$$

Assistance in interpreting this regression equation can be provided by calculating the values of $\hat{A}$ associated with the several foreman satisfaction conditions while job complexity and seniority are held fixed. Table 5 shows that the regression coefficients of the indicator

Table 5. Values of $\hat{A}$ from Equation (16) with C = 40 and SE = 5

Foreman Satisfaction, S	Conditional Mean Absenteeism, $\hat{A}$
1, Very Dissatisfied	2.333 = 2.158 + .175
2, Somewhat Dissatisfied	3.221 = 2.158 + 1.063
3, Neither Satisfied nor Dissatisfied	1.977 = 2.158 - .181
4, Fairly Well Satisfied	1.696 = 2.158 - .462
5, Very Satisfied	2.158

variables S_1, S_2, S_3, and S_4 represent the increments to $\hat{A}$ associated with the corresponding levels of foreman satisfaction, relative to the value of $\hat{A}$ associated with the fifth level of foreman satisfaction.

Some appreciation of the advantages of using indicator variables can be gained by comparing the regression equation (16) to one relating absenteeism to job complexity, seniority, and foreman satisfaction, S, itself,

$$\hat{A} = 4.435 - .019C - .055SE - .415S. \tag{17}$$

In (17) we see that $\hat{A}$ decreases by .415 whenever S increases by one. This suggests that the average absenteeism of workers who are somewhat dissatisfied is .415 less than that of workers who are very dissatisfied, and that a similar comparison of workers who are very satisfied with those who are fairly well satisfied leads to the same difference in average absenteeism. Although this situation may not be realistic, it is required by the direct use of the variable S in the regression equation (17). The use of indicator variables in equation (16), on the other hand, allows a more flexible representation of the changes in average absenteeism with increasing foreman satisfaction.

In selecting indicator variables for a multiple regression equation, care must be taken to avoid a situation called singularity or unidentifiability. If one attempts to find the regression equation which relates absenteeism to the seven explanatory variables C, SE, S_1, S_2, S_3, S_4, and S_5 by using a typical regression computer program, then one will get none of the usual output but only a cryptic comment such as "matrix singular" or "equation unidentified." The problem is that no unique regression equation is determined by the least squares criterion and there are in fact infinitely many different equations which fit the data equally well. This situation occurs whenever one of the explanatory variables can be written as a linear function of the other explanatory variables, and it is the case in our example, because we have

$$S_5 = 1 - S_1 - S_2 - S_3 - S_4.$$

The difficulty can be eliminated by excluding the redundant variable S_5 from the equation. In general, whenever a set of indicator variables is used to represent a set of mutually exclusive conditions or circumstances like degrees of foreman satisfaction, one of the logically possible indicator variables should be excluded from the regression equation.

An indicator variable can also be used as a dependent variable in a regression equation. Suppose it is company policy to review the performance of any employee having three or more occasions of absenteeism. Let A* denote an indicator variable that is one if absenteeism is three or larger and zero otherwise. The regression equation relating A* to job complexity and seniority can be found in the usual way,

$$\hat{A}^{*} = .645 - .041C - .019SE.$$

As usual, $\hat{A}^{*}$ is interpreted as the conditional mean of the dependent variable A*, but in this context A* takes on only the values 0 and 1, so its mean is equal to the conditional probability that A* takes on the value 1. Therefore $\hat{A}^{*}$ is interpreted as the conditional probability that an employee has three or more occasions of absenteeism, given his level of job complexity and seniority. In general, when the dependent variable of a regression equation is an indicator variable, then the regression equation is regarded as giving the conditional probability that the indicator variable is 1, given the levels of the explanatory variables.

Special care must be exercised in using an indicator variable as a dependent variable. Because the indicator variable takes on only the values zero and one, the errors of the regression equation cannot be a sample from a normal distribution. Moreover, it can be shown that the variance of the errors is not constant. The usual assumptions of linear regression analysis are therefore not satisfied. However, if the sample size is sufficiently large (greater than 30 for many applications), the nonnormality of the errors causes little difficulty. If, in addition, $\hat{A}^*$ is within the interval .2 to .8 for most of the observations, then the variance will be approximately constant and ordinary regression analysis is usually satisfactory. If either of these two conditions is violated, then the methods of logit or probit analysis can be utilized.

One additional method of expanding the types of equation that can be fitted to data using regression analysis should be discussed, namely, the use of interaction variables. An interaction variable is simply the product of two candidate explanatory variables. The role of an interaction variable can be seen by comparing the following two regression equations,

$$(18) \qquad \hat{A} = 3.07 - .015C - .063SE$$

and

(19) $\hat{A} = 3.41 - .023C - .159SE + .002C*SE.$

Equation (18) is said to be additive because $\hat{A}$ is expressed as the sum of two components, one depending on job complexity and the other depending on seniority. Equation (19) is said to be nonadditive because of the interaction variable C*SE. Table 6 shows illustrative values of $\hat{A}$ calculated from both the additive equation (18) and the nonadditive equation (19). In the additive equation the change in $\hat{A}$ as one moves from SE equal to 1 to SE equal to 5 is -.25, regardless of the value of C; this is illustrated for three values of C in Table 6(a). However, under the nonadditive equation the corresponding changes in $\hat{A}$ vary with the level of C (Table 6(b)). In general, the nonadditive equation defines a much more complex relationship between the dependent variable and the explanatory variables. Because of the complexity of nonadditive equations, it is much more difficult to obtain useful graphs of the data and its relationship to the regression equation.

All of the methods discussed in this section enhance the flexibility of regression analysis and usually require little added effort. It is very important to keep all these techniques in mind when analyzing data.

Table 6. Values of $\hat{A}$ for Given Values of C and SE

(a) As Determined by Equation (18)

SE	9	2.20	1.90	1.60
	5	2.46	2.16	1.86
	1	2.71	2.41	2.11
		20	40	60
			C	

(b) As Determined by Equation (19)

SE	9	1.88	1.78	1.68
	5	2.36	2.10	1.84
	1	2.83	2.41	1.99
		20	40	60
			C	

However, if skillfully applied, ordinary multiple regression analysis using candidate variables directly will often produce satisfactory results.

Hypothesis Testing in Data Analysis

We introduce the principal problem addressed in this section by an example based on the absenteeism data. Consider the following two regression equations, previously shown as equations (16) and (1), respectively:

$$(20)\quad \hat{A} = 3.048 - .017C - .042SE + .175S_1 + 1.063S_2 - .181S_3 - .462S_4$$

and

$$(21)\quad \hat{A} = 3.07 - .015C - .063SE.$$

A question of obvious practical importance arises: Is the first equation better, in some sense, than the second equation?

One way to approach this question is to compare the multiple correlation coefficients of the two regression equations, say R for (20) and R' for (21). If R is much larger than R', then the first of the equations would probably be preferred. In practice, however, it is often difficult to decide whether the difference between R and R' is sufficiently large to enable the researcher to choose between the equations. In this case, for example, R = .5522 and R' = .4214 and it is not clear whether the increase of R for (20) over R' for (21) warrants the inclusion of the four additional variables.

Before continuing our discussion it will be helpful to replace equations (20) and (21) with the following more general equations:

$$(22)\quad \hat{Y} = b_0 + b_1X_1 + \ldots + b_{p'-1}X_{p'-1} + b_{p'}X_{p'} + \ldots + b_{p-1}X_{p-1}$$

and

$$(23)\quad \hat{Y} = b_0' + b_1'X_1 + \ldots + b_{p'-1}'X_{p'-1}.$$

Here (22) is a regression equation involving p variables, and (23) is a regression equation involving p' variables. We call (22) the full equation and (23) the reduced version of (22). It is assumed that p > p' and that all of the explanatory variables of (23) are included in (22). For convenience we also assume that the first p'-1 explanatory variables included in (22) are the explanatory variables of (23). Thus (22) includes all the explanatory variables of (23) together with p - p' additional explanatory variables. Both of these regression equations are assumed to have been computed from data comprised of n observations of the variables. The multiple correlation coefficients corresponding to (22) and (23) will be denoted as R and R' and the sum of the squared residuals of these two equations as SSE and SSE', respectively.

In order to make further progress, it is necessary to give a careful statement of the experimental situation assumed to underlie the data. We assume that there is some "true" relationship or model between all the variables in the full equation (22) of the form

$$(24) \quad Y = \beta_0+\beta_1X_1+\ldots+\beta_{p'-1}X_{p'-1}+\beta_{p'}X_{p'}+\ldots+\beta_{p-1}X_{p-1}+\varepsilon$$

The $\beta_0,\ldots,\beta_{p-1}$ are unknown numbers called parameters and ε is a random variable which is normally distributed with mean zero and standard deviation σ. This implies that we view our data, which consist of n observations of the variables, as having been generated by an underlying process which advanced through the following steps:

(a) Either the researcher or a chance or deterministic mechanism selected the n values of the explanatory variables and these values are observed by the researcher.

(b) The chance mechanism selected a random sample of n values of ε, drawn independently from a normal distribution having mean zero and standard deviation σ. The researcher does not observe this sample.

(c) Then n values of Y were determined from (24) using true values of $\beta_0,\ldots,\beta_{p-1}$. These n values of Y are observed by the researcher but the true values of β_j are unknown to him.

The regression equation (22), computed from the data, is viewed as an estimate of the true relationship or model (24). In other words, the computer regression coefficients $b_0,\ldots,b_{p-1}$ are estimates of the corresponding parameters $\beta_0,\ldots,\beta_{p-1}$. These computed regression coefficients are determined by the underlying process; they would vary from sample to sample and are considered to be random variables having some probability distribution. Similarly, the observed multiple correlation coefficient R associated with (22) would vary from sample to sample and is also viewed as a random variable.

Using this formulation, our problem of choosing between the full equation (22) and the reduced version (23) can be approached by means of a statistical hypothesis test. If we can accept the hypothesis H_0 that each of the β_j corresponding to the explanatory variables in (22) which are not in (23) is equal to zero, then clearly (23) is the equation to be chosen. If we reject H_0, then (22) would be the preferred equation. Thus our null hypothesis H_0 is that the true coefficients $\beta_p,\ldots,\beta_{p-1}$ are all equal to zero and the alternate hypothesis H_1 is that at least one of these coefficients is nonzero.

The hypothesis H_0 can be tested by calculating the F-statistic using

$$F = \frac{\dfrac{R^2 - (R')^2}{p - p'}}{\dfrac{1 - R^2}{n - p}} \, .$$

It can be shown that H_0 implies that this statistic has a known probability distribution called the F-distribution, with p - p' and n - p degrees of freedom. Tables of this distribution are readily available. The hypothesis H_0 is rejected in favor of H_1 if the value of the F-statistic for our data is larger than some value determined by the F-distribution and the chosen significance level of the test. Thus, we would prefer (22) over (23) if and only if the value of the F-statistic for the sample is large enough to cause us to reject H_0 at the chosen level of significance.

It can also be shown that the value of the F-statistic may be calculated from the sum of the squared residuals of the two regression equations, SSE and SSE',

$$F = \frac{\dfrac{SSE' - SSE}{p - p'}}{\dfrac{SSE}{n - p}} \, .$$

We return to equations (20) and (21) to illustrate this procedure. Here R = .5522, R' = .4214, p' = 3, p = 7, and n = 77. We see that the two explanatory variables in (21) are the first two explanatory variables in (20). The hypothesis H_0 in this case is

$$H_0: \quad \beta_3 = \beta_4 = \beta_5 = \beta_6 = 0$$

and the alternate hypothesis H_1 is that at least one of the $\beta_3, \ldots, \beta_6$ is nonzero. Inspection of (20) indicates that $\beta_3, \ldots, \beta_6$ are the parameters corresponding to the four indicator variables $S_1, \ldots, S_4$. We choose the 5 percent level of significance for this test and calculate

$$F = \frac{\dfrac{R^2 - (R')^2}{p - p'}}{\dfrac{1 - R^2}{n - p}} = \frac{\dfrac{(.5522)^2 - (.4214)^2}{4}}{\dfrac{1 - (.5522)^2}{70}} = 3.21 .$$

According to tables of the F distribution, the value of F for p - p' = 4 and n - p = 70 degrees of freedom is 2.50. The F-statistic of 3.21 for the sample is greater than this value, so we reject H_0 and accept H_1. Thus we conclude that the difference between R and R' is statistically significant and that the full equation (20) is preferred over the reduced version (21).

It can be seen, however, that H_0 could be accepted at the more stringent 1 percent level of significance;

the difference between R and R' is not now statistically significant and one would prefer (21) over the full equation (20). As is often the case, the significance level chosen for a test has great influence on the decision that is subsequently made. A widely used level of significance is the 5 percent level.

This hypothesis testing procedure, called an F-test of the significance of additional explanatory variables (or an F-test for choosing between two regression equations of the type (22) and (23)), is more general that it first appears to be. A variety of other conventional F-tests can be placed in the framework above and treated as a problem involving a choice between two regression equations. For example, in fitting a polynomial to data, one can use our procedure to test the significance of one or more higher order terms. We can take the full equation and reduced version to be respectively

$$\hat{Y} = b_0 + b_1 X + b_2 X^2 + b_3 X^3 \tag{25}$$

and

$$\hat{Y} = b_0' + b_1' X. \tag{26}$$

The null hypothesis is H_0: $\beta_2 = \beta_3 = 0$ and the alternate hypothesis H_1 is that at least one of the parameters β_2 or β_3 is nonzero.

The significance of the coefficients of the entire set of explanatory variables included in a multiple regression equation can also be tested by using

(27) $\hat{Y} = b_0 + b_1X_1 + b_2X_2 + b_3X_3$

and

(28) $\hat{Y} = b_0'$

as our pair of equations. The null hypothesis in this case is H_0: $\beta_1 = \beta_2 = \beta_3 = 0$ and H_1 is the alternate hypothesis that at least one of these parameter values is nonzero. In calculating the F-statistic in this case, $p' = 1$ and $R' = 0$ and we have $p - 1$ and $n - p$ degrees of freedom.

One can also test the significance of the coefficient of one or more interaction variables; for example, suppose we have

(29) $\hat{Y} = b_0 + b_1X_1 + b_2X_2 + b_3X_1{*}X_2$

and

(30 $\hat{Y} = b_0' + b_1'X_1 + b_2'X_2$;

it is clear that we can test H_0: $\beta_3 = 0$ against H_1: $\beta_3 \neq 0$.

Finally, one can test the significance of the coefficient of any single explanatory variable in a regression equation as well. Suppose we have

(31) $\hat{Y} = b_0 + b_1X_1 + b_2X_2 + b_3X_3$

and

(32) $\hat{Y} = b_0' + b_1'X_1 + b_3'X_3$;

this enables us to test H_0: $\beta_2 = 0$ against H_1: $\beta_2 \neq 0$. In this instance we have $p = 4$, $p' = 3$ so that $p - p' = 1$ and we use the F-distribution for 1 and $n - p$ degrees of freedom.

This F-test of a single coefficient is the same as the t-test for the null hypothesis $\beta_2 = 0$ against the alternative hypothesis that $\beta_2 \neq 0$ because the F-statistic for 1 and $n - p$ degrees of freedom is the square of the t-statistic for $n - p$ degrees of freedom. Some regression computer programs give the value of the F-statistic and others the value of a t-statistic. One can make use of the relationship $F = t^2$, perform either an F- or a t-test, and obtain identical results.

The F-test of a single coefficient, or the equivalent t-test, is used by most variable selection programs such as stepwise regression. In the forward selection phase the explanatory variable having the greatest partial correlation coefficient with the dependent variable is appended only if its regression coefficient is significantly different from zero, otherwise the procedure stops. In the backward elimination phase, the explanatory variable with the smallest partial correlation coefficient is delected only if its regression coefficient is not significantly different from zero.

APPENDIX A. ABSENTEEISM DATA BASE

Case Number	Absenteeism	Job Classification	Job Complexity	Base Pay	Foreman Satisfaction	Seniority	Age	Number of Dependents
1	0	14	45	3.86	4	3	28	2
2	1	22	76	5.74	4	10	42	1
3	0	21	56	3.08	1	9	40	5
4	2	22	76	5.74	3	7	34	2
5	0	9	70	5.92	3	14	39	2
6	1	7	69	4.31	3	9	44	0
7	1	21	56	3.78	4	3	40	1
8	1	21	56	2.70	4	1	35	0
9	2	19	43	4.99	1	9	32	0
10	1	22	76	3.63	3	1	41	1
11	3	11	30	3.02	2	1	27	1
12	2	15	50	4.88	4	9	40	0
13	1	17	10	2.80	4	1	30	1
14	3	7	69	4.48	2	4	35	0
15	2	12	67	5.61	3	3	33	1
16	0	7	69	4.44	1	4	32	1
17	4	9	70	5.34	2	8	37	1
18	7	1	13	4.17	2	1	26	2
19	3	25	16	5.87	3	3	36	2
20	2	8	52	5.39	1	5	28	2
21	2	8	52	4.87	1	16	40	1
22	4	24	3	4.04	2	2	26	0
23	2	18	6	3.38	3	4	38	2
24	0	12	67	6.42	3	6	33	1
25	3	17	10	2.66	3	1	26	0
26	3	6	89	7.51	3	18	48	0
27	3	23	21	2.83	2	2	34	1
28	0	28	34	4.27	3	4	26	1
29	2	4	12	6.47	4	6	40	2
30	3	9	70	4.71	2	2	34	2
31	1	7	69	4.39	3	11	49	2
32	4	1	13	3.77	2	1	35	1
33	2	11	30	4.28	4	13	51	5
34	1	19	43	3.19	2	1	25	1
35	3	5	8	4.40	2	2	29	0
36	2	3	69	5.03	2	2	34	2
37	4	11	30	2.84	4	1	36	2
38	4	16	23	2.81	2	1	31	2

Case Number	Absenteeism	Job Classification	Job Complexity	Base Pay	Foreman Satisfaction	Seniority	Age	Number of Dependents
39	4	25	16	4.57	4	1	28	2
40	3	26	11	3.60	3	1	32	2
41	2	25	16	3.89	3	1	30	2
42	6	15	50	4.24	1	2	30	2
43	3	15	50	3.61	3	2	28	0
44	1	3	69	6.79	3	4	31	4
45	2	17	10	3.53	3	2	34	1
46	1	19	43	4.86	3	26	54	1
47	1	4	12	3.93	4	1	28	0
48	3	22	76	5.20	2	5	28	2
49	2	21	56	3.22	3	2	35	2
50	0	18	6	3.88	3	8	43	4
51	0	5	8	4.73	5	3	29	4
52	1	14	45	3.37	4	2	32	3
53	3	19	43	3.84	3	5	31	3
54	6	16	23	2.64	3	1	26	1
55	3	27	1	5.18	5	7	46	1
56	2	10	82	5.58	3	1	23	1
57	2	27	1	3.94	3	1	20	0
58	4	27	1	3.84	5	1	35	3
59	3	9	70	5.80	3	4	32	2
60	0	22	76	5.00	3	6	34	0
61	0	10	82	7.47	3	7	38	0
62	1	15	50	4.21	3	9	33	2
63	1	9	70	6.56	3	8	45	1
64	1	2	81	4.60	3	5	27	1
65	2	9	70	5.61	3	9	33	4
66	3	27	1	5.35	4	2	30	3
67	2	5	8	3.51	5	1	32	2
68	2	16	23	3.27	4	2	24	3
69	2	23	21	3.67	4	12	47	4
70	2	20	82	6.54	3	7	33	5
71	1	12	67	6.82	4	28	54	3
72	0	2	81	5.00	3	18	45	2
73	1	19	43	5.50	3	6	40	0
74	4	18	6	3.58	3	3	21	1
75	3	1	13	5.44	2	8	29	4
76	2	8	52	5.24	1	7	31	1
77	3	8	52	3.26	3	1	27	1

Source: Computer simulation by one of the authors

APPENDIX B. ABSENTEEISM DATA BASE DOCUMENTATION

Variable Name	Symbol	Description
Case Number	i	(also called observation number)
Absenteeism	A	The number of distinct occasions that the worker was absent during 1975. Each occasion consists of one or more consecutive days of absence.
Job Classification	J	An integer identifying the twenty-nine different jobs included in the study: 1 = Foundry Molder, 2 = Automatic Screw Machine Operator, 3 = Aluminum Extrusion Inspector, 4 = Warehouse Order Picker, 5 = Heavy Hydraulic Press Operator, etc.
Job Complexity	C	An index ranging from zero to one hundred, measured according to procedures developed by Turner and Lawrence.*
Base Pay	P	Base hourly pay rate ($)
Foreman Satisfaction	S	Determined by employee response to the question: "How satisfied are you with your foreman?" 1 = Very dissatisfied 2 = Somewhat dissatisfied 3 = Neither satisfied or dissatisfied 4 = Fairly well satisfied 5 = Very satisfied
Seniority	SE	Number of complete years with the company on December 31, 1975.

*Turner, Arthur N. and Lawrence, Paul R. Industrial Jobs and the Worker (Boston: Harvard University, 1965).

Variable Name	Symbol	Description
Age	AG	Employee's age on December 31, 1975
Number of Dependents	D	Determined by employee response to the question: "How many individuals other than yourself depend on you for most of their financial support?"