

55

AD 661836

MAN 67-1

# CORRELATION AND REGRESSION TECHNIQUES

by

John W. McCloskey  
University of Dayton Research Institute  
Dayton, Ohio



October 1967

**Management Analysis Note . . .**

**Office for Laboratory Management**

**Office of the Director of Defense  
Research and Engineering  
Washington, D.C.**

**Best Available Copy**

This document has been approved  
for public release and sale; its  
distribution is unlimited.

D D C  
RECEIVED  
NOV 29 1967  
RECEIVED  
C. / 35

CORRELATION AND REGRESSION TECHNIQUES

by

John W. McCloskey  
University of Dayton Research Institute  
Dayton, Ohio

October 1967

Management Analysis Note 67-1

Office for Laboratory Management  
Office of the Director of Defense Research and Engineering  
Washington, D.C. 20301

## FOREWORD

This report was written in an attempt to give the scientist, engineer or laboratory manager who has a limited background in the area of statistics a better understanding of the methods of correlation and regression. A number of examples have been given to illustrate the methods that have been developed, together with numerous graphical representations to give the reader a pictorial description of many interlocking relationships. No prior knowledge of statistics is assumed, but some experience in mathematics beyond calculus is necessary to fully comprehend the theoretical development.

No attempt has been made in this article to discuss such topics as confidence intervals or tests of hypotheses involving the correlation coefficient or regression equation. While these topics are certainly important statistical concepts, it was considered desirable to restrict the scope of the text to interpretations of the principles of correlation and regression.

Over the past year, the author has been a statistical consultant to the Office for Laboratory Management in the Office of the Director of Defense Research and Engineering. He was motivated to prepare this review as a preface to a series of studies utilizing these statistical techniques that will be published in the near future.

CONTENTS

	<u>Page</u>
Foreword-----	iii
1. Introduction-----	1
2. Regression in Two Variables-----	2
3. Correlation in Two Variables-----	5
4. Describing the Data-----	9
5. Spearman's Rank Correlation Coefficient-----	15
6. Multiple Linear Regression-----	17
7. The Correlation Matrix-----	20
8. Multiple Correlation Coefficients-----	22
9. Geometrical Considerations-----	23
10. Partial Correlation Coefficients-----	25
Bibliography-----	29

## I. INTRODUCTION

The methods of regression have been used quite extensively in recent years in the prediction of certain random phenomena called variables. Frequently it is difficult to obtain observations on one variable, or observations can be obtained only after a considerable time delay. In such cases it is often desirable to establish a relationship between this variable and one or more other variables which can more easily be observed so that the former variable can be predicted. For example, a student's college point average is a variable which can be observed only after the completion of four years of work, while the student's high-school grades and scores on various college entrance examinations are available prior to his enrollment into college. Many educators have been interested in predicting college success from such precollege records. Another major concern in our society today is that of predicting a man's salary from age, educational background and type of work data. The personnel offices of large corporations must know what factors influence salaries so that they can remain competitive with other organizations.

In regression the variable which is being predicted is often referred to as the dependent variable, and the variables used in the predicting are referred to as the independent variables. The initial step in an investigation requires that observations be made on both the dependent and independent variables. Using this data a relationship is established which best describes the trend of the observations between the dependent and independent variables so that in the future it is only necessary to observe the independent variables to make a reliable prediction as to what the value of the dependent variable would be if observed. For example, in a college admissions office, a study might be made of current college seniors using point average as the dependent variable and the student's high school grades and scores on college entrance examinations as the independent variables so that only students with a high predicted point average would be admitted the following year. The relationship that is established using the observations from the dependent and independent variables will be called the regression equation, and the mechanics of establishing this relationship will be discussed in the text of this article.

It should be pointed out that in regression one is not necessarily seeking perfect prediction of the dependent variable. Often the dependent variable will be strongly related to one or possibly two variables and only weakly related to a number of others. This is exhibited in the salary problem, where age and educational level are strong factors and such things as personality traits, compatibility with fellow employees and others, while sometimes important, for most employees have little effect and are difficult to measure in terms of salary effect. In regression, therefore, one is only seeking a relationship involving those variables for which the dependent variable is strongly related. No interpretation of the term "strongly related" will be given here, since the interpretation may depend upon the particular investigation being conducted.

Having obtained a regression equation, it is often desirable to obtain a measure of the strength of the hypothesized relationship, where the term strength will be taken to mean the degree to which the data follows the regression equation. Several correlation techniques have, therefore, been developed to describe the strength of the regression equation and interpretations will be given as to the meaning of the results.

As a matter of notation, capital letters (such as  $Y$ ,  $X$ ,  $X_2$ ) will be used to designate variables and small letters (such as  $y_i$ ,  $x_{1i}$ ,  $x_{2i}$ ) used to designate particular observations from the variables. Small letters will also be used in writing the regression equations.

## 2. REGRESSION IN TWO VARIABLES

Suppose an investigator is interested in studying the relationship between salary and age of the professional employees at Company A. The investigator initially believes that a relationship exists because he conjectures that an older man, in general, will have more experience in his profession than his younger colleague, and therefore will be more valuable to his employer. It is realized by the investigator that age is by no means the only factor involved, but it is conjectured that at least some trend will be present. A plot is therefore made of the observations on the variable  $Y$  (salary in dollars/month) versus the observations on the variable  $X$  (age in years) for the professional employees of the company and presented in the upper graph of Figure 1 along with the data. A casual observation shows that there appears to be an increasing trend of salary with age, but the investigator wishes to make more than just a subjective appraisal of the data. A further examination indicates that the trend appears to be linear, so the investigator wishes to find an equation of the form

$$y = a + bx \quad (1)$$

which describes the trend of the observations from the variables  $X$  and  $Y$ . That is, for this set of data, determine constants  $a$  and  $b$  so that an employee's salary in dollars per month can be predicted by multiplying his age in years by the constant  $b$  and adding the constant  $a$ . It should be emphasized that the quantity  $a + bx$  only serves as a predicted salary for an employee of age  $x$  years based upon a trend determined by the entire professional population and that a particular individual's salary might differ considerably from its predicted value, owing to the influence of other factors besides age (such as educational background and personality traits) which affect an employee's salary.

The investigator is thus interested in determining the constants in Equation (1) so that this equation best represents the trend established by Company A. The phrase "best represents" could be interpreted quite

differently by various investigators, so it is desirable to introduce the standard technique of fitting a line to the data known as the method of least squares--that is, find constants  $a$  and  $b$  in Equation (1) such that the sum

$$S = \sum_{i=1}^n (y_i - a - bx_i)^2 \quad (2)$$

is minimized where the sum is taken over all pairs  $(x_i, y_i)$  representing the age and salary pair of the  $i$ th employee. It will be recalled that  $a + bx_i$  will represent the predicted salary for the  $i$ th employee, so the quantity

$$d_i = y_i - (a + bx_i)$$

represents the  $i$ th employee's salary deviation from the linear trend of the professional population of Company A. From Equation (2) the method of least squares is seen to find constants  $a$  and  $b$  to make the sum of the squares of these deviations as small as possible. The details will not be stated here, but it can be shown that the constants  $a$  and  $b$  satisfying the least squares criterion can be found for Equation (2) by obtaining the partial derivatives  $\partial S/\partial a$  and  $\partial S/\partial b$  and then solving simultaneously the two equations

$$\frac{\partial S}{\partial a} = 0 \quad \frac{\partial S}{\partial b} = 0$$

for  $a$  and  $b$ . The result will yield the values

$$b = \frac{\sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^n x_i^2 - n\bar{x}^2} \quad a = \bar{y} - b\bar{x} \quad (3)$$

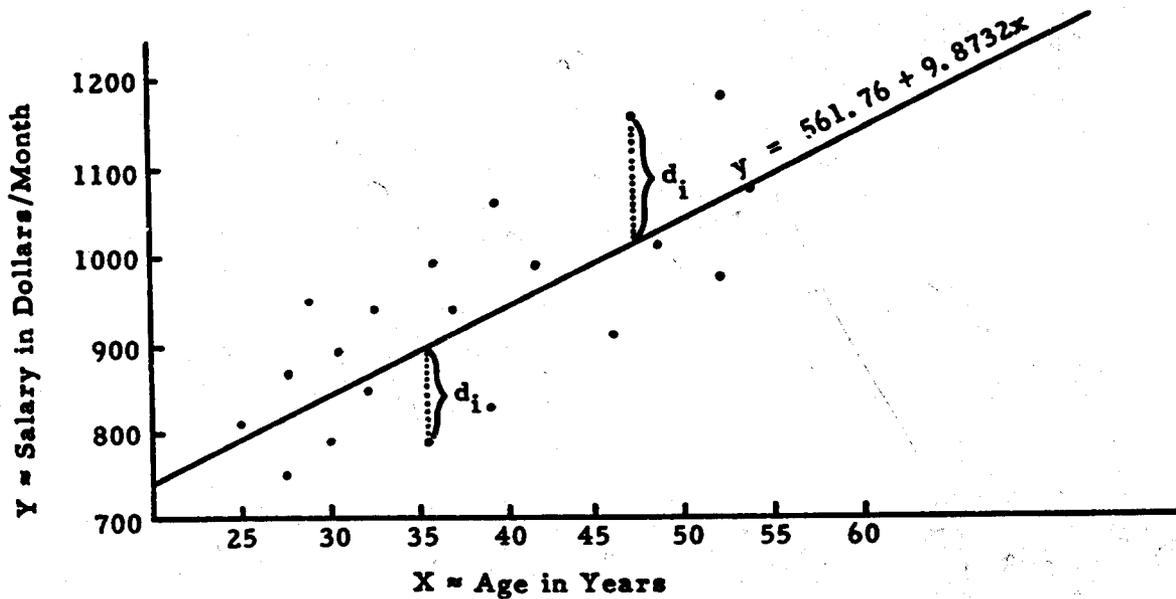
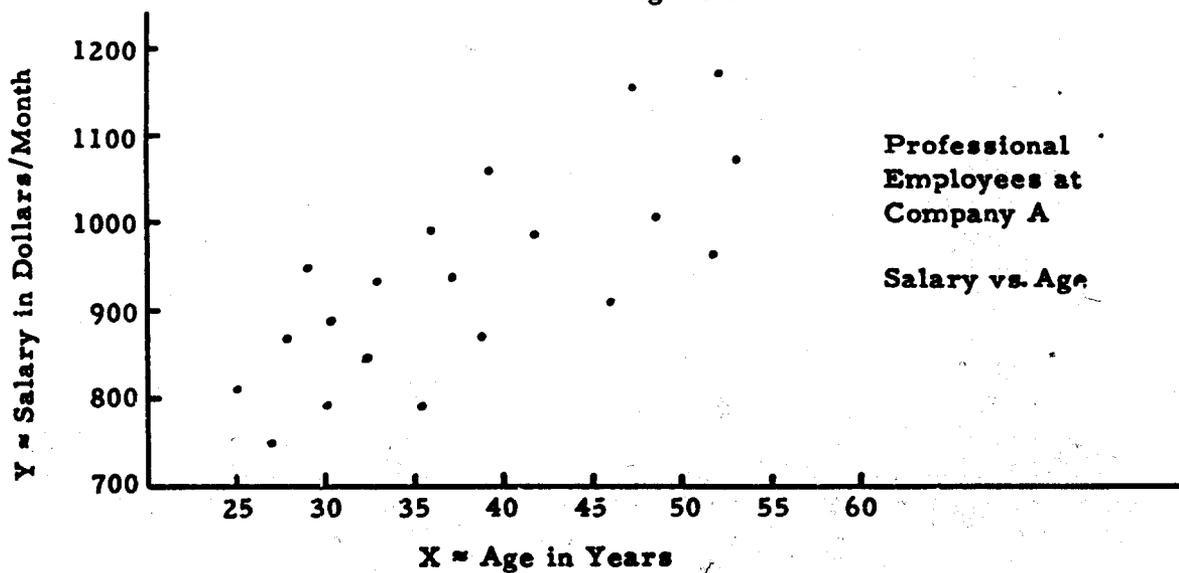
where  $\bar{x} = \sum_{i=1}^n x_i/n$  and  $\bar{y} = \sum_{i=1}^n y_i/n$ .

When  $a$  and  $b$  are calculated from (3), Equation (1) is known as the linear regression of  $Y$  on  $X$ . From (3) it is seen that the constant  $a$  is determined directly from  $b$  and the two means. Thus, a desirable alternative form of the linear regression can be obtained by the direct substitution for the constant  $a$  into Equation (1), yielding the modified form

$$y = \bar{y} + b(x - \bar{x}). \quad (3a)$$

This form is desired by many because it explicitly exhibits the means of the data used to obtain the regression equation.

Figure 1



Data Table

$i$	$x_i$	$y_i$									
1	25.0	810	6	30.3	792	11	36.4	894	16	47.2	1175
2	27.1	755	7	32.7	847	12	37.4	868	17	48.6	1016
3	27.4	863	8	33.2	926	13	39.2	1060	18	52.4	973
4	29.1	956	9	35.6	791	14	40.9	986	19	53.7	1083
5	30.2	890	10	35.8	995	15	45.8	917	20	54.6	1168

Using the data given in Figure 1, the linear regression has been calculated and this line plotted through the data in the second graph of Figure 1. Geometrically, the deviations  $d_i$  represent the vertical distance from the regression line to the point  $(x_i, y_i)$  as illustrated in the figure. The regression line has the properties that

$$\sum_{i=1}^n d_i = 0$$

and

$$\sum_{i=1}^n d_i^2$$

are minimized. It should be mentioned that, if  $X$  is considered to be the dependent variable instead of  $Y$ , then the regression of  $X$  on  $Y$  cannot be found by solving the equation  $y = a + bx$  for  $x$ . Rather, one wishes to find constants  $a'$  and  $b'$  in the equation  $x = a' + b'y$  such that the sums of squares of the horizontal distances from the points  $(x_i, y_i)$  to the curve  $x = a' + b'y$  are minimized. This solution for  $a'$  and  $b'$  may be considerably different from the constants found by the inversion of the equation  $y = a + bx$ , especially if the sample size is small.

### 3. CORRELATION IN TWO VARIABLES

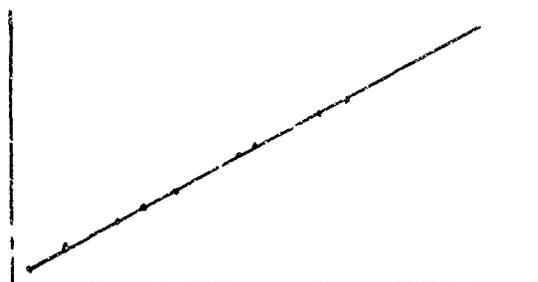
Having obtained a regression equation, the investigator is often interested in determining the strength of this relationship obtained by his regression technique. One of the most frequently used measures of this fit is the Pearson product moment correlation coefficient obtained from the observations on the variables  $X$  and  $Y$  by the equation

$$r = \frac{\sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}}{\sqrt{(\sum_{i=1}^n x_i^2 - n\bar{x}^2)(\sum_{i=1}^n y_i^2 - n\bar{y}^2)}} \quad (4)$$

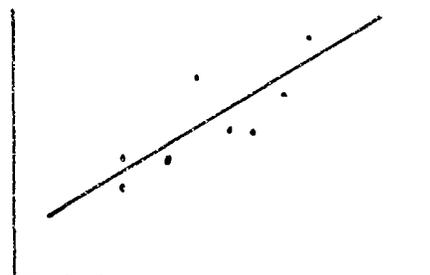
Some of the major properties of this correlation coefficient are as follows:

- (a) For any set of points  $-1 \leq r \leq +1$ .
- (b)  $r = +1$  if the points lie on a straight line with positive slope.
- (c)  $r = -1$  if the points lie on a straight line with negative slope.

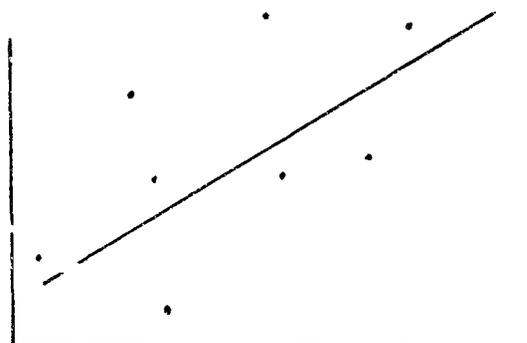
Figure 1.1



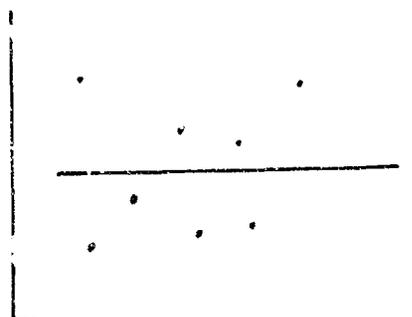
Perfect Positive Correlation  
 $r = +1$



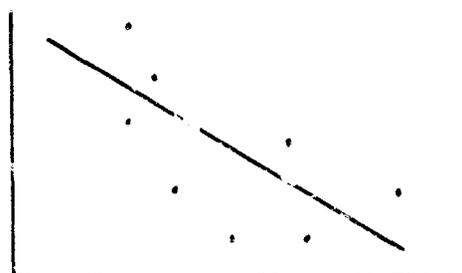
Strong Positive Correlation  
 $r$  approximately  $+ 0.80$



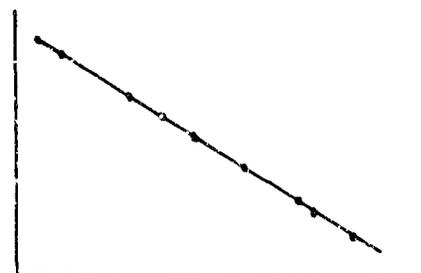
Weak Positive Correlation  
 $r$  approximately  $+ 0.20$



No Correlation  
 $r$  approximately zero



Moderate Negative Correlation  
 $r$  approximately  $- 0.50$



Perfect Negative Correlation  
 $r = -1$

Examples of Data Possessing Given Correlation Coefficients.

- (d)  $r$  will be close to zero if the data has no linear trend .
- (e) If each  $x_i$  is multiplied by a positive constant, the value of  $r$  is unchanged .
- (f) If a constant is added to each  $x_i$ , the value of  $r$  is unchanged .

Property (a) gives the range of the correlation coefficient, while properties (b), (c), and (d) give special significance to three particular values. In addition, a positive correlation less than one indicates that the slope  $b$  of the regression line is positive but that the data points do not fall on a straight line. Similarly, a negative correlation indicates the slope of the regression line is negative. The relationship between the slope of the regression line and the correlation coefficient is exhibited more explicitly by the equation

$$r = b \sqrt{\frac{\sum_{i=1}^n x_i^2 - n\bar{x}^2}{\sum_{i=1}^n y_i^2 - n\bar{y}^2}}$$

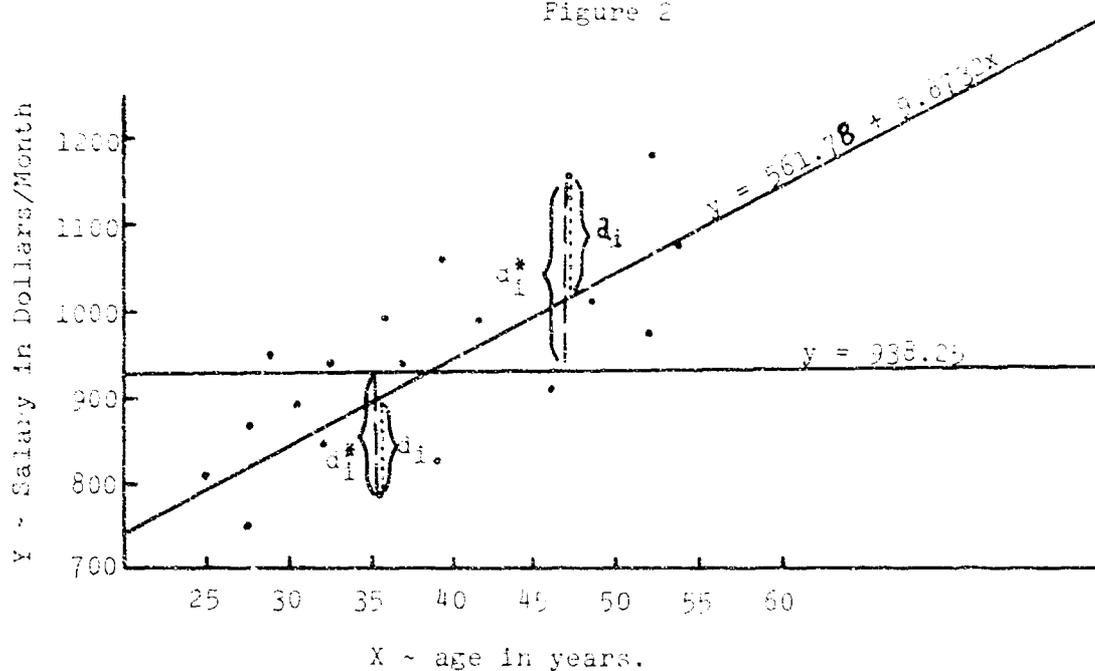
obtained from Equations (3) and (4). To further illustrate these properties, six hypothetical sets of data have been given in Figure 1.1 together with their corresponding correlations coefficients. Properties (e) and (f) indicate that certain linear transformations of the  $x_i$ 's do not alter the value of the correlation coefficient. Such transformations are often very helpful in simplifying computations, especially if the correlation is calculated by hand or with a desk calculator. Property (e) also implies that the value of the correlation coefficient does not depend upon the units of the  $x_i$ 's. That is, if  $x_i$  is a length, the same correlation will be obtained whether the  $x_i$ 's are recorded in inches, feet, centimeters or miles.

To give a specific meaning for all values  $-1 \leq r \leq +1$  it can be shown by squaring both sides of Equation (4) that

$$r^2 = 1 - \frac{\sum_{i=1}^n (y_i - a - bx_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (5)$$

where  $a$  and  $b$  are from Equation (3). Noting that  $b = 0$ ,  $a = \bar{y}$  is a solution which makes the sum of squares in the numerator equal to the sum of squares in the denominator, the solution given by (3) must of necessity give a sum of squares in the numerator which is no larger than the denominator, since this solution was obtained so as to minimize the numerator. The square of the correlation coefficient  $r^2$  thus has range  $0 \leq r^2 \leq 1$  and is often referred to

Figure 2



$$\bar{x} = 38.13 \quad \bar{y} = 938.25$$

Regression Line

$$y = 561.78 + 9.8732x$$

Regression in Modified Form

$$y = 938.25 + 9.8732(x - 38.13)$$

Correlation Coefficient

$$r = 0.76787$$

Coefficient of Determination

$$r^2 = 0.58963$$

as the coefficient of determination. Using

$$d_i^* = y_i - \bar{y}$$

to represent the deviation of the  $i$ th observation from the mean  $\bar{y}$ , it can be seen from Equation (5) and Figure 2 that  $r^2$  can be interpreted as the fraction reduction in the variation of the variable  $Y$  when the sum of squares is measured from the regression line of  $Y$  on  $X$  instead of from the mean  $\bar{y}$ . Another way of expressing this would be to say that the linear regression of  $Y$  on  $X$  explains  $100 r^2\%$  of the variation of the variable  $Y$ .

In the salary versus age data given in Figures 1-2 the correlation coefficient was calculated from Equation (4) and found to be 0.768. From this, the coefficient of determination was calculated and found to be 0.590. Thus, for Company A, the linear dependence of salary on age explains 59.0% of the variation in salary.

Another useful measure of the strength of the regression equation is the standard error of estimate  $S_e$  defined by the equation

$$S_e = \sqrt{\frac{1}{n-2} \sum_{i=1}^n (y_i - a - bx_i)^2} \quad (6)$$

where  $a$  and  $b$  are defined by (3). This measure is based upon the square of the deviations of the points from the regression line and therefore measures the spread of the points about the regression line. The quantity  $S_e$  is equal to zero if all points lie on a straight line and is positive otherwise, with larger values of  $S_e$  indicating weaker linear trends.

#### 4. DESCRIBING THE DATA

The quantities  $b$ ,  $r$  and  $S_e$  are all useful in describing characteristics of the data, but care must be taken to avoid misinterpretations. The correlation coefficient  $r$  is greatly dependent upon the slope of the regression line  $b$ , so that two sets of data which look quite different on a graph may have nearly identical correlation coefficients. To more clearly exhibit this property, salary versus age data has been gathered on the professional employees of companies A, B, and C in Figure 3, and a graph of this data presented in Figure 4. Below the data in Figure 3, and also in Figure 4, a number of statistics have been calculated in an attempt to adequately summarize the data from the three companies. A comparison between companies A and B indicates that if the standard error of estimate,  $S_e$ , is held fairly constant, an increase in the slope  $b$  will cause an increase in the correlation coefficient  $r$ . Note that companies B and C have similar slopes, but in this case an increase in the standard error of estimate,  $S_e$ ,

Figure 3

Company A			Company B		Company C	
i	$x_i$	$y_i$	$x_i$	$y_i$	$x_i$	$y_i$
1	25.0	810	25.1	926	26.6	887
2	27.1	755	27.7	796	28.2	863
3	27.4	863	28.7	853	28.4	841
4	29.1	956	29.2	903	30.4	848
5	30.2	890	31.2	764	33.7	886
6	30.3	792	32.8	829	33.9	844
7	32.7	847	33.9	771	36.2	912
8	33.2	926	34.2	919	36.4	850
9	35.6	791	37.1	797	37.2	877
10	35.8	995	37.2	845	38.6	892
11	36.4	894	37.4	952	42.7	870
12	47.4	868	42.2	837	43.1	921
13	39.2	1060	43.7	912	45.0	883
14	40.7	986	47.0	860	47.1	879
15	45.8	917	48.1	1017	48.3	942
16	47.2	1175	50.1	808	49.7	906
17	48.6	1016	51.8	962	52.5	939
18	52.4	973	54.0	1036	54.0	895
19	53.7	1083	54.3	810	55.0	933
20	54.6	1168	57.1	902	57.6	955
$\bar{x} = 38.13$			$\bar{x} = 40.14$		$\bar{x} = 41.20$	
$\bar{y} = 938.25$			$\bar{y} = 874.95$		$\bar{y} = 891.15$	
Slope of Regression Line						
$b = 9.8732$			$b = 2.7206$		$b = 2.6546$	
Correlation Coefficient						
$r = 0.76787$			$r = 0.34836$		$r = 0.74870$	
Coefficient of Determination						
$r^2 = 0.58963$			$r^2 = 0.12135$		$r^2 = 0.56055$	
Standard Error of Estimate						
$S_e = 79.12$			$S_e = 75.36$		$S_e = 23.31$	

Figure 4

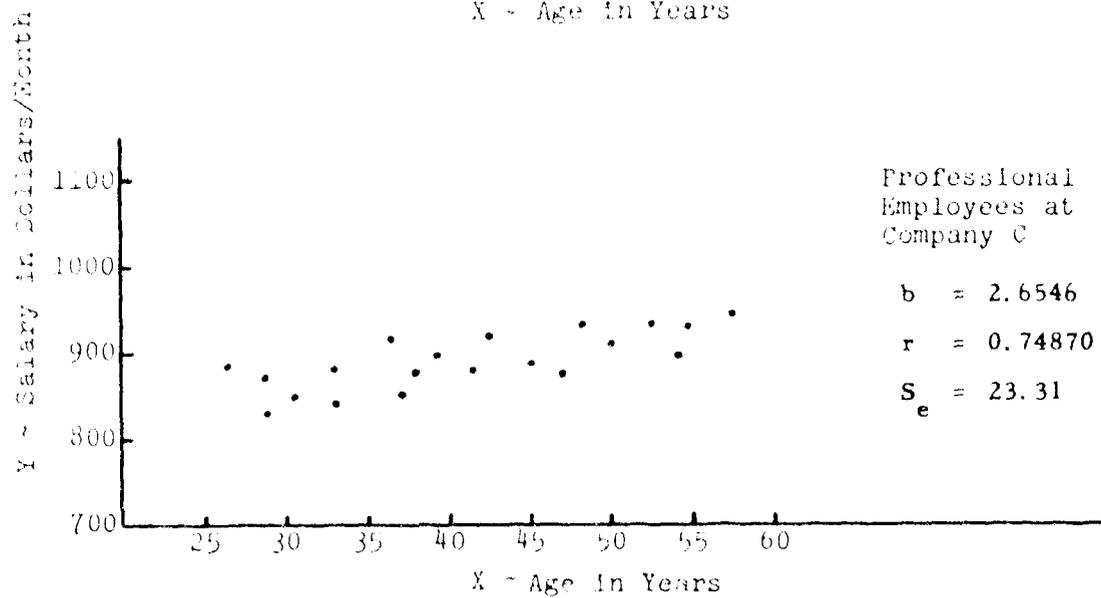
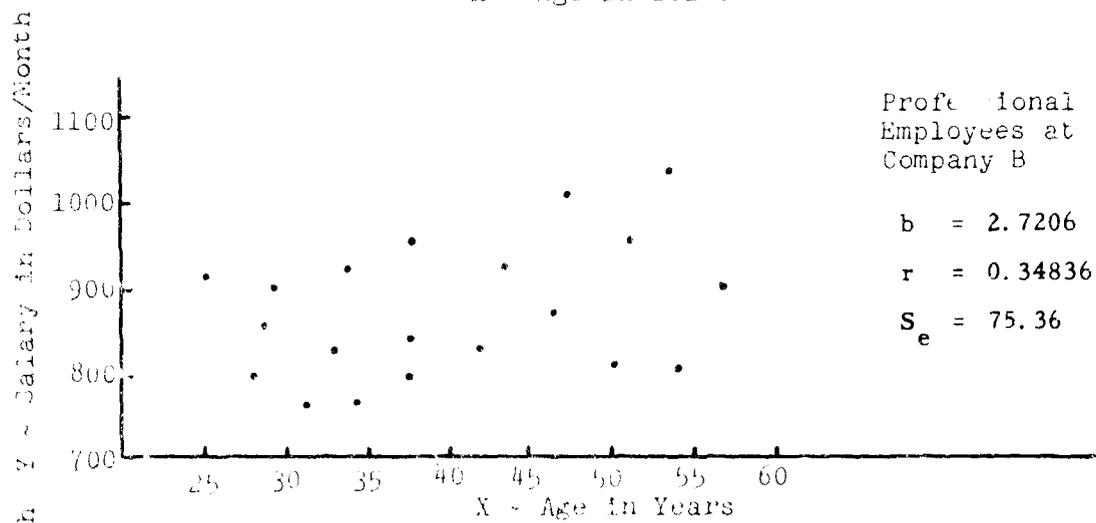
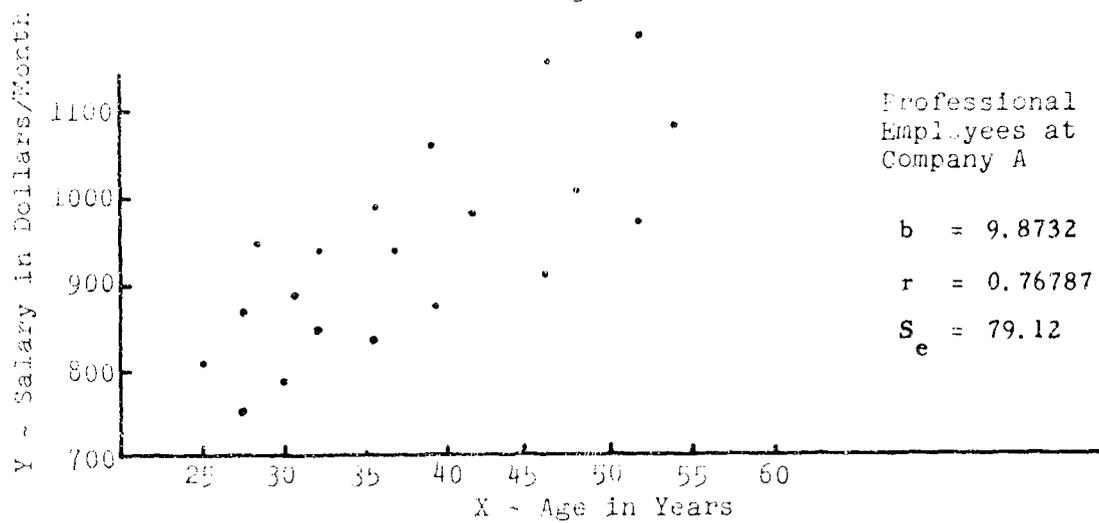
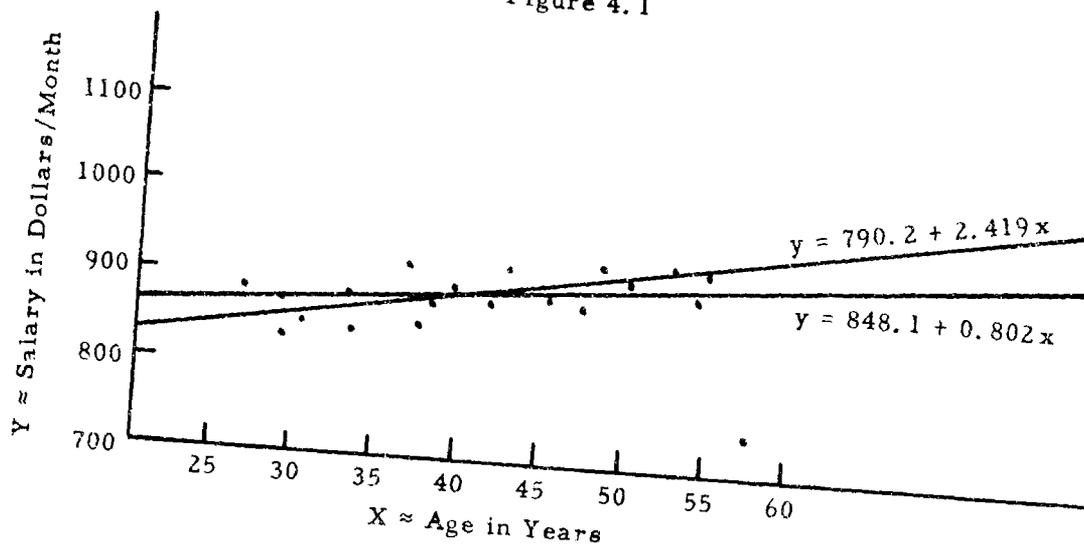


Figure 4.1



Statistics with  
Bad Points  
Included

$b = 0.802$

$r = 0.1812$

$S_e = 43.19$

Slope of Regression Line

Correlation Coefficient

Standard Error of Estimate

Statistics with  
Bad Points  
Excluded

$b = 2.419$

$r = 0.6960$

$S_e = 23.34$

produces a corresponding decrease in the correlation coefficient,  $r$ , for Company B. A comparison between A and C shows that these two effects can be neutralized. That is, both companies have similar correlations  $r$ , but company A has a much larger standard error of estimate,  $S_e$ , and also larger slope  $b$ . This last comparison illustrates that two sets of data can have similar correlation coefficients but exhibit different linear trends. This results because the correlation coefficient as shown by Equation (5) measures only a relative reduction in the sum of squares of the variable  $Y$  due to the linear regression of  $Y$  on  $X$ .

In summary, the three quantities  $S_e$ ,  $b$  and  $r$  individually give the investigator only limited information about the data, but together they provide a good description of the linear relationship between the dependent and independent variables. That is,  $b$  gives the slope of the regression line,  $S_e$ , a measure of the spread about the regression line and  $r^2$  the relative reduction in the sum of squares due to the linear regression.

Another interesting aspect of the data in Figures 3-4 is that the predicted salary for a man of age 30, which can be found by substituting the value  $x = 30$  into Equation (3a) for each company, is about the same for all three companies, as shown in the table below.

Predicted Salary at Age 30 in Dollars/Month	
Company A	858
Company B	847
Company C	861

Thus the starting salary for a young professional might well be the same at all three companies, but because of the advancement policy of the individual companies, the yearly increases might vary drastically from company to company.

Finally, it should be mentioned that when one uses the correlation and regression techniques, special care should be taken to eliminate all errors from the data. The method of least squares is especially sensitive to extraneous points, and even a few bad points among a few hundred can drastically alter the results. As an illustration, let us suppose that the salary of the last individual in Company C was recorded erroneously as \$755/month in Figure 3. Figure 4.1 shows a plot of the regression line with this bad point included in the data in addition to the recalculation of the regression equation after this point has been eliminated. The effect of this one bad point is surprising unless one is familiar with the mathematical analysis. It is, therefore, often advisable to plot the data on a graph and

Figure 5

$x_i$	$Rx_i$	$y_i$	$Ry_i$	$Rx_i - Ry_i$	$(Rx_i - Ry_i)^2$
25.0	1	810	4	-3	9
27.1	2	755	1	+1	1
27.4	3	863	6	-3	9
29.1	4	956	12	-8	64
30.2	5	890	8	-3	9
30.3	6	792	3	+3	9
32.7	7	847	5	+2	4
33.2	8	926	11	-3	9
35.6	9	791	2	+7	49
35.8	10	995	15	-5	25
36.4	11	894	9	+2	4
37.4	12	868	7	+5	25
39.2	13	1060	17	-4	16
40.9	14	986	14	0	0
45.8	15	917	10	+5	25
47.2	16	1175	20	-4	16
48.6	17	1016	16	+1	1
52.4	18	973	13	+5	25
53.7	19	1083	18	+1	1
54.6	20	1168	19	+1	1
	<u>210</u>		<u>210</u>	0	302

$$r^* = 1 - \frac{6(302)}{20(399)} = 1 - 0.2271 = 0.7729$$

check the validity of any point that does not appear to follow the trend established by the majority of the data points.

## 5. SPEARMAN'S RANK CORRELATION COEFFICIENT

Another measure of the correlation between two variables that can be used is Spearman's rank correlation coefficient. To obtain this measure for a set of data defined by the pairs  $(x_i, y_i)$ , the coordinates must be ranked with respect to the two variables; thus, the smallest  $x_i$  is given rank one, the second smallest rank two, and continuing until the largest  $x_i$  is given rank  $n$ . Define, therefore,  $R_{x_i}$  to be the rank of the coordinate  $x_i$  among the  $n$  observations on the variable  $X$  and similarly define the rank  $R_{y_i}$ . Using then the data  $(R_{x_i}, R_{y_i})$  in Equation (4), the results give Spearman's rank correlation coefficient. However, since the data now consists only of integer values, Equation (4) can be simplified in this special case to give the more common equation for the calculation of the rank correlation  $r^*$  as follows:

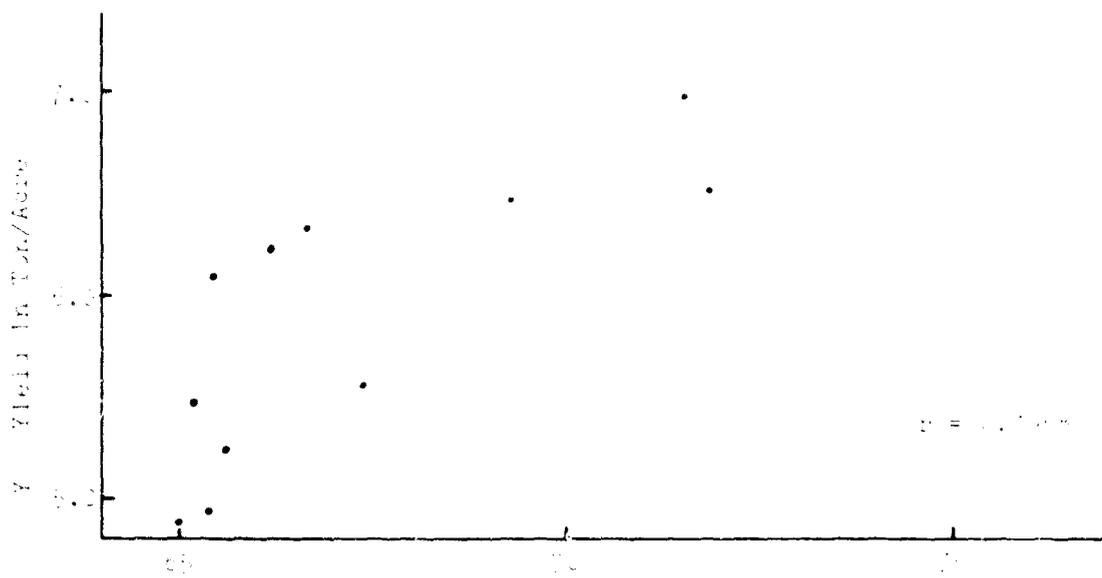
$$r^* = 1 - \frac{6 \sum_{i=1}^n (R_{x_i} - R_{y_i})^2}{n(n^2 - 1)} \quad (7)$$

An example of the calculation of the rank correlation coefficient is given in Figure 5 using the salary versus age data of Company A. The rank correlation has also been calculated for Companies B and C using Equation (7) and a comparison with the product moment correlation given in the table below for the three companies.

	Product Moment Correlation Coefficient	Rank Correlation Coefficient
Company A	0.7679	0.7729
Company B	0.3484	0.3158
Company C	0.7487	0.7098

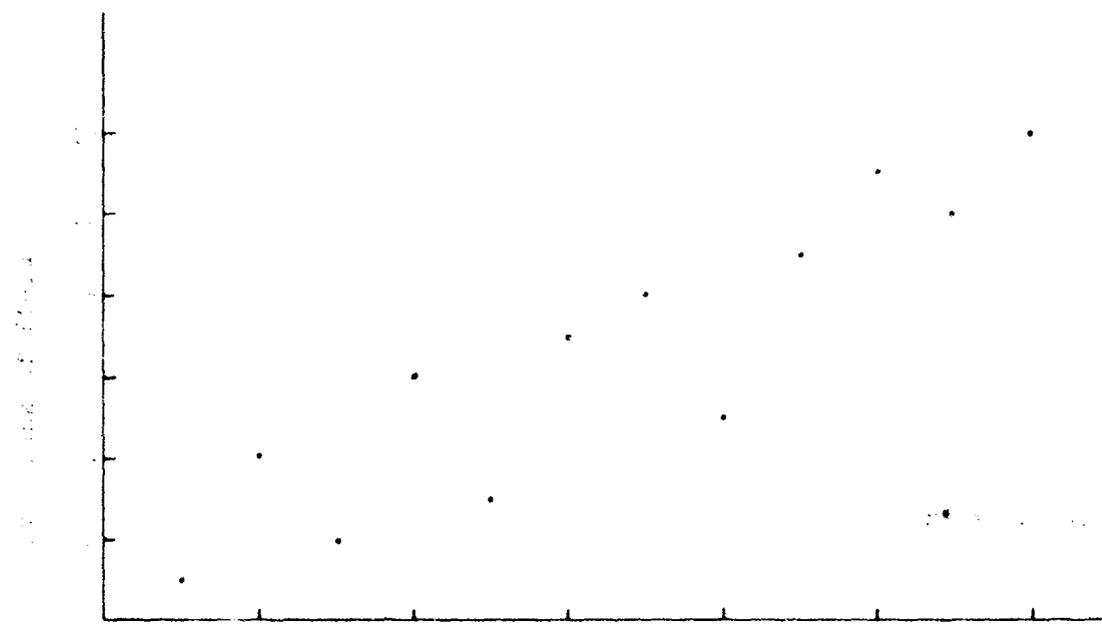
For the data just considered, the rank and product moment correlations were quite close, but this need not be the case. An example will now be given in an effort to more explicitly define the relationship which exists between the two correlation coefficients. The upper graph in Figure 6 is a plot of yield versus temperature for a given crop over a twelve-year period, where an individual point represents the yield for a given year plotted against the average temperature during the growing season for the given year. The product moment correlation was calculated

Figure 1  
original data



X = Average Temperature in Degrees Fahrenheit during drawing down

Ranked data



X = Average Temperature in Degrees Fahrenheit during drawing down

for this data using Equation (4) and found to be  $r = 0.7958$ . The data points are now ranked with respect to both their  $x$  and  $y$  coordinate values and plotted in the lower graph in Figure 6. This ranking procedure is essentially a transformation of the original data which preserves the rank of the observations but which distorts the spread of the data to give observations which are exactly one unit apart in both the horizontal and vertical directions. This transformation is introduced to simplify calculations, but, as shown in the example of Figure 6, this simplification is often produced at the expense of considerable distortion in the relative spread of the data. Using the ranked data, the rank correlation coefficient can be calculated from Equation (1) or from the simplified form given by Equation (7) and found to be  $r^* = 0.9091$ . A comparison indicates that the transformation to ranked data produces a change of more than 0.11 in the correlation coefficient in this case.

In general, if the original data is evenly spread with respect to the two variables, then  $r$  and  $r^*$  will be very close to each other, since little distortion is produced by the transformation to ranked data. If, however, a plot of the original data indicates that the points appear in clusters, it is entirely possible that  $r$  and  $r^*$  will be considerably different.

#### c. MULTIPLE LINEAR REGRESSION

Suppose the investigator feels the dependent variable  $Y$  is related to two or more independent variables and therefore wishes to use a function of several variables to predict the dependent variable  $Y$ . For  $k$  independent variables  $X_1, X_2, \dots, X_k$ , when the prediction equation takes the form

$$y = a_0 + a_1x_1 + a_2x_2 + \dots + a_kx_k \quad (8)$$

and if the coefficients  $a_0, a_1, \dots, a_k$  are determined in the least squares sense so as to minimize  $S$  defined by the equation

$$S = \sum_{i=1}^n (y_i - a_0 - a_1x_{1i} - \dots - a_kx_{ki})^2$$

then Equation (8) is known as the multiple linear regression of  $Y$  on  $X_1, X_2, \dots, X_k$ . As before, the solution can be found by obtaining a simultaneous solution to the system

$$\frac{\partial S}{\partial a_j} = 0 \quad \text{for } j = 0, 1, 2, \dots, k$$

Figure 7

Data Table

In a given area for the  $i$ th year define

$y_i$  = yield in tons/acre for a given crop

$x_{1i}$  = rainfall in inches during growing season

$x_{2i}$  = average temperature in degrees Fahrenheit during growing season

Year	$y_i$	$x_{1i}$	$x_{2i}$	$R_{y_i}$	$R_{x_{1i}}$	$R_{x_2}$
1	6.50	8.0	72.20	5	5	7
2	6.20	7.0	71.40	3	4	6
3	7.50	12.0	66.60	9	9	1
4	7.10	11.2	69.10	8	8	4
5	5.90	6.0	76.10	2	2	10
6	5.50	4.8	74.40	1	1	9
7	6.60	9.4	73.30	6	6	8
8	7.80	13.3	67.80	10	10	2
9	6.90	9.8	68.20	7	7	3
10	6.30	6.4	71.30	4	3	5

which upon simplification will yield the system

$$\begin{aligned}
 & a_0 + a_1 \sum_{i=1}^n x_{1i} + a_2 \sum_{i=1}^n x_{2i} + \dots + a_k \sum_{i=1}^n x_{ki} = \sum_{i=1}^n y_i \\
 & a_0 \sum_{i=1}^n x_{1i} + a_1 \sum_{i=1}^n x_{1i}^2 + a_2 \sum_{i=1}^n x_{1i} x_{2i} + \dots + a_k \sum_{i=1}^n x_{1i} x_{ki} = \sum_{i=1}^n x_{1i} y_i \quad (9) \\
 & \dots \\
 & a_0 \sum_{i=1}^n x_{ki} + a_1 \sum_{i=1}^n x_{ki} x_{1i} + a_2 \sum_{i=1}^n x_{ki} x_{2i} + \dots + a_k \sum_{i=1}^n x_{ki}^2 = \sum_{i=1}^n x_{ki} y_i
 \end{aligned}$$

For large values of  $k$  this system can best be solved by matrix methods, but if  $k$  is small, a solution can be obtained by substitution or by Kramer's Rule.

An alternative form often used for the linear regression is obtained by solving the first equation of the system (9) for  $a_0$  and making a direct substitution into Equation (8). This yields the equation

$$y = \bar{y} + a_1(x_1 - \bar{x}_1) + a_2(x_2 - \bar{x}_2) + \dots + a_k(x_k - \bar{x}_k) \quad (10)$$

Again, this form is often preferable to Equation (8) because it explicitly exhibits the means of the variables used to obtain the regression equation.

It should be noted that the variables  $X_i$  need not be independent. In fact, quite the contrary; it is possible for one variable to be a nonlinear function of some other variable. For example, it is possible to have  $X_2 = X_1^2$  or  $X_3 = X_1 + 3X_2^2$ . However, linear combinations of variables, such as  $X_3 = X_1 + 2X_2$  where the exponents on the independent variables are all one, are not permitted. In this case, the matrix of the system of Equations (9) is singular and the inverse does not exist; or, in other words, the coefficients  $a_i$  are not uniquely determined when any of the variables is a linear combination of some of the other variables.

Consider now an example employing the concept of multiple regression. Suppose an investigator wishes to predict the yield of a crop from rainfall and temperature data given in Figure 7.

Assuming the prediction equation has the form given by Equation (8) with  $k = 2$ , the system of Equations (9) reduces to the following three equations:

$$10a_0 + 87.9a_1 + 710.3a_2 = 66.3$$

$$87.9a_0 + 843.73a_1 + 6179.08a_2 = 600.44$$

$$710.3a_0 + 6179.08a_1 + 50537.19a_2 = 4692.08$$

Obtaining a simultaneous solution to this system for the coefficients  $a_0$ ,  $a_1$ ,  $a_2$  and then substitution into Equation (10) yields the regression equation

$$y = 6.63 + 0.21413(x_1 - 8.79) - 0.0382(x_2 - 71.03). \quad (11)$$

In a system like that which was just considered, where there is more than one independent variable, Equation (1) cannot be used directly to calculate the strength of the relationship given by Equation (11). Therefore, the methods of correlation will be generalized to handle this situation.

## 7. THE CORRELATION MATRIX

Suppose a problem is considered for which there is one dependent variable  $Y$  and two independent variables  $X_1$  and  $X_2$  as in the preceding example. Define  $r_{YX_1}$  to be the Pearson product moment correlation coefficient calculated from the observations on the variables  $Y$  and  $X_1$  using Equation (4) while ignoring the observations on the variable  $X_2$ . That is, calculate the correlation coefficient  $r_{YX_1}$  between  $Y$  and  $X_1$  as if the observations on  $X_2$  had never been taken. Similarly, define the product moment correlation between all other pairs of the three variables, and define  $r_{YY}$  to be the correlation of the variable  $Y$  with itself while ignoring the observations from the other two variables. The correlation matrix is then defined to be a listing of the product moment correlation for all pairs of variables presented in the form

$$\begin{bmatrix} r_{YY} & r_{YX_1} & r_{YX_2} \\ r_{X_1Y} & r_{X_1X_1} & r_{X_1X_2} \\ r_{X_2Y} & r_{X_2X_1} & r_{X_2X_2} \end{bmatrix}$$

Since from Equation (4), it can be shown by symmetry that

$$r_{XY} = r_{YX}$$

and that

$$r_{XX} = 1$$

it is customary to present only the upper half of the matrix in the form

$$\begin{bmatrix} 1 & r_{YX_1} & r_{YX_2} \\ & 1 & r_{X_1X_2} \\ & & 1 \end{bmatrix} \quad (12)$$

Using the data given in Figure 7 and the form given by (12) the following correlation matrix was obtained:

$$\begin{bmatrix} 1 & +0.9831 & -0.8781 \\ & 1 & -0.8312 \\ & & 1 \end{bmatrix} \quad (13)$$

This matrix indicates that there is a strong direct relationship between yield and rainfall and a strong inverse relationship between yield and temperature as well as a strong inverse relationship between the independent variables, rainfall and temperature. The main purpose of the correlation matrix is to present the correlation between all pairs of variables in a standard form so that the interlocking relationships may be observed. The concept of the correlation matrix may be generalized to any number of independent variables.

As another illustration of how closely the rank correlations approximate the product moment correlations, the quantities  $R_{y_i}$ ,  $R_{x_1i}$ ,  $R_{x_2i}$  were calculated as shown in Figure 7. Using Equation (7), the rank correlations were calculated for all pairs of variables and presented in the following rank correlation matrix:

$$\begin{bmatrix} 1 & +0.9879 & -0.8546 \\ & 1 & -0.8424 \\ & & 1 \end{bmatrix}$$

This example shows that with only ten points, a close approximation is obtained to the product moment correlations using rank correlation methods. It should be remembered that there is a fairly even spread of the data in this case, however, and one cannot expect the approximation to be as good if the data points are clustered.

## 8. MULTIPLE CORRELATION COEFFICIENTS

An examination of the correlation matrix (13) in view of Equation (5) indicates that in this example  $100 \times (0.9831)^2\% = 96.6\%$  of the variation in the yield can be explained by the linear regression of Y on  $X_1$ . However, suppose the investigator feels that temperature also has an effect on yield and decides to use both temperature and rainfall data to obtain a prediction equation for yield. From the data given in Figure 7 and using the form of Equation (8), he calculates the regression of Y on  $X_1$  and  $X_2$  and obtains Equation (11). It now becomes important to him to obtain a measure of the strength of this prediction equation to determine how much improvement in the prediction procedure has resulted from the use of both variables in the regression equation. An examination of the correlation matrix (13) indicates that yield and temperature are indeed related, but, since rainfall and temperature are also related, there remain some questions as to whether both variables are needed in predicting yield.

Therefore, generalizing Equation (4) to two independent variables, define the square of the multiple correlation coefficient  $R_{Y(X_1 X_2)}$  by the equation

$$R_{Y(X_1 X_2)}^2 = 1 - \frac{\sum_{i=1}^n (y_i - a_0 - a_1 x_{1i} - a_2 x_{2i})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (14)$$

where  $a_0, a_1, a_2$  are obtained by the solution of the system (9) with  $k = 2$ . The quantity  $R_{Y(X_1 X_2)}^2$  thus gives the fraction reduction in the variation of the variable Y explained by the linear regression of Y on  $X_1$  and  $X_2$ . Some properties of the multiple correlation coefficient are as follows:

- (a) For any set of data points  $0 \leq R_{Y(X_1 X_2)} \leq 1$ .
- (b) If  $R_{Y(X_1 X_2)}$  is close to one then the linear regression of Y on  $X_1$  and  $X_2$  is a good prediction equation for Y.
- (c) If  $R_{Y(X_1 X_2)}$  is small then the linear regression of Y on  $X_1$  and  $X_2$  is not a good prediction equation for Y.

Calculating now the multiple correlation coefficient for the regression Equation (11) using the data in Figure 7 together with Equation (14) it was found that

$$R_{Y(X_1 X_2)} = 0.9890 \quad (15)$$

This result indicates that  $100 \times (0.9890)^2\% = 97.8\%$  of the variation in yield can be explained by the linear regression of  $Y$  on  $X_1$  and  $X_2$ . Previously, it was shown that  $96.6\%$  of the variation in yield could be explained by the regression of  $Y$  on  $X_1$ , so that a subtraction indicates that an additional  $1.2\%$  of the variation in yield can be explained with the addition of the variable  $X_2$  to the prediction equation.

Equation (4) can be further generalized to  $k$  variables where the square of the multiple correlation coefficient  $R_{Y(X_1 X_2 \dots X_k)}$  is defined by the equation

$$R_{Y(X_1 X_2 \dots X_k)}^2 = 1 - \frac{\sum_{i=1}^n (y_i - a_0 - a_1 x_{1i} - \dots - a_k x_{ki})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (16)$$

where  $a_0 \dots a_k$  are obtained by the solution of the system (9).

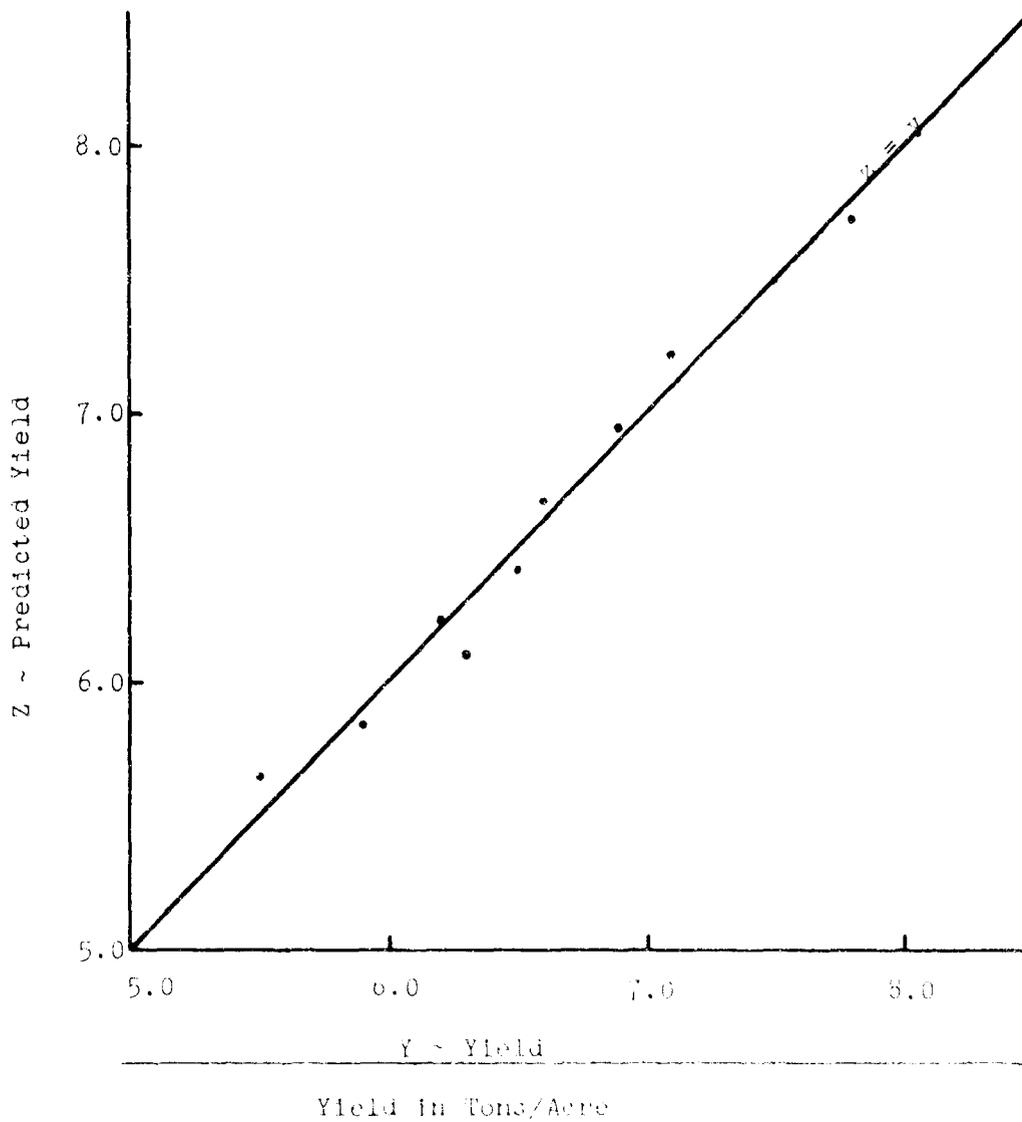
#### 9. GEOMETRICAL CONSIDERATIONS

The geometrical representation of the regression equation  $Y$  on  $X_1 \dots X_k$  is that of a  $k$ -dimensional hyperplane in  $k+1$ -dimensional Euclidian space which is best in the least squares sense. This geometrical representation, although mathematically interesting, has little practical use. There is, however, another geometrical approach which is very useful in exhibiting some important characteristics of the data. Consider the example given in Figure 7, and define  $(x_{1i}, x_{2i}, y_i)$  to be the rainfall, temperature and yield triple for the  $i$ th year. Using the ten triples, the regression of  $Y$  on  $X_1$  and  $X_2$  was calculated for this data and given in Equation (11). Using the triple  $(x_{1i}, x_{2i}, y_i)$ , define  $z_i$  to be the predicted yield for the  $i$ th year obtained by substituting the values of the independent variables  $X_1$  and  $X_2$  into the regression equation (11) for the  $i$ th year, yielding the equation

$$z_i = 6.63 + 0.21413(x_{1i} - 8.79) - 0.0382(x_{2i} - 71.03) \quad (17)$$

A plot of  $Z$  versus  $Y$  for the ten years is given in Figure 8 together with a tabulation of  $z_i$  using Equation (17). This graph demonstrates the ability of the regression equation to predict the dependent variable. That is, if the dependent variable  $Y$  is strongly related to the linear regression of  $Y$  on  $X_1$  and  $X_2$ , then  $z_i$  will be a good prediction of  $y_i$  and the points  $(z_i, y_i)$  will be close to the line  $z = y$ . If, on the other hand,  $Y$  is only weakly related

Figure 8



	Actual	Predicted
1	$y_1$	$z_1$
1	6.50	6.42
2	6.20	6.23
3	7.50	7.49
4	7.10	7.22
5	5.90	5.84
6	6.50	6.65
7	6.60	6.67
8	7.80	7.72
9	6.90	6.95
10	6.30	6.11

to the regression of Y on  $X_1$  and  $X_2$ , then the plot will exhibit a greater spread about the line  $z = y$ . Figure 8 also can be used to divide the YZ plane into two regions divided by the line  $z = y$ . Any point in the region above the line  $z = y$  has an observed value  $y_i$  smaller than predicted by the regression equation while all those points below the line  $z = y$  have an observed  $y_i$  larger than predicted. Therefore, any unusual point can be detected merely by noting those points with the largest deviations from the line  $z = y$ .

Another interesting result as a consequence of Figure 8 comes to light if the product moment correlation coefficient  $r_{YZ}$  is calculated between the variables Y and Z. It should be remembered that the constants  $a_0$ ,  $a_1$ ,  $a_2$  were originally chosen so as to minimize the sum

$$\sum_{i=1}^n (y_i - a_0 - a_1 x_{1i} - a_2 x_{2i})^2 = \sum_{i=1}^n (y_i - z_i)^2 \quad (18)$$

Thus, for the calculation of  $r_{YZ}$ , the solution for the constants a and b from Equation (5) is  $a = 0$ ,  $b = 1$  and because of the relationship given by Equation (18)

$$r_{YZ} = R_{Y(X_1 X_2)} \quad (19)$$

In general, these results state that the multiple correlation coefficient  $R_{Y(X_1 \dots X_k)}$  can be interpreted as the product moment correlation  $r_{YZ}$  between the dependent variable Y and the variable Z, the linear combination of the independent variables  $X_1 \dots X_k$  given by the regression equation.

## 10. PARTIAL CORRELATION COEFFICIENTS

When the relationship between two variables is being determined, quite frequently the true relationship between these two variables is disguised by a common relationship to a third variable. For example, in the problem just considered, it was found that the correlation between yield and temperature was  $-0.8781$ . An examination of the correlation matrix (13) shows that both temperature and yield are strongly related to the third variable, which is rainfall in this case. One might therefore ask the question, "Is the relationship between yield and temperature as strong as indicated by the correlation coefficient, or is this correlation coefficient strengthened by the strong dependence of both variables upon rainfall?"

It does not seem unreasonable to consider the possibility that high rainfall as a cause produces an effect of high yield and low temperature, and that, in reality, temperature only appears to affect yield because of its

Figure 9

Year	Yield			Temperature		
	Actual $y_i$	Predicted by rainfall $\bar{y}+b(x_{1i}-\bar{x}_1)$	$\Delta y_i$	Actual $x_{2i}$	Predicted by rainfall $\bar{x}_2+b'(x_{1i}-\bar{x}_1)$	$\Delta x_{2i}$
1	6.50	6.43	+ .07	72.20	71.75	+ .45
2	6.20	6.19	+ .02	71.40	72.65	-1.25
3	7.50	7.43	+ .07	66.60	68.12	-1.52
4	7.10	7.23	- .13	69.10	68.85	+ .25
5	5.90	5.94	- .04	76.10	73.56	+2.54
6	5.50	5.64	- .14	74.40	74.65	- .25
7	6.60	6.78	- .18	73.30	70.48	+2.82
8	7.80	7.75	+ .05	67.80	66.94	+ .86
9	6.90	6.88	+ .02	68.20	70.11	-1.91
10	6.30	6.04	+ .26	71.30	73.20	-1.90

product moment correlation matrix (13). With a little patience, it can be shown from Equation (4) that

$$r_{\Delta Y \Delta X_2} = \frac{r_{YX_2} - r_{YX_1} r_{X_2 X_1}}{\sqrt{(1 - r_{YX_1}^2)(1 - r_{X_2 X_1}^2)}} \quad (25)$$

Calculation of the partial correlation from Equation (25) is often preferable, since no information beyond the correlation matrix is required.

In the problem under consideration it was shown that 96.6% of the variation in yield was explained by the linear regression of yield on rainfall. Thus, by a subtraction from 100%, 3.4% of the variation in yield was not explained by its linear regression on rainfall. The square of the partial correlation coefficient  $r_{YX_2 \cdot X_1}$  gives the fraction of the unexplained variation in yield which can be explained by the variation in temperature. That is,  $100 \times (-0.5983)^2\% = 35.3\%$  of the variation in yield, unexplained by the linear regression of yield on rainfall, can be explained by temperature variation. Or equivalently, an additional  $100 \times (0.034)(0.353)\% = 1.2\%$  of the variation in yield can be explained by temperature variation over and above what is explained by rainfall. Adding these results together,  $(96.6 + 1.2)\% = 97.8\%$  of the variation in yield can be explained by rainfall and temperature together. It will be remembered that this is exactly the same result that was obtained with the use of the multiple correlation coefficient given by Equation (15). Thus, the relationship which exists between the multiple correlation coefficient and the partial correlation coefficient can be expressed by the equation

$$R_{Y(X_1 X_2)}^2 = r_{YX_1}^2 + (1 - r_{YX_1}^2) r_{YX_2 \cdot X_1}^2 \quad (26)$$

Formulae have been developed here for the partial correlation coefficient for the elimination of the effect of one variable in considering the relationships that exist among a set of variables. The theory can be extended to the elimination of more than one variable, and equations for this process are given by Kendal, (4).

dependence upon rainfall. To examine the true dependence of yield upon temperature one would like to eliminate the effect of rainfall in examining the variation in yield and temperature. It is not immediately obvious how this can be accomplished, because rainfall is not a controllable variable; however, one acceptable way of eliminating the effect of rainfall will now be considered.

Using Equation (3) calculate the regression of Y on  $X_1$  and use the form of the regression equation given by Equation (3a). Define

$$\Delta y_i = y_i - \bar{y} - b(x_{1i} - \bar{x}_1) . \quad (21)$$

$\Delta y_i$  is the difference between the actual yield and that predicted by the regression equation of Y on  $X_1$ . Similarly define  $\Delta x_{2i}$  to be the difference between the actual average temperature and the temperature predicted by the regression of  $X_2$  on  $X_1$ . The quantity  $\Delta x_{2i}$  is given by the equation

$$\Delta x_{2i} = x_{2i} - \bar{x}_2 - b'(x_{1i} - \bar{x}_1) . \quad (22)$$

$\Delta y_i$  and  $\Delta x_{2i}$  designate the variation in yield and temperature, respectively, which cannot be explained by the linear regression on rainfall. These quantities are calculated for this example and are presented in Figure 9.

Calculating now the product moment correlation coefficient between  $\Delta Y$  and  $\Delta X_2$  given by Equation (4), one obtains

$$r_{\Delta Y \Delta X_2} = -0.59830 . \quad (23)$$

This quantity  $r_{\Delta Y \Delta X_2}$  is known as the partial correlation of Y and  $X_2$  with the effect of the variable  $X_1$  removed. A more common notation that is used in most texts for the partial correlation as above is  $r_{YX_2 \cdot X_1}$  yielding the notational identity

$$r_{YX_2 \cdot X_1} = r_{\Delta Y \Delta X_2} \quad (24)$$

It is not necessary to calculate the partial correlation coefficient in the way indicated in Figure 9. Rather, it can be calculated directly from the

#### BIBLIOGRAPHY

Suggested texts for further reading on the topics considered in this article are listed below:

- (1) Brownlee, K. A. *Statistical Theory and Methodology in Science and Engineering*. New York: Wiley, 1965, 2nd edition.
- (2) Draper, N. R., and Smith, A. *Applied Regression Analysis*. New York: Wiley, 1966.
- (3) Hoel, P. G. *Introduction to Mathematical Statistics*. New York: Wiley, 1962.
- (4) Kendall, M. G., and Stuart, A. *The Advanced Theory of Statistics*. Hafner Publishing Company, 1961, II.
- (5) Mendenhall, William. *Introduction to Probability and Statistics*. Belmont, California: Wadsworth Publishing Company, Inc., 1963, 2nd edition.