ED 364 603 TM 020 851

AUTHOR	Murthy, Kavita
TITLE	What Makes r Positive or Negative?: An Exploration of Factors that Affect r with an Emphasis on Insight and Understanding.
PUB DATE	Nov 93
NOTE	21p.; Paper presented at the Annual Meeting of the Mid-South Educational Research Association (22nd, New Orleans, LA, November 9-12, 1993).
PUB TYPE	Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	*Analysis of Covariance; Data Analysis; *Educational Research; Elementary Secondary Education; Higher Education; Influences; Mathematical Models; *Research Methodology
IDENTIFIERS	*Linear Relationships; *Pearson Product Moment Correlation

ABSTRACT

The Pearson product-moment correlation, r, is commonly applied in educational research. Almost all researchers realize that r ranges between -1.00 and +1.00, and that negative coefficients indicate that the bivariate relationship is inverse. Researchers also recognize that the Pearson r only evaluates linear relationship, and is not sensitive to curvilinear relationship. However, few researchers, if pressed, could explain exactly what makes r negative or positive, from a mathematical point of view, even though most researchers know what such results mean. The present paper explores the factors that affect r, including those that impact its sign. The reasons for preferring r over the covariance are explored. Small data sets and graphs are employed to make the discussion concrete. Four figures, four tables. (Contains 8 references.) (Author)

What Makes r Positive or Negative?: An Exploration of Factors that Affect r with an Emphasis on Insight and Understanding

Kavita Murthy

Texas A&M University 77843-4225

Paper presented at the annual meeting of the Mid-South Educational Research Association, New Orleans, LA, November 12, 1993.

ABSTRACT

The Pearson product-moment correlation, r, is commonly applied in educational research. Almost all researchers realize that r ranges between -1.00 and +1.00, and that negative coefficients indicate that the bivariate relationship is inverse. Researchers also recognize that the Pearson r only evaluates linear relationship, and is not sensitive to curvilinear relationship. However, few researchers, if pressed, could explain exactly what makes r negative or positive, from a mathematical point of view, even though most researchers know what such results mean. The present paper explores the factors that affect r, including those that impact its sign. The reasons for preferring r over the covariance are explored. Small data sets and graphs are employed to make the discussion concrete.

Many of the problems of the behavioral sciences go beyond the description of a single variable in its various forms. Rather, most studies within the field of education or psychology are frequently called upon to determine the relationships among two or more variables. For example, college administrators are very concerned with the relationship between high-school grade point averages and Scholastic Aptitude Test scores and performance at college. Do students who do well in high school or who score high on the SAT also perform well in college? Conversely, do poor high-school students or those who perform poorly on the SAT also perform poorly in college?

As soon as one raises questions concerning the relationships among variables, we are thrust into the area of correlation. To express quantitatively the extent to which two variables are related we need to calculate the correlation coefficient. The coefficient of correlation, r, is a statistical summary that represents the degree and direction of relationship between two variables (Glass & Hopkins, 1984). There are many types of correlation coefficients (Haber, Runyon & Badia, 1970). The decision to employ one of them with a specific set of data depends on factors such as: a) the type of scale of measurement in which each variable is expressed, b) the nature of the underlying distribution (continuous or discrete), and c) the characteristics of the distribution of the scores (linear or non-linear). Examples of various correlation coefficients include: a) point biserial, b) Spearman r, and c) Pearson r.

According to Edwards (1973), no matter which correlational technique is used, all have certain characteristics in common: First, two sets of measurements are obtained on the same individuals (or events), or on pairs of individuals who are matched on some basis. Second, the values of the correlation coefficients vary between -1.00 and +1.00. Both extremes represent perfect relationships between the variables and 0.00 represents the absence of a relationship. Third, a positive relationship means that individuals obtaining high scores on one variable tend to obtain high scores on a second variable. The converse is also true; that is, individuals scoring low on one variable tend to score low on a second variable. Fourth, a negative relationship means that individuals scoring low on one variable tend to score high on a second variable. Conversely, individuals scoring high on one variable tend to score low on a second variable. Fifth, a high correlation between variables does not, as such, establish a causal link between variables.

The Pearson product moment correlation, is commonly applied in educational research. Almost all researchers realize that r ranges between -1.00 and +1.00, and that negative coefficients indicate that the bivariate relationship is inverse. Researchers also recognize that the Pearson r only evaluates linear relationship, and is not sensitive to curvilinear relationship. However, few researchers, if pressed, could explain exactly what makes r negative or positive, from a mathematical point of view, even though most researchers know what such results mean. The present paper explores the factors that affect r, including those that impact its sign. The reasons for preferring r over the covariance are explored. Small data sets and graphs are employed to make the discussion concrete.

Pearson's Product-moment Correlation Coefficient

The most commonly used statistical index for the relationship between two variables is the Pearson product-moment correlation coefficient, which is sometimes called the correlation coefficient, correlation, or intercorrelation (Allen &Yen, 1979). The symbol for a sample correlation coefficient for variables X and Y is $r_{xy.}$ Sample correlations are defined using the following formula:

$$r = \frac{\left(\sum\left(X-\overline{X}\right)\left(Y-\overline{Y}\right)\right) \big/ (n-1)}{\overline{)sd_x \bullet sd_y}}$$

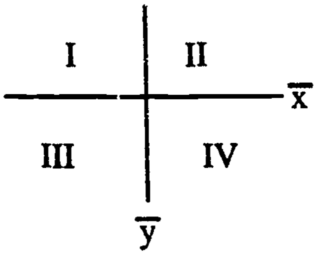
The numerator in this formula for the correlation is called the covariance, and is the average product of the deviations in X and Y, where a deviation is a distance from the mean. By multiplying the deviation of each individual's score from the mean of the X-variable by its corresponding deviation on the Y-variable and then summing and averaging the cross products, yields the covariance (Glass &Hopkins, 1984). The denominator in the formula is the product of the standard deviations of X and Y. The standard deviation is a measure of variability (Allen & Yen, 1979), and is defined as the square root of the sum of the squared deviations from the mean divided by the number of scores you have, minus one (for sample statistics). The formulas for the standard deviation of X and Y are:

$$sd_x = \sqrt{\frac{\sum\left(X-\overline{X}\right)^2}{N\text{-}1}}$$

$$sd_y = \sqrt{\frac{\sum\left(Y-\overline{Y}\right)^2}{N\text{-}1}}$$

It is important to note that the standard deviation can never be negative. The standard deviation is really the square root of the variance, which is a squared statistic. By squaring the deviations from the mean, and then summing them, the variance has eliminated the impact of a negative sign on the denominator portion of the correlation coefficient calculations. Thus, covariance and r for a given data set always have the same sign.

The most common way to visually represent the relationship between two variables is by using a scatter plot. Each point on this plot represents a pair of scores for each case, or individual. By plotting these points on a Cartesian plane, along both the horizontal or X-axis (abscissa) and the vertical or Y-axis (ordinate), it is possible to actually see whether you have a positive or inverse relationship between variables X and Y. The Cartesian plane divides the graph into four distinct quadrants using the mean score on Y to define a horizontal line and the mean score on X to define a vertical line. Quadrants one and two are located above the X-axis, and quadrants three and four fall below the X-axis, and are listed from left to right.:



The covariance included in the correlation formula determines in which quadrant the scores will lie. Suppose, that people who score above the mean on variable X also score above the mean on variable Y. These people will be located in quadrant II, will have positive deviation scores, and their cross-products will be positive. Similarly, the people who score below the mean on both variables will have negative deviation scores, and their cross-products will be positive as well. As a result the scores for these examples will fall either in quadrant II or quadrant III, and the Pearson r for these examples will be positive.

When the scores are above the mean on one variable and below the mean on another variable, then the product of the two deviation scores will be negative, the numerator of the correlational formula will be negative, the scores will fall either in quadrant I or IV, and the Pearson r will be a negative number.

Finally, if the scores above the mean in the X-variable are approximately equally likely to be associated with scores above and below the mean on the Y-variable, then some of the cross-products on X and Y will be negative and some will be positive, causing the numerator to be near zero, which in turn leads to a near-zero correlation. In this case, the scores will lie in all four quadrants of the Cartesian plane.

Reasons for Standardizing the Covariance into r

While the covariance alone determines where the scores will lie, one might ask, "Why don't we just use the covariance and forget Pearson r ?" The reason for choosing Pearson r over the covariance is twofold. First, the covariance has no maximum or minimum scores and is heavily influenced by the linear relationship of X and Y. The Pearson r, on the other hand, is scaled, and has maximum (+1.00) and minimum (-1.00) cutoff points under which the correlation score must fall. Another reason for preferring the Pearson r over the covariance is that the covariance is influenced by the "spreadoutness" of X and Y. The Pearson r accounts for this by dividing by the standard deviation of both X and Y, thus eliminating the effect of the "spreadoutness" of X and Y.

Correlation coefficients are described in terms of their sign and their size. The sign of the correlation reflects the direction of the relationship, whereas the size of the correlation, which can vary from zero to one, reflects the strength of the relationship (Glass & Hopkins, 1984). The strength of the relationship translates into how well one variable can be predicted from another. The size of the correlation can be considered as a measure of how well the points in the scatter plot "hug a line". This line is called a regression line, and is calculated through the use of the following formula: y= a + bx , in which X and Y represent variables that change from individual to individual, and a and b represent constants for a particular set of data. More specifically, b represents the slope of a line relating values of Y to values of X. This is referred to as the regression of Y on X (Runyon & Haber,1988). The correlation coefficient is also related to this equation in that $b = r\left(\frac{sd_y}{sd_x}\right)$. From this formula, it can be shown that the correlation coefficient is actually a weight within the regression equation, and will influence where the line is drawn in the scatter plot.

When constructing regression lines, it is possible to see that the regression line will not pass through all the paired scores, except when r= +1.00 or r= -1.00. Otherwise, the regression line will pass among the paired scores in such a way as to minimize the squared deviations between the regression line (predicted scores) and the obtained scores. In conceptualizing the relationship between the regression lines and the magnitude of r, it might be helpful to think of the regression lines as rotating about the joint means of X and Y. When r= ±1.00, the regression line will pass directly through all the paired scores. However, as r becomes smaller, the regression line will rotate away from the "perfect" line of best fit, so that in the limiting case, when r=0, the line will become parallel to the x-axis. At this point the regression line for predicting Y from known values of X for all subjects will yield the prediction that each subject scored the mean of Y.

Heuristic Examples

To make the discussion of correlation and regression lines more concrete, small, hypothetical data sets have been created to demonstrate the effects of positive and negative scores on the correlation. Table 1 demonstrates that a positive relationship exists between the two variables, X and Y. The sum of the cross-products, or the covariance is a positive number, (414), and therefore the correlation is positive. Also, the quadrants have been calculated as well. For Data set #1, the paired scores for X and Y fall either in quadrant two or three. Figure 1 graphically represents this.

Insert Table 1 and Figure 1 about here

The regression line for Data set #1 is nearly "perfect", as reflected by the strength or magnitude of the correlation coefficient, (.9628). Most of the individual pairs of scores are "caught" by the regression line. Notice that the standard deviations of both X and Y are equal. This makes the regression line much easier to calculate. The regression line is simply the Y-intercept ; ˋus the correlation times the value of X. Because the mean of both variables in this data set is 0, the quadrants are delineated by the Y and the X axes. Because the standard deviations are equal, the slope will be equal to the correlation coefficient, which in this case would be, .9628. Additionally, the covariance reduces simply to (X*Y), since the means for both X and Y are 0. For all of the following data sets, the means will be equal to 0 and the standard deviations of X and Y will be equal to each other, to simplify the discussion without loss of generality.

For Data set #2, in Table 2, the scores have an inverse relationship. The pairs of scores fall in either quadrant I or IV, and the correlation coefficient becomes negative. Because the sum of the cross-products (X*Y) is a negative number, the sign of the correlation coefficient is negative. Again, the correlation coefficient is high (-.9710), and Figure 2 demonstrates the strength of this relationship. Most of the points "hug" the regression line as well.

Insert Table 2 and Figure 2 about here

In Table 3, the scores reflect a low positive correlation (.2174). Most of the scores lie in quadrant II and III, however one case is in quadrant I and another case lies in quadrant IV.

Insert Table 3 and Figure 3 about here

Notice in Figure 3 that the regression line does not "catch" any points directly. This is partly due to the two outlier scores in quadrants I and IV. If these scores had not existed, the sum of the cross-products would have been much higher, (28), thus yielding a "stronger" correlation. In fact, the correlation would have been approximately .85. This illustrates how much the correlation and regression equation is influenced by each case.

Each individual score holds a certain amount of "weight", and directly impacts that calculation of the covariance and correlation. An example of how much these scores influence the outcome of these calculations is demonstrated in the last example.

Just like the data in Table 3, Table 4 shows that most of the scores lie in quadrant II or III. But this time, the regression line is drawn through quadrants I and IV! The regression line has actually flipped and become a strong inverse correlation (-.8228) rather than a moderate positive correlation.

Insert Table 4 and Figure 4 about here

Upon closer inspection of the data, we find that in Data set #4, the two outlier cases, are extremely high in value as compared to the other cases. As a result, cases one and two completely "take over", so to speak, the calculation of the covariance, and turn the sum of the cross-products into a large negative number. From a mathematical perspective, it becomes clear why the line of best fit rotates into its new position. Scores farther from the Cartesian coordinate for the two means exert more influence on the numerator of the correlation coefficient, because the influence of each pair of scores is a weighted function of the distance of the scores from the group means. This rotation could very well mislead the researcher into believing that there is a strong inverse relationship between all the variables on X with all the variables on Y, when in fact, a more accurate description of the data would yield a moderate positive correlation.

Other Factors the Affect r

The correlation coefficient is also influenced by many other factors, not otherwise inherent in the formula for correlation itself. Attenuation influences on *r* include departure from linearity, departures from both variables being similarly distributed, using instruments with lower reliability, and using data in which either variable has a restricted range (Dolenz-

Walsh, 1992). First, If X and Y have any degree of curvilinear relationship, the value of r will underestimate the true degree of relationship between the two variables (Glass & Hopkins, 1984). Second, departures from similar distribution shapes can result in conservative underestimates of relationship. Therefore unless two variables have exactly the same distribution, it is simply not possible to obtain a perfect Pearson correlation between the two variables (Nunnally, 1967). Third, Measurement error lends to the attenuation of the Pearson r (Busby & Thompson, 1990). Reliability coefficients of the two variables being correlated establishes a ceiling for the correlation coefficient for a given data set. For this reason, it is important to assess the reliability of the scores in hand on both variables being correlated. Finally, the variance of a sample heavily influences the correlation (Glass & Hopkins, 1984). If a broader range of subjects is studied, the correlation will increase; if a narrower range of subjects is studied, the correlation will decrease. Subject pools that are homogeneous underestimate the magnitude of the relation between the variables and represent a restricted sample (Allen & Yen, 1979).

Conclusions

The Pearson product-moment correlation coefficient is an integral part of educational research. Given the importance of Pearson r, it remains necessary to understand the many factors that affect r. The present paper has explained from a mathematical perspective what makes r positive or negative, with an emphasis on insight and understanding. The reasons for preferring r over the covariance were explored. Small heuristic data sets and graphs were employed to make the discussion concrete.

References

Allen, M.J. & Yen, W.M. (1979). Introduction to measurement theory. Monterey, CA: Brooks/Cole Publishing Company.

Busby, D., & Thompson, B. (1990, January). Factors attenuating Pearson's r: A review of basics and some corrections. Paper presented at the annual meeting of the Southwest Educational Research Association, Austin, TX.

Dolenz-Walsh, B. (1992, January). Factors that attenuate the correlation coefficient and its analogs. Paper presented at the annual meeting of the Southwest Educational Research Association, Houston, TX.

Edwards, A.L. (1969). Statistical analysis. NewYork: Holt, Rinehart and Winston.

Glass, G.V. & Hopkins, K.D. (1984). Statistical methods in education and psychology. (2nd ed.). Englewood Ciiff, NJ: Prentice-Hall, Inc.

Haber, A., Runyon, R.P., & Badia, P. (1970). Readings in statistics. Reading, MS: Addison-Wesley.

Nunnally, J.C. (1967). Psychometric theory. New York: McGraw-Hill Book Company.

Runyon, R.P. & Haber, A. (1988). Fundamentals of behavioral statistics. (6th ed.) New York: Random House.

Table 1

Data Set #1

ID	X	Y	X*Y	Quadrant	Regression Points
1	3	3	9	2	2.8884
2	9	5	45	2	8.6651
3	5	9	45	2	4.8140
4	10	10	100	2	9.6279
5	-3	-3	9	3	-2.8884
6	-5	-5	25	3	-4.8140
7	-9	-9	81	3	-8.6651
8	-10	-10	100	3	-9.6279
Sum	0	0	414		
Count	8	8	8		
Mean	0.0000	0.0000	51.7500		
Std. Dev.	7.8376	7.8376	37.7463		
Pearson r	0.9628		Y-intercep	0	

Table 2
Data Set #2

ID	X	Y	X*Y	Quadrant	Regression Points
1	-1	1	-1	1	-.9710
2	-4	4	-16	1	-3.884
3	-4	6	-24	1	-3.884
4	-6	4	-24	1	-5.826
5	1	-1	-1	4	.9710
6	4	-4	-16	4	3.884
7	4	-4	-16	4	3.884
8	6	-6	-36	4	-5.826
Sum	0	0	-134		
Count	8	8	8		
Mean	0.0000	0.0000	-16.75		
Std. Dev.	4.4401	4.4401			
Pearson r	0.-.9710		Y-intercep	0	

Table 3
Data Set #3

ID	X	Y	X*Y	Quadrant	Regression Points
1	3	-3	-9	2	.6522
2	-3	3	-9	2	-.6522
3	-3	-3	9	2	-.6522
4	-2	-2	4	2	-.4348
5	-1	-1	1	3	-.2174
6	1	1	1	3	.2174
7	2	2	4	3	.4348
8	3	3	9	3	.6522
Sum	0	0	10		
Count	8	8	8		
Mean	0.0000	0.0000	1.2500		
Std. Dev.	2.5635	2.5635	7.0255		
Pearson r	0.2174		Y-intercep	0	

Table 4
Data Set #4

ID	X	Y	X*Y	Quadrant	Regression Points
1	-12	12	-144	1	9.8736
2	12	-12	-144	4	-9.8736
3	-3	-3	9	3	2.4684
4	-2	-2	4	3	1.6456
5	-1	-1	1	3	.8228
6	1	1	1	2	-.8228
7	2	2	4	2	-1.6456
8	3	3	9	2	-2.4684
Sum	0	0	-260		
Count	8	8	8		
Mean	0.0000	0.0000	-32.5		
Std. Dev.	6.7188	6.7188	68.8871		
Pearson r	-.8228		Y-intercep	0	

Figure 1.

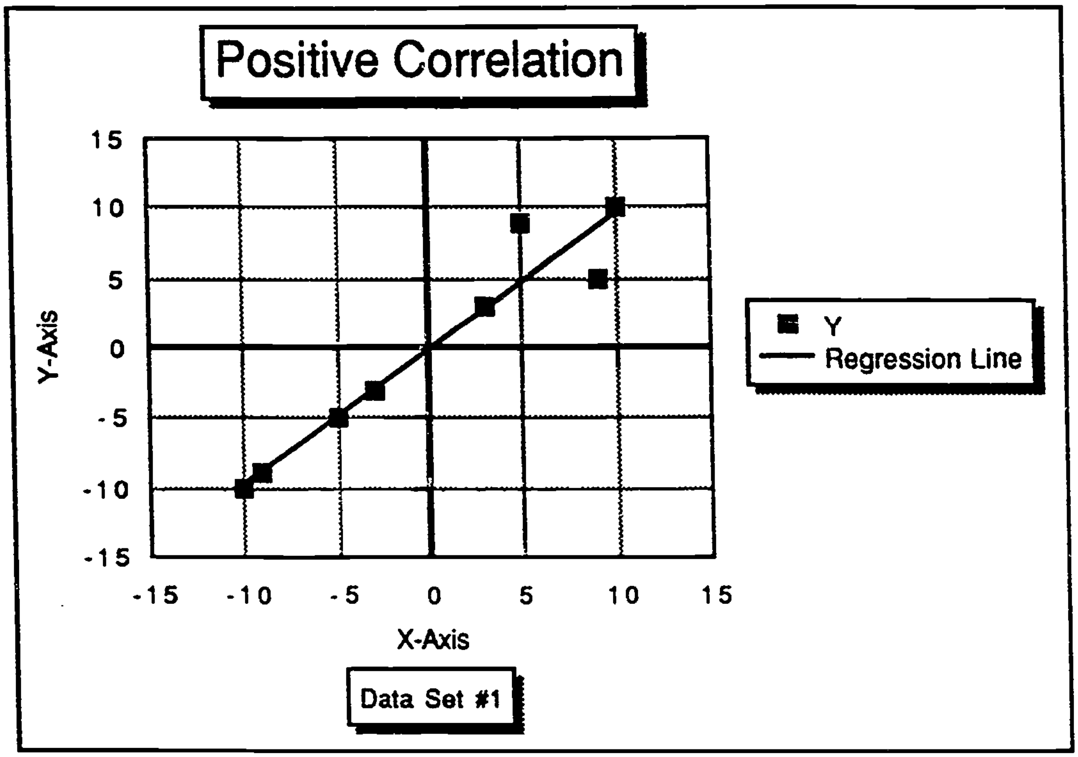

Positive Correlation
15
10
5
0
-5
-10
-15
Y-Axis
-15
-10
-5
0
5
10
15
X-Axis
■ Y
— Regression Line
Data Set #1

Figure 2.

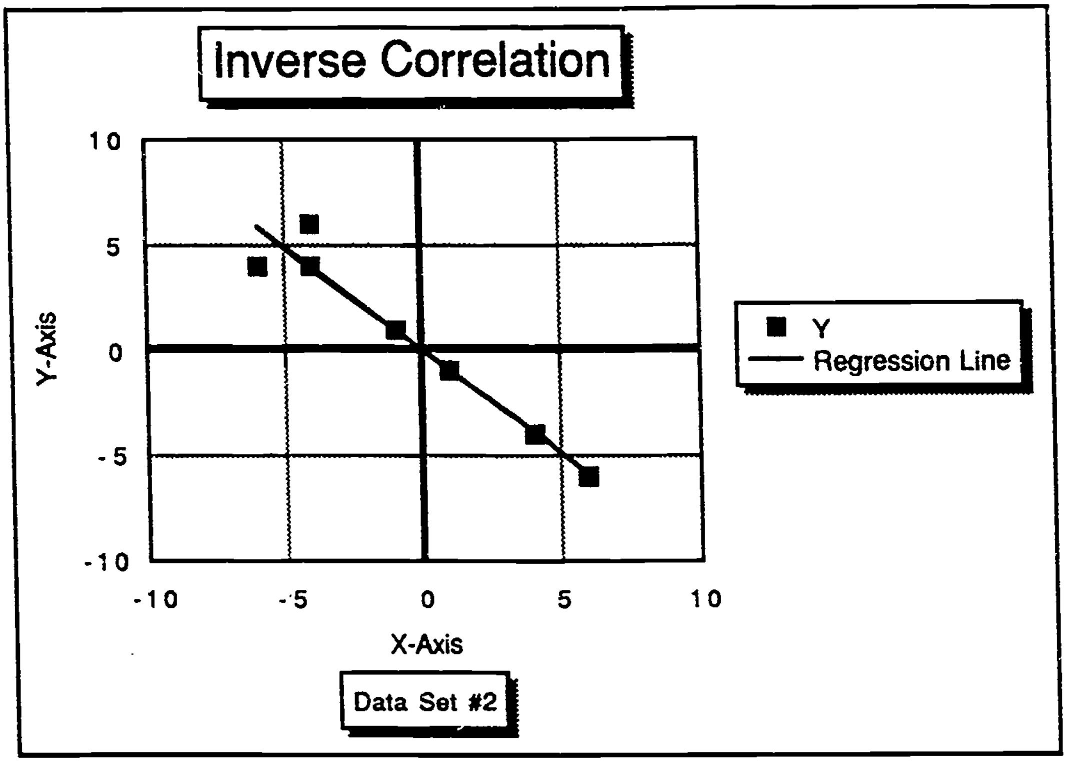

Inverse Correlation
10
5
0
-5
-10
Y-Axis
-10
-5
0
5
10
X-Axis
■ Y
— Regression Line
Data Set #2

Figure 3.

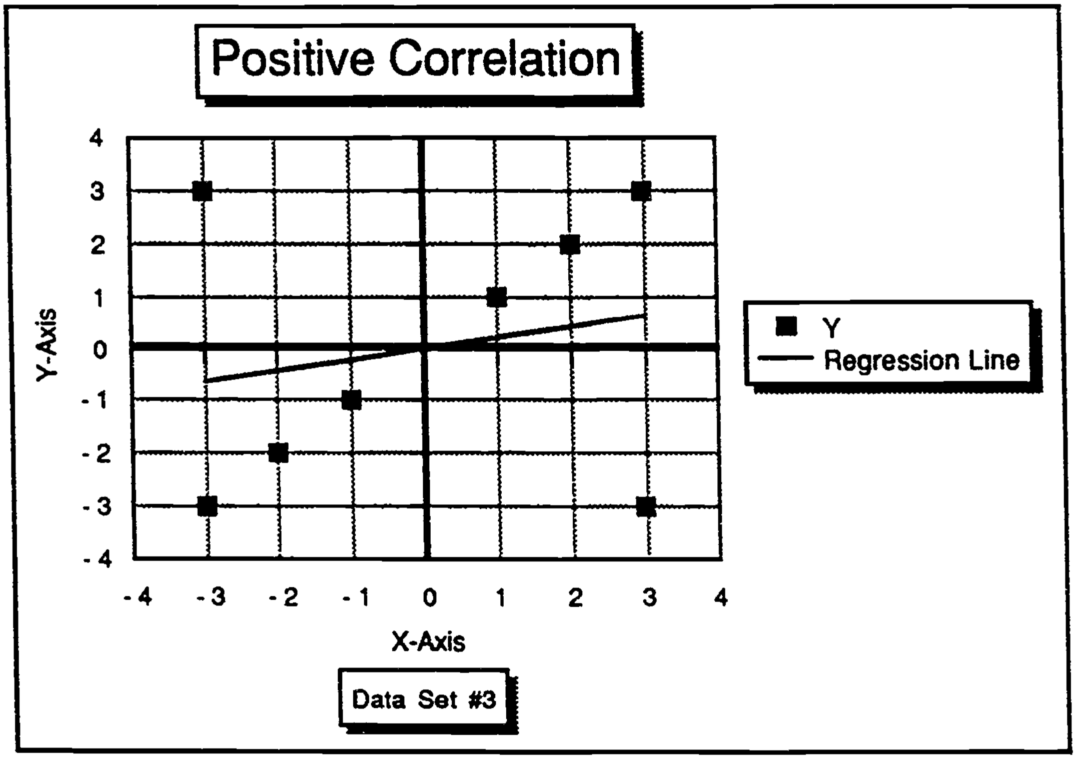
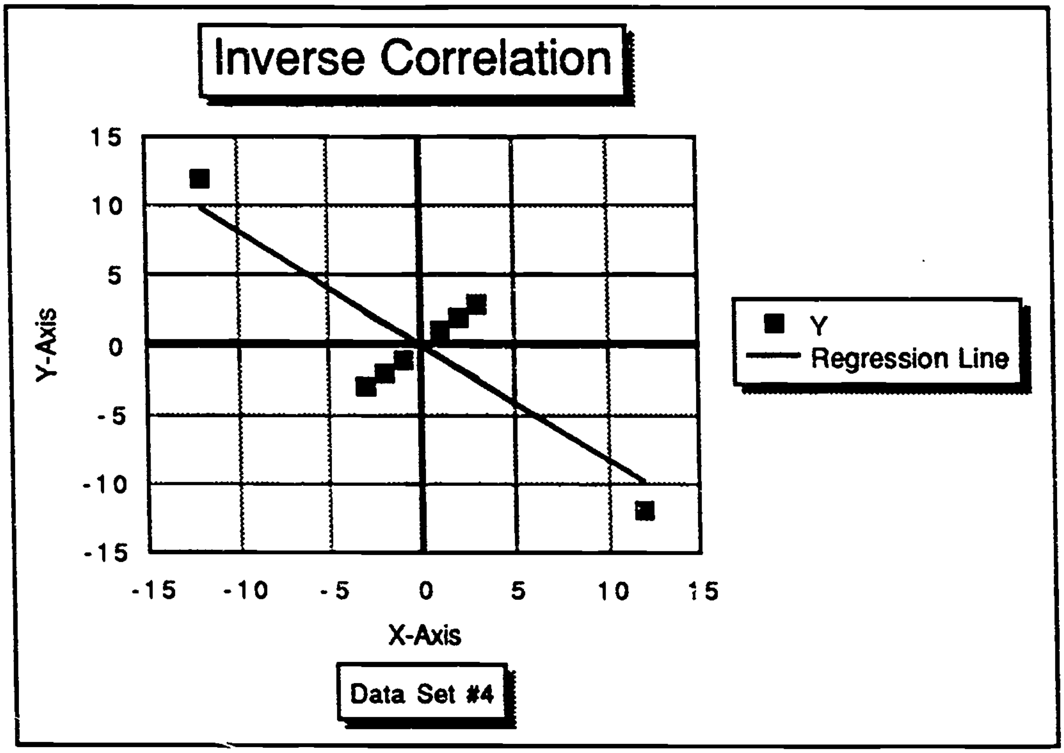

Positive Correlation
4
3
2
1
0
-1
-2
-3
-4
Y-Axis
-4
-3
-2
-1
0
1
2
3
4
X-Axis
■ Y
— Regression Line
Data Set #3

Figure 4.


Inverse Correlation
15
10
5
0
-5
-10
-15
Y-Axis
-15
-10
-5
0
5
10
15
X-Axis
■ Y
— Regression Line
Data Set #4