PSYCHOMETRIKA—VOL. 11, NO. 1
MARCH, 1946

A NEW DESCRIPTIVE STATISTIC: THE PARABOLIC CORRELATION COEFFICIENT

CHARLES C. PETERS
PENNSYLVANIA STATE COLLEGE

This article proposes a new descriptive statistic related to the second order parabola in the same manner in which the familiar correlation coefficient is related to the regression coefficient. The parabolic r describes in standard terms simultaneously the general trend of the regression and the extent and nature of its curvilinearity, and is relatively easy to compute and easy to communicate. Formulas for its standard error and its limits are derived and it is applied to a number of regressions.

It is a well-known fact that the correlation coefficient expresses the slope of a *straight* regression line on a squeezed scale, and that it is an inappropriate and misleading descriptive statistic where the relation between two sets of paired variates is not rectilinear. Nevertheless it has been recklessly used without regard to the rectilinearity of the regression. That has been in part due to the absence of any convenient descriptive statistic for curvilinear correlational relationships. We have long had, of course, the equations of parabolas of various orders and also a number of other forms of curves. But these have not been in the form of standard units so as to make their descriptions of the regressions readily communicable. It is the purpose of this paper to describe a new statistic related to the equation of the second-order parabola in the same manner as the correlation coefficient is related to the regression coefficient, to give machinery for translating it into standard meaning, and to develop for it a standard error formula and a proof of its limits. I have named this statistic *the parabolic correlation coefficient.* It provides a dual description of the regression: (1) an index of the slope (i.s.) exactly equal in numerical value to the correlation coefficient; and (2) an index of corvilinearity (i.c.) The index of slope is the first derivative of the equation of the second-order parabola evaluated at a certain point to be described in this article, while the index of curvilinearity is one half of the second derivative, each adjusted to a standard meaning by a suitable squeezing of the scale.

The equation for the second degree parabola is

$$Y = a + bX + cX^2. \tag{1}$$

We wish to get values for the a, b, and c such that the sum of the squares of the errors made in predicting Y-scores with this equation will be a minimum. In the case of any individual with an X-score in the class i the error would be

$$(Y_i - \hat{Y}_i) = Y_i - (a + bX_i + cX_i^2).$$

We want the sum of the squares of these errors to be a minimum. Hence we square the expression at the right and sum it termwise. But we shall write the terms in a different order, merely for convenience and, for the sake of simplicity, shall drop the subscripts. The sum of the squares of the errors is:

$$\begin{aligned}\sum Y^2 + c^2\sum X^4 + b^2\sum X^2 + Na^2 - 2c\sum X^2Y - 2b\sum XY \\ - 2a\sum Y + 2ac\sum X^2 + 2ab\sum X + 2bc\sum X^3.\end{aligned}$$

In order to determine values for a, b, and c that will make this expression for the sum of the squares of the errors a minimum, we must equate to zero the partial derivatives with respect to these three terms. These partial derivatives are in order (each equated to zero and each divided through by 2):

$$\begin{aligned} c\sum X^4 + b\sum X^3 + a\sum X^2 - \sum X^2Y &= 0, \\ c\sum X^3 + b\sum X^2 + a\sum X - \sum XY &= 0, \\ c\sum X^2 + b\sum X + Na - \sum Y &= 0. \end{aligned}$$

This set of simultaneous equations must be solved for the three unknowns, a, b, and c. A number of different outcomes, all completely equivalent in value, are possible according to the method of handling the algebra. One of the most convenient solutions is the following:

$$c = \frac{(\Sigma X^2 \Sigma Y - N\Sigma X^2 Y)(N\Sigma X^2 - [\Sigma X]^2) + (N\Sigma XY - \Sigma X \cdot \Sigma Y)(N\Sigma X^3 - \Sigma X^2 \Sigma X)}{(N\Sigma X^3 - \Sigma X^2 \Sigma X)^2 + ([\Sigma X^2]^2 - N\Sigma X^4)(N\Sigma X^2 - [\Sigma X]^2)}. \quad (2)$$

$$b = \frac{(N\Sigma XY - \Sigma X \cdot \Sigma Y) - c(N\Sigma X^3 - \Sigma X^2 \Sigma X)}{N\Sigma X^2 - [\Sigma X]^2}. \quad (3)$$

$$a = \frac{\Sigma Y - c\Sigma X^2 - b\Sigma X}{N}. \quad (4)$$

These are the coefficients which are now to be put in equation (1) for predicting Y scores.

But so far we have an operative tool, not a descriptive statistic. We shall turn next to getting the latter. The reader should get in mind a conventional correlation chart laying out the independent scores on the X-axis and the dependent scores on the Y-axis. We are concerned with two features about such outlay: (1) What is the general trend of the Y scores as we go up along the X-axis? and (2) To what extent is the regression line curved rather than straight, and in what direction is it curved? We can get an account of the first feature from the first derivative of the equation of the parabola and an account of the second feature from the second derivative.

The first derivative of equation (1) is $b + 2cX$. This gives the slope of a tangent to the curve at any point at which the derivative is evaluated. Its value is dependent upon its place along the X-axis, as is evidenced by the fact that it contains an X. If we evaluate it near the middle of the X-distribution, we get a tangent that will show the general trend pretty well for the whole distribution; and if, in addition, we make a proper adjustment in our scales, we can have a description of the general trend that compares in value with the coefficient of correlation. We shall adjust our scales by stretching or squeezing them so as to make the units equal in terms of variability on the two axes. This we accomplish by multiplying any increments along the X-axis (here the first derivative) by σ_x/σ_y. But we shall put this in summation form in the first term since we have all the other terms in summation form, but shall, for convenience in writing, let it stand in σ form in the other two terms. So re-writing our first derivative thus multiplied by the standard deviation ratio and separating it into three terms, and at the same time putting the $2c$ over the common denominator, and also substituting NM_x for $\sum X$, we have

$$i.s. = \frac{N\sum XY - \sum X \cdot \sum Y}{\sqrt{(N\sum X^2 - [\sum X]^2)\,(N\sum Y^2 - [\sum Y]^2)}} - \frac{c(N\sum X^2 - NM_x\sum X^2)}{N^2\,\sigma_x\,\sigma_y} + \frac{2cXN^2\,\sigma_x^2}{N^2\,\sigma_x\,\sigma_y}. \tag{5}$$

Inspection of the above expression will show that, if we evaluate it where the second and third terms aggregate exactly zero, we shall have left precisely the formula for the Pearsonian correlation coefficient. Since the denominators are the same, that will be where

$$2cXN^2\sigma_x^2 = c(N\Sigma X^3 - NM_x\Sigma X^2).$$

Solving this for X, we get

$$X = \frac{\Sigma X^3 - M_x\Sigma X^2}{2N\sigma_x^2}. \tag{6}$$

So if we evaluate the first derivative at the position on the X-axis where X has the value indicated in (6) on our adjusted scale, we shall always get for this part of our descriptive statistic exactly the numerical value of the coefficient of correlation. Investigation of this expression will show that it reduces to the mean in the case of a symmetrical distribution of the X-variable; but we can easily get for the needed critical point at which to evaluate it a more generalized value.

The expression as it stands in (6) is in score form. In order to investigate its properties we shall transform it to deviation form. The capital X represents a score and the lower case x a deviation from the mean of the X's. So

$$X = x + M_x.$$

Hence

$$\begin{aligned} \Sigma X^3 &= \Sigma x^3 + 3M_x\Sigma x^2 + 3M_x^2\Sigma x + NM_x^3; \\ M_x\Sigma X^2 &= M_x\Sigma x^2 + 2M_x^2\Sigma x + NM_x^3. \end{aligned} \tag{7}$$

Subtracting the second equation in (7) from the first to get the numerator of (5), remembering that no matter what the shape of the distribution $\Sigma x = 0$, and simplifying, we have

$$X = \frac{2M_x\Sigma x^2 + \Sigma x^3}{2N\sigma_x^2} = M_x + \tfrac{1}{2}\sigma_x\sqrt{\frac{(\Sigma x^3)^2}{N^2\sigma^6}}. \tag{8}$$

The term under the radical above is, in the Pearson terminology, β_1. But because the sum of the odd powers may have either the plus or minus sign before squaring, and because we wish to preserve the effect of the sign, we shall write the equation as follows:

$$X = M_x \pm \tfrac{1}{2}\sigma_x\sqrt{\beta_1}. \tag{9}$$

For a normal distribution, or for any symmetrical distribution, β_1 is zero. So, in order to get for this part of our descriptive statistic a value exactly equal to the coefficient of correlation, we must evaluate it exactly at the mean if the X-distribution is normal or otherwise symmetrical. If the distribution is not symmetrical, the critical point would be near the mean but a little off to the extent involved in (9). But in the computational work the worker need not be concerned about that. If he will merely compute

$$i.s. = \frac{N\Sigma XY - \Sigma X \cdot \Sigma Y}{\sqrt{(N\Sigma X^2 - [\Sigma X]^2)\,(N\Sigma Y^2 - [\Sigma Y]^2)}}$$

it will follow as an automatic consequence that he has evaluated the derivative at the position indicated in (9) and that he has the slope of a tangent to the curve exactly parallel to the fitted line the slope of which would be the coefficient of correlation. It would, of course, be plausible always to evaluate the derivative at the mean, but then our value for the *i. s.* would not always have a familiar meaning. It would seem worth while to sacrifice that point for the practical advantage of continuing to carry in our parabolic correlation coefficient an element exactly equal to the well-known Pearsonian r.

The above discussion provided an index of slope for our descriptive statistic. We need for our second element an index of the curvilinearity. This requires the second derivative. The second derivative of equation (1) is $2c$. But for certain practical reasons discussed below we shall take for our *i.c.* half of that derivative, namely c. But we also want it expressed in terms of our adjusted scales in order that it may have a standard meaning for all types of distributions to which it is applied. That means that we must multiply it by the ratio σ_x^2/σ_y. When this multiplier is put in summation terms and used with equation (2) we get for the *i.c.* equation (10a) below. When used with another of the three possible solutions for c, we get equation (10b) below. In practice it is desirable to employ both of these formulas as an arithmetic check. They should give identical outcomes.

$$i.c. = \frac{(\Sigma X^2 \Sigma Y - N\Sigma X^2 Y)(N\Sigma X^2 - [\Sigma X]^2) + (N\Sigma XY - \Sigma X \cdot \Sigma Y)(N\Sigma X^3 - \Sigma X^2 \Sigma X)}{(N\Sigma X^3 - \Sigma X^2 \Sigma X)^2 + ([\Sigma X^2]^2 - N\Sigma X^4)(N\Sigma X^2 - [\Sigma X]^2)} \cdot \frac{(N\Sigma X^2 - [\Sigma X]^2)}{N\sqrt{N\Sigma Y^2 - [\Sigma Y]^2}}. \quad (10a)$$

$$i.c. = \frac{(\Sigma X^2 Y \cdot \Sigma X - \Sigma X^2 \Sigma XY)(N\Sigma X^2 - [\Sigma X]^2) + (N\Sigma XY - \Sigma X \cdot \Sigma Y)([\Sigma X^2]^2 - \Sigma X^3 \Sigma X)}{(\Sigma X^4 \Sigma X - \Sigma X^3 \Sigma X^2)(N\Sigma X^2 - [\Sigma X]^2) + (N\Sigma X^3 - \Sigma X^2 \Sigma X)([\Sigma X^2]^2 - \Sigma X^3 \Sigma X)} \cdot \frac{(N\Sigma X^2 - [\Sigma X]^2)}{N\sqrt{N\Sigma Y^2 - [\Sigma Y]^2}}. \quad (10b)$$

The reader may be alarmed at the elaborateness of this formula for the index of curvilinearity. Its application does, it must be admitted, require some work. But it will be observed that the same quantities recur in it, in consequence of which the labor is less than would at first appear. The quantities in all of the parentheses lend themselves to very economical procedures on a calculating machine. There are needed, besides the N, the following moments and product moments:

$$\Sigma X$$
$$\Sigma Y$$
$$\Sigma X^2$$
$$\Sigma Y^2$$
$$\Sigma XY$$
$$\Sigma X^2 Y$$
$$\Sigma X^3$$
$$\Sigma X^4$$

The first five of these are needed in the computation of a coefficient of correlation; only the last three are additional. They can be very easily and quickly obtained when working with a correlation chart, though no correlation chart is really needed for this statistic as it is for the correlation ratio and the Blakeman Test. Getting them on a calculating machine requires going through the scores one additional time as compared with the correlation coefficient, and the total time to calculate a parabolic r from raw scores on a calculating machine is about twice that required for a Pearsonian coefficient of correlation!* The needed moments and product moments can also be obtained on the I.B.M. machines, but this requires that the X^2 values be punched in the cards as well as the X values.

So our parabolic correlation coefficient will contain always two terms, one giving in familiar form an index of the general slope and the other giving an index of the form and the degree of curvilinearity. If the regression is really rectilinear, the second element will be zero and the parabolic correlation coefficient will automatically reduce to the Pearsonian r. In any case the parabolic correlation coefficient will automatically carry notice of the degree of departure from rectilinearity, and we can, if we wish, test that departure for statistical significance by the method explained below. A parabolic r of, say, $-.163 - .167$ means that there is in our relation a moderate over-all downward trend (i.s. $= -.163$) and there is an upward hump of moderate amount in the curve (i.c. $= -.167$). A parabolic r of $-.163 +$

* Convenient worksheets for the parabolic correlation coefficient may be obtained from the Department of Psychology and Education of the Pennsylvania State College at one cent each.

.675 would mean that there is a general downward trend and also a distinct downward hump. Thus the nature of the regression can be visualized from the statement of the parabolic r without the need for a correlation chart.

What are the maximum and minimum limits between which our index of curvilinearity lies? My proof of these limits rests upon two postulates: the *i.c.* can be a maximum or a minimum only (1) when the index of slope is zero, and (2) when the regression law operates perfectly to place each Y-value exactly on the curve when allocated to its appropriate position along the X-axis. Hence, where σ_y is the standard deviation of the actual Y-values and σ_{y_c} is the standard deviation of the Y's on the regression line,

$$\sigma_y = \sigma_{y_c} = \left[\frac{\Sigma(a + bX + cX^2)^2}{N} - M_y^2\right]^{\frac{1}{2}}.$$

When this expression is ~~evaluated it~~ results in moments in terms of ΣX^4, ΣX^3, ΣX^2, and ΣX. These are in score form; they can be more easily simplified if transformed to deviation form. Let $X = x + M_x$, where the lower case x is a deviation from the mean of the X's. Raising this to the 4th power, summing, and dividing by the N of the denominator above yields terms in $\Sigma x^4/N$, $\Sigma x^3/N$, $\Sigma x^2/N$, and $\Sigma x/N$. The last two are, of course, σ_x^2 and zero. The first is $\beta_2\sigma_x^4$ and the second is $\sigma_x^3\sqrt{\beta_1}$. The X^3 can be correspondingly treated. When these values are substituted, and also values for a and for b from equations (3) and (4), and cognizance is taken of the fact that the *i.s.* must be zero, the whole expression reduces to

$$\sigma_y = c\sigma^2\sqrt{\beta_2 - \beta_1 - 1}.$$

The algebra involved in this is straightforward substitution but to carry it through requires several pages, which I do not deem it necessary to consume here, because anyone who is interested can carry through the proof himself from the above hints. We now substitute this value for the denominator in formula (10). The σ_x^2 is canceled by the numerator of the same fraction. The c standing in the substituted denominator here is canceled by the lengthy fraction constituting the left-hand factor of equation (10), which is exactly c. Hence the whole reduces to the very simple form

$$i.c.^{\text{max.}} = \frac{1}{\sqrt{\beta_2 - \beta_1 - 1}}.$$

The minimum value is had by using the negative sign with the square root. The β's here are the Pearsonian measures of kurtosis and of skewness, respectively, of the X-variable (the independent variable).

For a normal distribution ($\beta_2 = 3$ and $\beta_1 = 0$) the maximum limit of the *i.c.* would be .707. . . . But for a rather platykurtic distribution the *i.c.* could reach ± 1.00 or even, though rarely in practice, exceed those limits. It is because of the desirability of pairing two elements in our parabolic r that are confined to substantially the same limits that we chose half of the second derivative for our *i.c.* instead of the second derivative as it stood. It is, perhaps, a further justification that the *i.c.* then becomes just the c of the parabolic equation, though on a squeezed scale.

Statistical workers accustomed to the conventional testing for reliability will ask for standard error formulas for this statistic. Since the parabolic correlation coefficient is merely a descriptive statistic and has no functional meaning as a whole, we do not need a standard error formula for it as a whole, though one could be developed if wanted. We are interested only in testing the reliability of its two parts. For testing the hypothesis of a true slope of zero, the standard error of the index of slope where we have evaluated it is merely that long familiar for testing the corresponding hypothesis in the correlation coefficient:

$$S.E._{i.s.} = \frac{1}{\sqrt{N-1}}.$$

But we desire a formula for the standard error of the index of curvilinearity. Deming* has shown that the sampling variance of the coefficient of X^2 in a second-order parabola, c, is

$$S.E._c^2 = \frac{\sigma_y^2}{(N-3)\,(\sum x^4/N - \sigma^4)}, \tag{11}$$

where σ_y^2 is the variance of the Y-array in the sample and the x's in the denominator are deviations from the X-mean. But that form would be too inconvenient in use. We want a form that can be easily applied mentally as soon as we hear the index of curvilinearity and know the N. We can simplify (11) by assuming normality of distribution in the X-variable. Normality is not a very violent assumption in the X-variable, although it would be for the Y-variable if there is marked curvilinearity.

$\sum x^4/N\sigma_x^4$ is β_2 in the Pearson terminology. So $\sum x^4/N = \beta_2\,\sigma_x^4$. But in a normal distribution, $\beta_2 = 3$. So the first term in the parentheses in (11) equals $3\sigma^4$. This combines with the second member to make $2\sigma^4$. Making this substitution and taking the square root, we have for the standard error of the coefficient c,

* Deming, W. E. Statistical adjustment of data. John Wiley and Sons, 1943.

$$S.E._c = \frac{\sigma_y}{\sigma_x^2 \sqrt{2(N-3)}}. \qquad (12)$$

But our index of curvilinearity is $\frac{\sigma_x^2}{\sigma_y} c$. On the principle that $\sigma_{ax} = a\sigma_x$, we would then have for the standard error of our index of curvilinearity:

$$S.E._{i.c.} = \frac{\sigma_x^2}{\sigma_y} \cdot \frac{\sigma_y}{\sigma_x^2 \sqrt{2(N-3)}} = \frac{1}{\sqrt{2(N-3)}}. \qquad (13)$$

In mathematical theory this has the same meaning and use as the standard error of any other statistic, and is to be divided into the c to get the standard error ratio. But in this sort of situation, as in many others, the worker should be warned against the customary acceptance of the hypothesis that there may be no true curvilinearity until the standard error ratio reaches the five per cent or the one per cent level. A test of statistical significance is properly employed to caution the worker against proceeding to action on the basis of his findings when there still remains considerable possibility that the difference tested may have arisen by chance fluctuation in sampling. In an experiment on methods of teaching, for example, one should withhold confident action even when the odds are 90 to 10 that the true difference is on the side on which he found it, for they are still 10 to 90 that the true difference might turn out to be on the other side. But in respect to taking warning that there may be true curvilinearity in our regression the danger of over-hasty action lies on the opposite side; here the unwarrantedly hasty action would consist in disregarding the evidence with low probability that there may be true curvilinearity and proceeding to treat the regression as rectilinear. So, whereas we hold forth for a standard error ratio of two or two and a half before we claim statistical significance in an experiment, we would better take warning from a standard error ratio of one in this connection (as, indeed, in a number of other situations of like character). The actual numerical size of the *i.c.* is more important that its standard error ratio. An *i.c.* of .06 or .10 or higher arithmetically shows an appreciable amount of curvilinearity.

In the table below are reported a number of parabolic *r's*. These are not a random sample nor are they selected as only the ones showing curvilinearity; they are ones computed from data in which it seemed plausible that we might find curvilinearity. In order to appreciate the phenomena involved in them, it may be well to approach the table by speculating about what might be expected to be the relation. A number of the parabolic *r*'s are ones computed between scores in subject fields made at a single long examination by

Parabolic Correlation Coefficients Between Indicated Pairs of Variates

Variates	*i.s.*	*i.c.*	$S.E._{i.c.}$	*Ratio*
History on mathematics	+ .08	—.168	.067	2.49
Percentage of votes for Willkie by counties in Pa. on density of population	—.291	+.208	.089	2.34
Percentage of votes for Willkie in 1940 by Counties in Ohio on density of population	—.402	+.011	.077	.14
Otis Classification test on age at eighth-grade level	—.163	—.167	.045	3.71
Literature on grade-point average	+.347	—.138	.067	2.04
Literature on mathematics	+.113	—.125	.067	1.85
Belief in superstitions on I.Q.	—.199	—.054	.021	2.57
Belief in superstitions on no. of magazines taken in home	—.145	—.037	.021	1.76
Belief in superstitions on no. of semesters taken in science	—.101	+.033	.021	1.50
Impulsiveness on success in nursing	—.265	—.125	.072	1.74
Indecision on success in nursing	—.140	—.054	.072	.75
Worry on success in nursing	—.027	—.041	.072	.57
Literature on science	+.408	—.111	.067	1.64
Grade-point average on math.	+.321	—.111	.067	1.64
Mathematics on literature	+.113	—.109	.067	1.61
English on grade-point average	+.528	—.081	.067	1.17
Washburne control scores on academic achievement ratio	—.034	+.093	.041	2.27
Washburne alienation scores on academic achievement	+.126	+.052	.041	1.27
Happiness scores on academic achievement	+.042	+.053	.041	1.29
Number of siblings on academic achievement ratio	—.091	—.006	.041	.15
Science on general intelligence	+.306	+.068	.067	1.01
English on mathematics	+.162	—.062	.067	.92
English on history	+.463	—.062	.067	.90
Mathematics on history	+.082	+.022	.067	.32
Moore-Nell academic achievement test on behavior inventory	+.033	—.010	.051	.19
English total on intelligence	+.728	+.022	.067	.33

113 sophomores at the Pennsylvania State College on the Carnegie Foundation Achievement Tests. As an individual increases his ability in mathematics, what is likely to happen in his measured abilities in history? The two abilities may be expected to increase together at the low end on account of the common elements of reading ability, general intelligence, etc. But in the later stages those persons who go in much for mathematics are likely not to be going in as much for history, and hence at the upper levels the curve relating the two could be expected to bend downward from the straight regression line that

would have fitted the early stages. Thus there would be a humped up curve for the regression line, represented by a negative index of curvilinearity. That is exactly what inspection of the first line in the table shows to be the fact. Nearly all the other parabolic r's bear out similarly the hypotheses one would set up for them as plausible. In the table the first term in each pair denotes the dependent variable and the second term the independent variable. Thus *history on mathematics* shows the manner in which the history scores regress as one goes up the scale of mathematics scores.

We have talked in terms of a second-order parabola. Parabolas of higher order could, of course, also be used and the descriptive statistic would need to have as many terms as the order of the parabola. But beyond the second order the computational work would greatly increase and the difficulty of communicating the results and of visualizing their meaning would likewise increase. It would also be possible to set up descriptive statistics in terms of other curves, such as the Gompertz or other exponential curves. But they also would be more awkward to compute and to communicate. Moreover, the second-order parabola will serve pretty well to detect and describe curvilinearity through the ranges through which we usually measure. We are likely to be measuring growth some time after its initial period, when it is in the upper bend of the S of the Gompertz curve; and in relations which are fitted by curves of the form $Y = e^{ax}$, we are also likely to be working within a range in which the second-order parabola will not fit too badly. But for actual prediction rather than rough description we should resort to the fitting of the needed type of curve. It would thus appear that the constants of a second-order parabola will make at least a satisfactory start in a more adequate description of regression than the sole resort to correlation coefficients upon which we have hitherto relied. Although our experience suggests that a disturbing amount of curvilinearity in regression is far from the rule, it does occur sufficiently frequently that the careful researcher will wish to be on the lookout for it. It is particularly likely to be missed when correlation coefficients are run on a calculating machine, or on the I.B.M. machines from ungrouped scores, where, consequently, there is no inspectional warning of the curvilinearity.