THE SAMPLING VARIABILITY OF LINEAR AND CURVILINEAR REGRESSIONS

A FIRST APPROXIMATION TO THE RELIABILITY OF THE RESULTS SECURED BY THE GRAPHIC "SUCCESSIVE APPROXIMATION" METHOD

By

MORDECAI EZEKIEL[1]

Many statistical problems involve determining the change in one variable with changes in each of several others, all operating at the same time. Linear multiple correlation provides a method of making this determination, on the assumption that all the relations are linear. In many problems this assumption is not valid. To determine curvilinear relations without making assumptions as to the type of each curve except that it be a continuous function, a method of successive approximations by graphic fitting was presented six years ago; and it was demonstrated empirically that in cases of high correlation this method successfully determined the underlying curves.[2] It was also pointed out that multiple regression curves could be fitted by the least-squares method. if specific parabolae or other first-degree equations were assumed for each variable, following methods previously sug-suggested by Yule.[3]

1. Formerly Senior Agricultural Economist, United States Department of Agriculture.
2 Ezekiel, Mordecai. A Method of Handling Curvilinear Correlation for any Number of Variables. Quart. Pub., Amer. Stat. Assoc., XIX, No. 148, Dec., 1924.
3. Yule, G. U. "On the Theory of Correlation," Jour. Roy. Sta. Soc., Vol. LX, p. 817 (1897). Apparently Wicksell had also suggested fitting regression curves to several variables simultaneously. Wicksell, S. D., Annals of Math. Stat., Vol. I, No. 1, pp. 3-15. Feb., 1930.

Institute of Mathematical Statistics is collaborating with JSTOR to digitize, preserve, and extend access to *The Annals of Mathematical Statistics.* www.jstor.org

The advantage claimed for the successive approximation method was that it did not require assumptions as to the specific type of each curve, but instead permitted each regression to be indicated by the observations themselves.

A new measure, the "index of multiple correlation," was suggested to measure correlation for curvilinear regressions in the same way that the coefficients of multiple correlation measured it for linear regressions.

No measure of the reliability of the net regression curves or of the index of correlation, was provided in the initial article. The usefulness of the results secured by this method has therefore been limited by the inability to state the confidence that could be placed in them even when based on a random sample, or to judge how large a sample would be necessary to infer, within any stated limits of precision and probability, the relations existing in the universe from which that sample was drawn.

This paper reports an attempt to determine the sampling error of multiple regression curves and indexes of correlation obtained by the successive approximation process, under conditions of simple sampling . The experimental method has been used to investigate the variability of results from successive samples drawn from the same universe under specified conditions and to establish error formulae inductively. These experiments, representing the solution of over 150 multiple curvilinear correlation problems, indicate the possibility of establishing approximate expressions for the reliability of multiple regression curves and indexes of multiple correlation.[1] The results, however, are not fully consistent, and the error formulae are not completely satisfactory. The experimental results are therefore given in full, in the hope that the attention of mathematicians may be attracted to this problem, and that the tentative formulae may be modified to provide more rigorous and exact measures of the reliability of the curvilinear regressions and correlations.

1. The extensive computations involved in this investigation were carried through by Helen L. Lee and Della E. Merrick, and by others of the staff of the Division of Farm Management, U. S. Department of Agriculture. Credit is due them for their intelligent and loyal assistance.

PART I.—COEFFICIENTS AND INDEXES OF CORRELATION.

1. THE REDUCTION OF THE "DEGREES OF FREEDOM" BY FREE-HAND SMOOTHING.

When a line is fitted to a series of paired observations by the use of the formulae $Y = a + bX$, the assumption is made that the straight regression line is adequate to describe the relation. Two parameters, one giving position to the line and the other slope, are required. For that reason, this equation will give a perfect fit to any two pairs of observations of X and Y. Furthermore, if the line is fitted to four pairs of observations, the determination of two parameters from four observations reduces the degree of freedom in obtaining the line from four to two; and the standard errors of the parameters must be determined with the number of degrees of freedom, N, equal to 2 instead of 4. Similarly, if a cubic parabola $Y = a + bX + cX^2 + dX^3$ were fitted to ten observations, there would be only 6 degrees of freedom after determining the four parameters, and the standard errors would be based on $N = 6$. In this case the four parameters determine position, slope, rate of change, and change in the rate of change.[1]

If instead of fitting a curve by the method of least squares or some other exact method, a free-hand curve is drawn by eye through the series of observations, it is necessary to make certain assumptions in drawing the curve, analogous to those represented in the parameters when more rigid methods are used. In addition to the basic assumption of continuity, these conditions may include:

(1) Whether the origin for $X = O$ will be at $Y = O$ or at some ordinate to be indicated by the data.

(2) Whether a straight line will be fitted (by ruler or thread) or whether a curve will be permitted.

1. The treatment of standard errors for small samples by "Student" and R. A. Fisher, as set forth in the latter's "Statistical Methods for Research Workers," give full recognition to these facts. Least square theory has always recognized that, for small samples, the number of parameters determined reduced the number of observations. See Wright, Thomas Wallace, and John Fillmore Hayford, "Adjustments of Observations," 1905, pp. 24-40, 132-133, and Merriman, Mansfield, "Method of Least Squares," 1911, pp. 80-82.

(3) If a curve, whether it will be limited (a) to a continuous arc of even curvature, (b) to a continuous parabola-like curve, (c) whether one or more inflections will be permitted, (d) whether the line will be so drawn as to minimize departures on the Y-axis, the X-axis, or at right-angles to the line itself.

It is evident that if a curve is drawn free-hand with its initial ordinate as indicated by the observations, with a continuous changing rate of curvature, and with no inflection, at least the three parameters of position, slope, and rate of change of curvature are represented, as shown by the corresponding equation for a parabola.

$$Y = a + bX + cX^2$$

It is true that the free-hand curve may involve still more parameters, but three is the minimum. While the number of parameters represented in any free-hand curve cannot be exactly determined, it can be roughly estimated by a process of reasoning similar to that indicated above; and any measure of the sampling reliability of such free-hand curves would be more reliable if it allowed for the number of parameters assumed than if it ignored this reduction of the degrees of freedom.

It should be noted that while the process of fitting curves free-hand involves the "taste" of the investigator, represented in the conditions he places on himself as previously mentioned, and on his skill in drawing the line under those conditions, the process of fitting a curve by a mathematical formula also involves "taste" in deciding what formula to use. If the conditions placed on the free-hand fitting are the same as those represented in the mathematical equation, the results may agree within the significant limits of error, and, therefore, either may be satisfactory for practical purposes.[1]

When coefficients of correlation or coefficients of multiple correlation are obtained from samples with a limited number of cases, the reduction in the number of degrees of freedom by the two or more parameters in the regression equation makes the observed correlation

1. Note the witty discussion of free-hand versus mathematical curves in the presidential address by E. B. Wilson, Proceedings: American Statistical Association, March, 1930.

tend to exceed the true correlation in the universe from which the sample was obtained. Accordingly, even the usual linear correlation coefficients, if obtained from small samples, tends to exceed the true values. Adjustments to correct for this factor will be considered before going to the more complicated problem of adjustments in observed indexes of correlation.

2. Bias in Coefficients of Correlation

Determining a coefficient of correlation from a finite sample reduces by 2 the number of degrees of freedom present. As a consequence, there is a tendency for the computed correlation to exceed the true correlation in the universe, and a corresponding tendency for the computed standard error of estimate to fall below the true value. Exact measures of the "most likely" value of the correlation coefficient were given by Soper and others in 1917[1] and an elaborate method was provided for estimating it.[2]

Where a coefficient of multiple correlation for n_2 independent variables is determined from a finite sample of n' independent observations, the degrees of freedom are reduced by the $n_2 + 1$ parameters represented in the regression equation. If $n' = n_2 + 1$, the number of observations exactly equals the number of parameters to be obtained, the least square solution reduces to a simultaneous solution of the n' observation equations, and the coefficient of multiple correlation comes out 1.00 regardless of the presence or absence of correlation in the universe.

R. A. Fisher called attention to this problem in 1924 and suggested an approximate adjustment of the observed correlation from limited samples by the equation

$$(1) \qquad 1 - \bar{R}^2 = \frac{n' - 1}{n' - n_2 - 1}\left(1 - R^2\right),$$

1. Soper, H. E., Young, A. W., Cave, B. M., Lee, A., and Pearson, K. On the Distribution of the Correlation Coefficient in Small Samples. A cooperative Study. Biometrika. Vol. XI, Part IV, May, 1917, pages 352-359.

2. Locus, Cit., pp. 374-375.

where n' and n_2 have the same meanings as above, R is the correlation observed in the sample, and $\bar{R}$ is the most probable correlation in the universe.[1] This correction is very similar to that deduced independently by B. B. Smith in 1925, directly from the least square adjustment for number of constants.[2] In the same notation as above, Smith's adjustment:

$$\bar{R}^2 = 1 - \frac{1-R^2}{1-\frac{n_2}{n_1}}$$

may be stated

$$(2) \qquad 1-\bar{R}^2 \doteq \frac{n'}{n'-n_2}(1-R^2)$$

which differs from Fisher's formula only by the omission of the -1 from both numerator and denominator. In restating this formula a year ago[3] the present author modified it to the form

$$\bar{R}^2 = 1 - \frac{1-R^2}{1-\frac{n_2+1}{n'}}$$

or, stated in the same form as (1) and (2)

$$(3) \qquad 1-\bar{R}^2 = \frac{n'}{n'-n_2-1}(1-R^2)$$

This differs from both the previous equations in including the -1 term in the denominator but not in the numerator. The effect is thus to make the correction most severe; i. e., the corrected value departs still more from the uncorrected value than in either of the other forms.

---

1. Fisher, R. A., The Influence of Rainfall on the Yield of Wheat at Rothamstead—Phil. Trans. B. ccxiii, 89-142; 1924.

2. Smith, B. B. Forecasting the Acreage of Cotton. Jour. Amer. Stat. Assoc., March, 1925. Footnote on p. 41.

3. Ezekiel, Mordecai. Application of the Theory of Error to Multiple and Curvilinear correlation. Jour. Amer. Stat. Assoc., Supp., pp. 99-104, Vol. XXIV, No. 165-A. March, 1929.

The interpretation of correlation coefficients adjusted by any one of the three equations (1), (2), or (3) has been difficult because of lack of a definite explanation of the meaning of the adjusted coefficients. To determine their exact meaning, and to decide which one of the three forms of adjustment is most satisfactory, a study has been made of the relation of the adjusted values to the distribution of simple correlation coefficients when computed from random samples of various sizes drawn from universes with specified correlations. The "Cooperative Study" gives tables showing the exact theoretical frequency curves for zero order correlation coefficients, computed from samples of from 3 to 25, and 50, 100, and 400 observations, for true correlations ranging from 0 to .9, by tenths. Ordinates of the distributions of observed correlations are given for each value from $r = -1.00$ to 1.00 by .05 steps. With the frequency curve thus defined by as many as 41 ordinates, a rough integral of the curve was constructed by a cumulative summary of the ordinates. Then dividing by the total area, the proportion below any particular value was determined. When ρ (the true correlation in the universe) = 0, the summation was made from 0 in both directions to show the proportion of all samples showing correlations falling below the particular r, either plus or minus. When ρ exceeds 0, the summation was made from -1.00 to increasing values, to show the proportion of all the samples which show correlations falling below any particular value.

For each size sample investigated as described, more than 50 % of the theoretical observed correlations exceeded the true correlation. Thus for $\rho = .40$, with samples of 4, over 55 per cent of the samples showed r in excess of 40; 53 per cent with samples of 9; and about 51 per cent with samples of 50. But with $\rho = .80$, over 61 per cent of the samples showed r above .80 with samples of 4, 56 per cent with samples of 9, and 53 per cent with samples of 25. If we define the value which will be exceeded by exactly half the samples as the value which is most likely to be observed in any given sample, this "most likely" observed correlation is evidently in excess of the true value. The problem is to determine the adjustment equation, similar to eq. (1), (2), or (3), which will reduce the observed value to the correlation which exists in the universe from which it is most probable that that sample was drawn.

Frequency ogives (on a percentage basis) were constructed from the tables in the "Cooperative Study for $\rho = 0$, 0.2, 0.4, 0.6, 0.8, and 0.9, for $n' = 4$, 5, 9, 17, 25, 50, and 100. Equation (1) was then tested against these ogives, to determine what was the significance of the adjustment. For zero order correlations, equation (1) becomes

$$1-\bar{r}^2 = \left(\frac{n'-1}{n'-2}\right)(1-r^2)$$

Hence, with $\rho=0$ and $n'=9$, (r) would have to equal at least ± 0.35 for $\bar{r}$ to be 0. Comparing this value, 0.35, with the frequency ogive for $\rho = 0$, $n'=9$, it was found that only 35 per cent of the samples would give observed correlations larger than 0.35, or smaller than -0.35. Similarly for $\rho = 0.6$ and $n'=17$, r would have to be 0.63 for $\bar{r}$ to be .60. For these conditions, 49 per cent of the samples would give observed correlation in excess of .63. Carrying out this same comparison for all of the ogives constructed gives results as shown in the following tabulation.

Size of sample (n')	When correlation in sample is					
	0.0	0.2	0.4	0.6	0.8	0.9
4	0.42	0.29	0.36	0.42	0.49	0.51
5	.39	.29	.37	.43	.48	.50
9	.35	.30	.38	.44	.48	.49
17	.33	.32	.40	.45	.48	.49
25	.32	.34	.42	.46	.48	.49
50	.31	.37	.44	.47	.48	.50
100	.31	.40	.46	.48	.49	.50

Proportion of samples, of specified sizes, drawn from universes of specified correlations, which show correlations in excess of the true value in the universe, even after adjusting the observed correlation by the formula $\bar{r}^2 = 1 - \frac{n'-1}{n'-2}(1-r^2)$

These values are determined from the graphs based on a rough integration by successive summations, and slight errors may have entered in making the graphic interpolations. Hence the values cannot be regarded as precise. The error probably does not exceed .01 or .02 in any case, however, so the results are sufficiently exact to interpret the general effect of the correction formula.

It is evident from the table that when the true correlation is high, .80 or above, the probability of a value as large as that implied by the use of adjustment formula (1) is practically .50. Tests by the tables given in the "Cooperative Study" for the most probable value show that the probability becomes almost exactly .50 for larger samples and still higher correlations, the adjusted values by those tables and by the correction formula agreeing to the third or fourth decimal place.

Where the true correlation is low, however, the table indicates that the adjustment is too severe—that is, the probability of the true correlation in the universe being as high as the correlation shown after the adjustment is more than .50, and may be as high as .70 (for $n' = 4$ or 5 and $\rho = 0.2$). Even with this variation in the meaning of the adjusted value, however, equation (1) gives a valuable adjustment, since it indicates the probable correlation with almost exactly a .50 probability where the correlation is high, whereas it indicates the probable correlation with a higher probability—between .50 and .70—for those cases where the correlation is low and the standard error of the coefficient is correspondingly large.

Comparison of equations (2) and (3) with the frequency ogives showed that where n' was small, the adjustment was more severe in the case of (3), and less severe in the case of (2), and did not in either case tend to approximate the 50 per cent probability, except where n' was very large. In some cases equation (2) gives corrected values so low that such cases are likely to occur more than 50 per cent of the time, and accordingly the probability would be even less than .50 that the correlation is really as high as shown by the adjusted coefficient.

It may be concluded that equation (1) gives the most satisfactory simple method for adjusting coefficients of simple or multiple correlation to remove the positive bias. *The adjusted value* thus obtained may be defined as ***the value that most probably exists in the true universe, in the case of a high correlation, or a value slightly below the probable true value, in the case of a low correlation.***

The adjustment of the standard error of estimate may next be considered. When a standard deviation, σ_s, is calculated from the items in a sample of n' cases, the probable standard deviation of the items in the universe, σ_x, may be computed (following Fisher) as

$$\sigma_x^2 = \frac{n'\sigma_s^2}{n'-1}$$

So if the standard error of estimate is calculated by the usual formula

$$S_e^2 = \sigma_x (1 - R^2)$$

but the adjusted correlation, $\bar{R}$, is substituted for R, and the value just shown is used for σ_x, the equation becomes

$$S_e^2 = \frac{n'\sigma_e^2}{n'-1}\left[\frac{n'-1}{n'-n_2-1}(1-R^2)\right]$$

(4)

$$S_e^2 = \frac{n'\sigma_s^2}{n'-n_2-1}(1-R^2)$$

This is identical with the equation given by Fisher[1], though in different form.

3. Correcting for Bias with Indexes of (Curvilinear) Correlation

Where correlation is measured with respect to curvilinear regressions, the greater number of parameters represented in the regression curve increases the tendency for the observed correlation to exceed the actual and requires a more drastic correction of the observed values. Where the regression curve is determined by a definite equation, the number of parameters is known, and the observed correlation may be adjusted to the most probable true correlation by the use of equation (1), as before. Since the number of parameters, rather than the number of independent variables, now becomes of moment, the equation may be restated for curvilinear correlation

$$1-\bar{\rho}^2 = \frac{n'-1}{n'-m}(1-\rho^2)$$

using m to designate the number of parameters, and ρ and $\bar{\rho}$ to designate the observed and the adjusted index of correlation. This formula may be used either for simple or for multiple curvilinear correlation. Thus if the regression equation

$$X_1 = a + b_2 X_2 + b_2'(X_2^2) + b_3 X_3 + b_3'(X_3^2)$$

1. Fisher, R. A., Statistical Methods for Research Workers. 1928. P. 117, first equations; page 135, 2nd equation.

had been fitted, m would equal 5. For a sample of 20 observations and an observed multiple correlation of 0.80, the most probable true correlation would be but 0.74.

Where the regression curve or curves have been fitted free-hand, the observed correlation may be even more in need of adjustment than where a definite equation has been employed.[1]

It is true that the number of parameters which it would take to duplicate the free-hand curve by a definite mathematical function cannot be exactly determined without finding some equation which will exactly represent the curve. On the other hand, even an approximate estimate of the number of parameters which would be required provides a better basis for judging the probable true correlation than does the observed correlation taken alone. Such an approximate estimate may be made by considering how many degrees of position, change, or movement are represented in the graphic curve. The following list suggests some of these:

(a) Position
(b) Direction
(c) Change of direction
(d) Change in the change of direction

Where several different free-hand regression curves have been obtained by the method of successive approximation, the number of parameters represented by each one must be estimated separately. Only a single "position" parameter is required, since the origin of each regression is purely arbitrary, depending upon the constant in the regression equation, and the origin assumed for each of the other curves. That is, in the curvilinear regression equation

$$X_1 = a + f(X_2) + f(X_3) + f(X_4)$$

the value of a depends upon the origin used in graphing each of the functions.

Once the number of parameters represented in the regression

1. Ezekiel, Mordecai. Application of the Theory of Error to Multiple and Curvilinear Correlations. Jour. Amer. Stat. Assoc., March, 1929, Supp., pp. 99-104. Vol. XXIV, No. 165-A.

equation has been estimated, equation (4) may be used to adjust the observed correlation. Until more exact information is available, the explanation of the precise meaning of the adjusted value which has just been developed for the coefficient of linear correlation, may be assumed (by analogy) to apply to the adjusted index of (curvilinear) correlation as well.[1]

4 Sampling Accuracy in Coefficients of Correlation

Although equations (1) and (4) may be used to find the *most probable* correlation in the universe from which a given sample has been drawn, they do not give any measure of the *range within which* the true value probably lies, for any specified degree of probability.

It has long been recognized that coefficients of correlation, computed from small samples drawn from a universe in which some correlation exists, show a very skew distribution. Even for samples of a size most used in actual research—up to $n = 100$ or larger—the distribution is so skewed that the computed standard error of the correlation coefficient is of relatively little value. Even with fairly large samples the chances of the observed value departing from the true value by four or five times its standard error are very much greater than any interpretation based upon the normal curve would indicate.[2]

Recent investigations by "Student" and by R. A. Fisher have developed means of determining the reliability of correlation coefficients

1. The adjusted correlation corresponding to a given observed correlation, for any size of sample and value of m, may be more readily determined from a graphic chart, instead of eq. (1) or (4). Such a chart is shown in the appendix to "Methods of Correlation Analysis," by the present author, page 404. (John Wiley and Sons, 1930.)

2. "Student," On the Probable Error of a Correlation Coefficient. Biometrika, Vol. VI., p. 302, 1908
Soper, H. E., On the Probable Error of the Correlation Coefficient to a Second Approximation. Biometrika, Vol. IX, p. 91, 1913.
Fisher, R. A., Distribution of the Correlation Coefficients of Samples, Biometrika, 10, p. 507, 1915.
Soper, H. E., A. W. Young, B. M. Cave, A. Lee, K. Pearson. Distribution of Correlation Coefficients in Small Samples. Appendix 11, to the papers of "Student" and R. A. Fisher. Biometrika, XI, p. 328-413.

while allowing for the skewness of their distribution. That phase of the subject will not be developed in this article; it is referred to here merely to call attention to the fact that even after the most probable value for the true correlation has been determined, it may still be necessary to take account of how much confidence can be placed in that value—of how far the correlation obtained from the sample, even after adjusting as suggested, is likely to vary from the true correlation of the universe for any stated odds of probability.[1]

It must be recognized that the interpretation of the reliability of a correlation merely serves to indicate the significance that may be attached to the observed correlation, in view of the possibility of variation of the observed value from the true value in the universe due solely to random variation in sampling. If the conditions under which the sample is obtained do not fulfill the assumptions of simple sampling, then obviously Fisher's methods cannot be used unless the necessary reservations or modifications are added.

1. Fisher, R. A. On the "Probable Error" of a Coefficient of Correlation Deduced from a Small Sample. Metron, 1, No. 4, p. 3, 1921.—Statistical Methods for Research Workers, pp. 159-175, 2nd edition, 1928.—The General Sampling Distribution of the Multiple Correlation Coefficient. Proc. Roy. Soc., A. Vol. 121, pp. 654-673. 1928.
The methods developed by Fisher in the last of these articles have been made more readily available by the construction of graphic charts, both for simple and multiple correlations, which are given in the present author's "Methods of Correlation Analysis," pp. 400-403.

PART II.—LINEAR AND CURVILINEAR REGRESSIONS

1. Sampling Variability of Linear Regressions

Relatively little attention has been given in practical research work to the reliability of the regressions determined. Many correlation studies. especially where multiple correlation has been employed, have been misinterpreted because proper attention has not been given to the standard errors of the regression coefficients. As was pointed out recently,[1] this sampling variation may readily be so great in practical work as to invalidate the conclusions as to the effect of various variables, even when samples of considerable size are employed.

Fortunately, regression coefficients, derived from finite samples selected by random sampling, tend to be distributed in a normal distribution in the same way as does the arithmetic mean, so that elaborate devices necessary to allow for skewed distribution are not necessary. If the necessary corrections are made for the failure of the distribution to be normal when the number of degrees of freedom falls below 30, the standard error of a linear coefficient of gross regression or of partial regression may be employed with only the same restrictions as apply in the case of the arithmetic mean. More recently the formula for regression errors has been extended by Working, Hotelling, and Schultz to develop the standard errors of each constant for curves fitted by least-square methods.[2]

Where the regression is represented only by a plotted curve instead of by a definite equation, the reliability of the curve has been unknown. Obviously, it cannot be estimated from the constants represented in the curve, for they are unknown, and only their number

1. Ezekiel, Mordecai. The Application of the Theory of Error to Multiple and Curvilinear Correlations. Jour, Amer. Stat. Assoc. Proceedings, 19th annual meeting, Vol. XXIV, No. 165-A, pp. 99-104, March, 1929.

2. Working, Holbrook, and Hotelling, Harold. Applications of the Theory of Error to the Interpretation of Trends. Jour. Amer. Stat. Assoc. Proc., Vol. XXIV, 165-A, pp. 73-85, March, 1929.
Schultz, Henry. Discussion of above paper. pp. 86-88.
Schultz, Henry. The Standard Error of a Forecast from a Curve. Jour. Amer. Stat. Assoc., June, 1930.

may be roughly estimated. Some knowledge of the variability of such regression curves may, however, be obtained experimentally.

2. Outline and Summary of Experimental Study of Sampling Variability of Multiple Curvilinear Correlation Results

The study was conducted by first constructing a set of data in which a dependent variable, X_1, was related to several independent variables according to known curvilinear regressions, and in which a certain known portion of the variance of X_1 was not related to any of the independent variables. A second universe was then constructed with the same underlying functions, but with a different proportion of random variation in the dependent variable. Successive samples of various sizes were drawn at random from both "universes" and net (partial) regression curves and indexes of multiple correlation were computed separately for each sample. The net regression curves obtained in successive samples of the same size were compared with the true curves and with each other to see how far the results determined from the samples differed from the true values, and how much variance there was among them. The variability of the curves, for samples of different size, different true correlations, and different points along the curves, was then studied, and it was found possible to construct an error formula to estimate the standard error of the regression curves from the values obtained in the individual samples. Checking this formula by applying it to each of the samples previously determined, the actual errors were found to be in fair agreement with the estimated errors.

For a more rigorous test of the new error formula for regression curves, two new synthetic universes were constructed. Samples of various sizes were drawn from them, net regression curves computed separately for each sample, and the actual departures of the computed curves from the true curves checked against the error indicated by the new formula. The agreement in this test was not so good as in the previous case, although 66.5 per cent of the ordinates of the curves showed errors no greater than their computed standard errors, only 20.3 per cent fell between 1 and 2 times the computed values, while 7.5 per cent fell between 2 and 3 times, as compared to 68.3, 27.2 and 4.3, the proportions to be expected if the distribution were normal.

On the other hand, 5.8 per cent of the ordinates had errors exceeding 3 times the computed standard error, and some departures in excess of 5 times the computed standard error were obtained. It is evident from these results that either (a) the tentative formula is not adequate to estimate the standard errors of regression curves determined by the free-hand method, or (b) that net regression curves obtained by the successive approximation process are so unstable that their errors cannot be represented by a normal curve, and possibly may be impossible of estimation by any mathematical process. In the hope that the attention of others may be drawn to this problem, and a more satisfactory error formula be obtained, the experimental study is given subsequently in as full detail as possible.

The indexes of multiple correlation obtained from successive samples of the same size, were studied with respect to (1) bias and (2) variability. As has been previously reported[1], the indexes of multiply correlation show an average positive bias even larger than that of coefficients of multiple correlation. Indexes of multiple correlation apparently require a correction which takes into account both the number of observations and the estimated number of constants represented in the regression curves, according to equation (4) already discussed. Further study of the variability of the correlations showed that as far as could be judged from the relatively small number of replications of each size sample (5 to 16) they tend to have a standard error of the order of

$$\sigma_\rho = \frac{(1 - \rho^2)}{n' - m} \tag{5}$$

where n' and m have the same meaning as for equation (4), and where ρ represents the observed index of multiple correlation. If this very rough approximation for their sampling errors is found adequate, it would seem logical to expect Fisher's determination for the sampling error of multiple correlation coefficients to apply equally well to indexes of multiple correlation.

In concluding this summary, it must be reiterated that these conclusions are only tentativhe. They provide at least some indication of the reliability of curvilinear correlation results, for which previously

1. *Loc. Cit.*, Proc. Amer. Stat. Assoc., March 1929, p. 100.

nothing had been known. The error formulae are only first approximations, however, and in the case of the error of net regression curves, are such a poor approximation that much more work remains to be done before the results of such analyses can be used with anything like the degree of confidence that can be felt in older and more well-established statistical procedures.

DETAILS OF EXPERIMENTAL STUDY

3. CONSTRUCTION OF SYNTHETIC UNIVERSES

The set of data used in the initial sampling was constructed as follows:

1. Values for X_2 were obtained by taking the sum of values from two dice. The throws were repeated 500 times, giving 500 values.

2. To insure some curvilinear correlation between X_2 and X_3, values of X_3' were computed for each value of X_2, according to the following function.

Value of X_2	Value of X_3'	Value of X_2	Value of X_3'
2	3	8	6
3	4	9	6
4	5	10	7
5	5	11	8
6	6	12	9
7	6		

One die was then thrown, and the value for X_3 computed as the die reading $+X_3'$ $\left[\Sigma - \text{die reading} + f(X_2)\right]$.

3. Values for X_4' were then computed for each value of X_3, according to the following function:

Value of X_3	Value of X'_4	Value of X_3	Value of X'_4
3	4	10	0
4	3	11	0
5	2	12	0
6	1	13	0
7	1	14	0
8	1	15	0
9	1		

Again, one die was thrown, and the reading of the die added to the X'_4 value to get X_4. This gave a set of 500 values of X_2, X_3, and X_4, fairly normally distributed, with positive correlation between X_2 and X_3 ($r = + .534$); with a negative correlation between X_4 and X_2 ($r = - .489$); and between X_3 and X_4 ($r = - .234$); and with all of the inter-correlations more or less curvilinear.

4. Values for a dependent variable, X_1, were then calculated according to the relation

$$X_1 = f(X_2) + f(X_3) + f(X_4) + e \quad ,$$

where the values for each of the functions were read from the assumed regression curves tabled below, and where e was obtained by throwing two dice, and taking the sum of the readings.

VALUES FOR ASSUMED REGRESSION CURVES

X_2	$f(X_2)$	X_3	$f(X_3)$	X_4	$f(X_4)$
2	2.6	4	2.0	1	0.0
3	3.4	5	1.5	2	0.2
4	4.0	6	1.3	3	0.7
5	4.4	7	1.0	4	1.7
6	4.7	8	1.0	5	3.0
7	5.0	9	1.3	6	4.1
8	5.0	10	1.7	7	5.0
9	5.0	11	2.1	8	5.0
10	5.0	12	2.8	9	4.5
11	5.0	13	3.6	10	3.3
12	5.0	14	4.4	11	2.5
		15	5.2		

Values for a second dependent variable, Y, were obtained by using the same assumed regressions, but obtaining the value for e by throwing a single die, rather than two dice. This gave two sets of 500 observations, both identical as to the independent variables, but with different dependent variables, and with the true correlation higher in one universe than in the other, since the dependent variable included a smaller proportion of random variation in one case than in the other. The complete set of 500 paired observations are shown in Table A.

4. Drawing Random Samples

Thirty-one separate samples were drawn from each of the 2 "universes"; 5 samples of 100 observations each; 10 samples of 50 observations; and 16 samples of 30 observations. In making the drawings, slips numbered from 1 to 500 were mixed in a box, and drawn at random. They were stirred afresh between each drawing. In making the drawings for the X_1 universe, the slips were not returned to the box until each sample was completed; so that the same set of data would appear only once in each sample. In making the drawings for the Y universe, each slip was returned to the box as soon as its number was noted. In a few cases this resulted in the same observations appearing twice in the same sample. While 500 is not an "infinite" universe as compared to a sample of 100, the difference in the method of drawing appeared to make no practical difference in the variability in the two sets of samples. However, the fact that the samples made an appreciable proportion of the "universe" would mean that the variability in the observed results would not be quite as large as if drawn from an infinite universe. Using Bowley's statement of this the maximum effect[1], however, which would be for the samples of 100, would make the of the observed deviations about one-tenth smaller than it would have been if determined by drawings from an infinite universe of similar characteristics.

For, following Bowley,

$$\sigma_{s'} = \sigma_s \sqrt{1 - n_s / n_u}$$

Where, $\sigma_{s'} = \sigma$ of actual sample, from a finite universe

$\sigma_s = \sigma$ of a similar sample, from an infinite universe

n_s = number of cases in sample

n_u = number of cases in the finite universe

Hence where $n_u = 500$, $n_s = 100$, then $\sigma_{s'} = .894\,\sigma_s$

Since the effect of the limited universe on the variation in the results can thus be estimated, the results can be transformed to what they would probably have been had a much larger universe been available for study.

5. Curvilinear Regressions Determined from the Samples

Net regression curves were determined for each sample by the method of successive graphic approximations, and indexes of multiple correlation were computed for each set of curves. Each sample was carried through successive approximations until no further significant increase in correlation was found by further modifications of the curves. From 2 to 4 approximations were necessary, in various cases. The multiple correlation found for each sample at the first (linear) solution,

1. Bulletin Int. Institute Statistics, Proceedings, Rome, 1925. Annex by A. L. Bowley, Cambridge Univ. Press.

and for each successive set of curves, are shown in Table B. For the Y universe, a multiple correlation was run to adjust, by least squares, the slope of each regression curve according to the formula.[1]

$$Y = a + b_2'\left[f(X_2)\right] + b_3'\left[f(X_3)\right] + b_4'\left[f(X_4)\right]$$

The indexes of multiple correlation (necessarily higher than the previous indexes) as found by this process are also shown in Table B. The further study of the sampling variability of the regression curves was based on the set of regression curves for each individual sample which showed the highest correlation for that sample.

6. Errors in Regression Curves from the Samples

The net regression curves determined from each successive sample were all put on a comparable basis by adding a constant to each so that the central ordinate of each would equal the central ordinate of the corresponding true regression curve. The differences between the adjusted ordinates at other points along the curves and the true ordinates would then show the errors in the curves. That is, the difference between ordinates at the central value and the ordinates at other points along the curve, as shown for the curves determined from the samples, were compared with the same differences for the true curves.

This procedure centered attention on the reliability of the slope and shape of the curves, rather than on the accuracy of their position. It is true that in linear correlation, the a as well as the b of the formula $Y = a + bx$, is subject to sampling errors, and formulae have been devised to compute its standard error. In the present case, however, it seemed desirable to first solve the problem of the shape and slope of the curve, before attacking the further problem of its position.

The departures of the curves found in the several samples from the true values for each curve are shown in Table C, for selected ordinates. The central point of reference (and therefore the point of 0 error) was taken at approximately the mean value of each independent variable.

The individual samples were studied to see if there was any relation between the correlation observed in individual samples and the

1. See pages 445-447, Dec. 1924, Jour. Amer. Stat. Assoc., for the original discussion of this process.

errors in the regression curves. No relation whatever was found between the size of the correlation in the individual sample and the size of the errors for the sample so long as samples of the same size and drawn from the same universe were compared.

Standard errors for the linear partial regression coefficients were computed for each sample by the standard formula given by Yule, and, modified, by R. A. Fisher:

$$\sigma^2_{b_{12.34}} = \frac{\sigma_1^2(1 - R^2_{1.234})}{n'\sigma_2^2(1 - R^2_{2.34})}$$

When the actual errors in the regression curves for individual samples were compared with these standard errors, again no relation was found for samples of the same size and drawn from the same universe. For that reason it was decided to abandon further study of the characteristics of individual samples, and instead study the characteristics of each entire set of samples of the same size and from the same universe.

7. Derivation of Tentative Error Formula

Study of the errors showed that, so far as could be judged from the limited number of observations, they had a marked tendency to a normal distribution. However, to prevent undue weighting of single extreme cases, the average deviation was used instead of the standard deviation as a basis for summarizing the results shown by different samples of the several sizes. These average deviations are shown in Table 1 (page 298).

Each of these results would be expected. The true standard error of estimate for Universe X is 2.39, and for Universe Y is 1.80, or 75.3 per cent as large. It would therefore be reasonable to expect that, other things being the same, the errors in the ordinates of the regressions for Universe Y would average only three-quarters as large as the corresponding errors for Universe X. Stating each mean error (Table 1) in Universe Y as a percentage of the corresponding mean error in Universe X, and taking the geometric mean of these percentages, it appears that on the average the errors in Universe Y are 78.5 per cent as large as in Universe X, or in fair agreement with the proportion expected. The extent to which average error shown in Table 1 for the selected ordinates in Universe X are correlated with the average error for the corresponding ordinate in Universe Y are shown graphically in Figure 1.[1] It is evident that the individual group averages agree fairly well with the expected relation. Accordingly, it was concluded that any formula for the standard error of net regression curves would have, for one component, $\bar{S}_{1.234}$, the standard error of estimate for the dependent variable, just as does the formula for the probable error of a linear net regression coefficient, which is

$$\sigma^2_{b_{12.34}} = \frac{\bar{S}^2_{1.234}}{n'\sigma^2_2 \left(1 - R^2_{2.34}\right)}$$

TABLE 1.

Average deviation of errors in net regression curves, at selected ordinates for various sizes of sample.

X_2	$f(X_2)$	Universe X			Universe Y		
		16 samples of 30	10 samples of 50	5 samples of 100	16 samples of 30	10 samples of 50	5 samples of 100
3	11.4	1.66	1.19	0.34	0.90	0.82	0.50
5	12.4	0.93	0.63	0.24	0.48	0.50	0.26
7	12.9	0.00	0.00	0.00	0.00	0.00	0.00
9	13.0	0.72	0.56	0.34	0.38	0.32	0.24
11	13.0	1.48	1.09	0.77	0.71	0.58	0.50
X_3	$f(X_3)$						
5	12.5	1.65	1.25	1.04	1.38	0.52	1.10
7	12.0	0.84	0.44	0.38	0.38	0.18	0.16
9	12.3	0.00	0.00	0.00	0.00	0.00	0.00
11	13.1	0.61	0.47	0.24	0.41	0.56	0.52
13	15.6	1.35	0.93	0.82	1.56	1.24	0.88
X_4	$f(X_4)$						
2	10.2	0.69	0.80	0.38	0.58	0.80	0.50
3	10.7	0.52	0.54	0.48	0.39	0.36	0.34
4	11.7	0.35	0.36	0.40	0.30	0.16	0.22
5	13.0	0.00	0.00	0.00	0.00	0.00	0.00
6	14.1	0.28	0.29	0.12	0.34	0.54	0.22
7	15.0	0.72	0.67	0.40	0.68	0.68	0.54
8	15.0	1.50	1.09	0.76	1.14	0.92	0.66
9	14.5	2.26	1.60	1.00	1.34	1.25	0.74

1. This and subsequent figures will be found at the conclusion of the paper.

It is evident from Table 1 and from Figure A, which shows the data graphically for Universe X, (a) that in general the larger the sample the smaller the average error; (b) that the further from the center ordinate, the larger the error; and (c) since the errors in Universe X were usually larger than in Universe Y, that the lower the true correlation, the larger the error.

The influence of sample size may next be considered. The number of observations is involved in two ways in the results shown in Table 1. In the first place, the average error tends to vary somewhat inversely with the size of sample. But in addition, it tends to vary with the distance from the central ordinate. Since the independent variables were composed of elements derived from dice readings, their distribution was roughly normal. As a result, the number of observations upon which the regression curves were based was largest toward the center portions, and thinned out toward the extremes. In the graphic approximation method of determining the curves, each portion of the curve is determined from the cases falling within that portion, rather than from all the cases as a whole. Accordingly, it seems logical to try to relate the observed differences in the average deviation of the errors to differences in the number of cases from which they were determined, rather than to the total size of sample.

There is no precise range within which the observations can be said to be considered in free-hand fitting. Instead of trying to measure the exact *number* of cases within any specified range, therefore, it seemed desirable to establish a measure of the *concentration* of observations at any point along the curve. Thus, for example, if within a given interval of X_1, with a group interval of u units, there are n_u observations, we can express the concentration of observations at the mid-point of that group by the relation

$$n_k = n_u \left(\frac{\sigma_x}{u_x} \right)$$

If the group-interval is taken equal to the standard-deviation of the variable, n_k will be simply the number of cases falling within that group. If, however, the group-interval is made either larger or smaller than the standard-deviation, this equation will measure the *concentration* of observations *in terms of the number per standard-deviation range.* In a rectangular distribution, changing the value of u_x

would change the size of n_o to a corresponding extent, so the value of n_k would be independent of the group-interval selected. In a normal distribution, however, n_k would be only an approximation of the true value which would be secured from the theoretical distribution when the total number of cases was made very large and v_x was made infinitely small.

On the basis of the foregoing reasoning, it was thought that the differences in the average deviations within each universe as shown in Table 1, might be explained by differences in the number of cases which each *portion* of each curve was based upon. In sampling theory the dispersion of values of a constant determined from successive samples ordinarily varies with $\frac{1}{\sqrt{n}}$, rather than with $\frac{1}{n}$, hence, in this case, it was tentatively assumed that the value $\frac{1}{\sqrt{n_k}}$ would be a component of the formula for the error of ordinates of regression curves. This hypothesis was tested by adjusting the average shown in Table 1 by multiplying each of these by the factor $\frac{1}{\sqrt{n_k}}$, determining the n_k in each case from the true distribution of that variable in the whole universe, and from the total number of cases in the samples. These average differences would presumably reflect the true distribution of each independent variable in the original universe, since the variations in distribution in different samples would tend to cancel out. We may therefore use the distribution of the entire universe to indicate the average distribution within samples of specified sizes drawn from that universe. The calculation of n_k for each ordinate in accordance with this method is shown in Table 2.

TABLE 2

Calculation of n_k values for selected ordinates and various sizes of samples.

Group	Number of cases (n_v)				Value of n_k [1]		
	In entire Universe	30	50	100	30	50	100
X_2							
3	32	1.92	3.2	6.4	2.171	2.803	3.964
5	55	3.30	5.5	11.0	2.846	3.674	5.197
9	53	3.18	5.3	10.6	2.794	3.607	5.102
11	38	2.28	3.8	7.6	2.366	3.055	4.370
X_3							
5	9	0.54	0.9	1.8	1.081	1.396	1.974
7	70	4.20	7.0	14.0	3.015	3.892	5.505
11	91	5.46	9.1	18.2	3.437	4.437	6.276
13	18	1.08	1.8	3.6	1.529	1.974	2.792
X_4							
2	62	3.66	6.2	12.2	2.771	3.577	5.060
3	76	4.56	7.6	15.2	3.093	3.993	5.648
4	79	4.74	7.9	15.8	3.153	4.071	5.757
6	91	5.46	9.1	18.2	3.385	4.370	6.181
7	55	3.30	5.5	11.0	2.631	3.397	4.804
8	26	1.56	2.6	5.2	1.809	2.335	3.303
9	16	0.96	1.6	3.2	1.419	1.832	2.591

1. Computed from formula $n_k = n_v \left(\frac{\sigma_x}{v_x}\right)$, with $v_x = 1$, $\sigma_2 = 2.455$; $\sigma_3 = 2.164$; $\sigma_4 = 2.098$, $v_x = 1$, since the frequencies for 3 include 2.5 to 3.5; for 5, 4.5 to 5.5, etc.

TABLE 3

Average deviation of errors in net regression curves, at selected ordinates, adjusted to error per unit observation per standard-deviation range

Group	Universe X			Universe Y		
X_2	[1] 30	50	100	30	50	100
3	3.60	3.34	1.35	1.95	2.30	1.98
5	2.65	2.31	1.25	1.37	1.84	1.35
7	0.00	0.00	0.00	0.00	0.00	0.00
9	2.01	2.02	1.73	1.06	1.15	1.22
11	3.50	3.33	3.33	1.68	1.77	2.16
X						
5	1.78	1.75	2.05	1.49	0.73	2.17
7	2.53	1.71	2.09	1.15	0.70	0.88
9	0.00	0.00	0.00	0.00	0.00	0.00
11	2.10	2.09	1.51	1.41	2.48	3.26
13	2.06	1.84	2.29	2.39	2.45	2.46
X						
2	1.91	2.86	1.92	1.61	2.86	2.53
3	1.61	2.16	2.71	1.21	1.44	1.92
4	1.10	1.47	2.30	0.95	0.65	1.27
5	0.95	1.27	0.74	1.15	2.36	1.36
6	0.00	0.00	0.00	0.00	0.00	0.00
7	1.89	2.28	1.92	1.79	2.31	2.59
8	2.71	2.55	2.51	2.06	2.15	2.18
9	3.21	2.93	2.59	1.90	2.29	1.92

When the values in Table 1 are multiplied by the corresponding n_k values, from Table 2, the adjusted values shown in Table 3 are obtained. Averaging together all the values in Table 3, average adjusted errors of 1.89 are secured for samples of 300 cases, 2.04 for samples of 50, and 1.98 for samples of 100 cases. It is evident that most of the difference due to different sizes of samples has been eliminated. However, even after this adjustment, the errors tend to increase as the ordinate departs from the assumed point of origin at the center. This same relation holds for linear regression lines. The standard error of any point on a regression line (in relation to the origin at $M_y = 0$) is $\sigma_b x$, and hence increases directly as x increases. A line

1. Number of observations in each of the successive samples.

continues out with the slope given it by b, and any error in b has a progressive influence on the accuracy of the line. The free-hand curve, on the contrary, is more flexible, and does not continue in any determinate direction. Hence it would hardly be supposed that the errors in the ordinates of the curve would increase with increasing values of X so rapidly as does the standard error of the straight line. The errors shown in Table 3 may be tested with respect to this hypothesis by averaging, for each universe, the errors shown by the three sizes of samples for the several selected ordinates and relating the resulting averages to the departures from the assumed means. To put these departures in comparable terms for the three variables, they may be stated in terms of standard deviation units. Carrying these operations through, the data appear as shown in Table 4.

TABLE 4

Average adjusted deviation of errors at selected ordinates, contrasted with departure from origin

Group	Departure from origin	Departure $\frac{}{\sigma}$	$\sqrt{\frac{d}{\sigma}}$	Average adjusted errors Universe X	Universe Y
X_2					
3	4	1.63	1.06	2.76	2.08
5	2	0.81	0.90	2.07	1.52
7	0				
9	2	0.81	0.90	1.92	1.14
11	4	1.63	1.06	3.39	1.87
X_3					
5	4	1.85	1.36	1.86	1.46
7	2	.92	0.96	2.11	91
9	0				
11	2	.92	0.96	1.90	2.38
13	4	1.85	1.36	2.06	2.43
X_4					
2	3	1.41	1.20	2.23	2.33
3	2	.95	0.97	2.16	1.52
4	1	.48	0.69	1.62	.96
5	0				
6	1	.48	0.69	.99	1.62
7	2	.95	0.97	2.03	2.23
8	3	1.43	1.20	2.59	2.13
9	4	1.91	1.38	2.91	2.04

It is evident from Table 4 that the average error, adjusted for size of sample, increased as the departure from the origin increased. This is shown more clearly in Figure 2, where the average error is plotted against the departures from the origin. This figure, however, indicates that the relation is not linear, as the errors do not increase in proportion. When the average errors are plotted against the departures on semi-log paper, however, as shown in Figure 3, the relation is substantially linear, and is of such an order as to suggest that the errors vary with the square-root of the departures, rather than the departures themselves. The line drawn in on each chart, with such a slope as to coincide with the square roots, parellels the relation fairly well, so from this it may be concluded that another constituent of the error formula will be

$$\sqrt{\frac{\text{Units departure from origin}}{\sigma_x}}$$

If the origin is made at the mean of X, the independent factor, X_2, X_3, etc., this segment of the error formula may be stated (using $x = X - M_x$)

$$\sqrt{\frac{x}{\sigma_x}}$$

Each of the adjusted errors shown in Table 4 may be further adjusted by dividing each one by $\sqrt{\frac{dep}{\sigma_x}}$, the value shown in the third column. They may also be adjusted to allow for the difference in the original standard errors of estimate in the two universes, as noted earlier. The standard error in Universe Y was 1.80 and Universe X, 2.39, so the errors may be made comparable by dividing those from each universe by the corresponding standard error of estimate. Performing these two operations, the average deviations of the errors appear as shown in Table 5. These average deviations are now so adjusted as to eliminate differences due to (1) number of observations in each portion of the distribution, (2) departure from origin, and (3) standard error of estimate in the universe. As stated in Table 5, the average deviations are in per cent of the deviations that would have been estimated from an equation representing the three elements discussed.

TABLE 5

Average deviation of errors at selected ordinates adjusted for n_k, $\sqrt{\frac{x}{\sigma_x}}$ and s_e

Group	Universe X	Universe Y	Average
Average X_2 - 3	1.09	1.09	
5	0.96	0.94	
9	0.89	0.70	
11	1.33	0.98	
Average X_2	1.07	0.93	1.00
X_3 - 5	0.57	0.60	
7	0.92	0.53	
11	0.83	1.38	
13	0.63	0.99	
Average X_3	0.74	0.88	0.81
X_4 - 2	0.78	1.08	
3	0.93	0.97	
4	0.98	0.77	
6	0.60	1.30	
7	0.83	1.28	
8	0.90	0.99	
9	0.88	0.82	
Average X_4	0.84	1.03	0.94

Averaging all values for each variable, as shown in Table 5, there still remains some difference in the average errors. The errors for $f(X_3)$ are smaller on the average than the errors for either of the other variables, while those for $f(X_2)$ are larger. This suggests that some element other than those already considered influences the errors, and that it differs with individual independent variables.

The formula for the standard error of a linear net regression coefficient contains the term

$$\frac{1}{\sqrt{1 - R^2_{2.34}}}$$

which allows for the intercorrelation between the independent variables. The more closely an independent variable may be estimated from the other independent variables, the less accurately its net regression line can be determined. The same relation might be expected to hold true of multiple regression curves. We can test this by comparing the average adjusted errors, just computed, withh the intercorrelation, as follows:

Regression	Mean Adjusted Error	$\sqrt{1-R^2}$	Mean Error / $1-R^2$
$f(X_2)$	1.00	$\sqrt{1-R^2_{2.34}} = 0.757$	0.76
$f(X_3)$	0.81	$\sqrt{1-R^2_{3.24}} = 0.844$	0.68
$f(X_4)$	0.94	$\sqrt{1-R^2_{4.23}} = 0.870$	0.82

It is evident that the means vary somewhat inversely with the $\sqrt{1-R^2}$ values. They may therefore each be multiplied by the corresponding $\sqrt{1-R^2}$ value to secure the final adjusted values, as shown in the last column[1]. This column now shows the average deviation of the errors actually observed stated in per cent of an estimated error computed from a theoretical equation composed of the four elements developed separately.

The average deviations of the observed errors varies from 68 to 82 per cent of the estimated error in each case, as contrasted to the value of 80 per cent to be expected if the equation gave the standard error. This is consistent with the fact that the standard error of estimate is included as the initial value in the equation. Furthermore, since the samples were drawn from a limited universe, the variation observed would tend to be slightly less than if they were drawn from an infinite universe with the same characteristics, which is consistent

1. This demonstration is by no means convincing proof of the need of including this adjustment. After this final adjustment, the discrepancy between the smallest and largest average errors, 0.68 and 0.82, is still as great as it was between the smallest and largest before, 0.81 and 1.00. On logical grounds, however, some such adjustment for the closeness of inter-relation between the independent variables is necessary, and by analogy, this method seems a possibility. It may be, however, that the index of (curvilinear) multiple correlation, $P_{2.34}$, should be used in the adjustment, rather than the coefficient of multiple correlation,

with the observed values falling mostly a little below the expected value of 0.80. The elements considered in estimating the error may therefore be said to give the standard error of the regression curves.

By a combination of induction and deduction, of which the foregoing is a condensed re-statement, a tentative formula for the standard error of the ordinates of a net regression curve was constructed from the four elements developed separately. They may be combined as follows[1]:

$$\text{I.} \qquad e_{f(X_2)} = \left(\bar{S}_{1.234}\right)\left(\frac{1}{\sqrt{n_k}}\right)\sqrt{\frac{x}{\sigma_x}}\left(\frac{1}{\sqrt{1-R^2_{2.34}}}\right)$$

or writing n_k out in full,

$$\text{II.} \qquad = \left(\bar{S}_{1.234}\right)\left(\frac{u}{\sigma_x n_u}\right)\sqrt{\frac{x}{\sigma_x}}\left(\frac{1}{\sqrt{1-R^2_{2.34}}}\right)$$

Hence

$$\text{III.} \qquad e^2_{f(X_2)} = \frac{\bar{S}^2_{1.234}\, u_x\, x}{n_u \sigma_x^2 \left(1-R^2_{2.34}\right)}$$

8. Testing Tentative Formula by Samples Drawn from the Original Universe

The formula which has just been shown was derived from the *average* errors shown by all the samples, using the known facts about each universe—the standard error of estimate, the frequency distributions and the standard deviations of each independent variable, and the inter-correlations among the independent factors in working out the estimated errors. But for practical use in estimating the reliability of regressions determined from a single sample all that would be known about the universe would be what could be inferred from that sample,

1. Equation (III) may be restated in a simpler form for practical computation, and the operations of working out the standard error for selected ordinates along the net regression curves may be organized in a systematic manner, as shown in the author's "Methods of Correlation Analysis," pages 384 to 389.

and the standard errors of the regression curves would have to be computed from the values so obtained. The next step of the experiment, therefore, was to calculate the standard error separately for each sample in turn, using only the values obtained from each one. These computations were made for each independent variable for each abscissa listed in Table C. The actual error of the regression curve at that point was then compared to the calculated standard error, and the ratio

$$\frac{\text{Observed error}}{\text{Calculated standard error}} = T$$

computed for each selected abscissa. If the computed error was the true standard error of the regression curve, these ratios should then be distributed according to the normal curve, and should have a standard deviation of 1.00.

The test was first applied to all the samples from both universes without including the term $1 - R^2_{2.34}$ in the error formula.

The standard deviation of the ratios σ_T was calculated separately for each selected abscissa of each independent variable with results as follows:

TABLE 6

Standard Deviation of Ratios of Actual Errors to Calculated Errors, as shown by 62 separate samples

Value of independent variable	Errors in $f(X_2)$	Errors in $f(X_3)$	Errors in $f(X_4)$
2			0.96
3	1.13		1.06
4			0.98
5	1.10	0.81	
6			1.19
7		0.82	1.21
8			1.23
9	0.89		1.11
11		1.13	
13	1.34	0.87	
All values	1.19	0.94	1.10

It is evident (1) that σ_T does not tend to increase appreciably as the abscissa departs from the mean of the independent variable; and (2) that the results based on the errors computed from individual samples are on the average quite consistent with those based on the facts from the universe. This is shown more fully in the following comparison:

Regression	Errors from individual samples		Errors from entire universe; mean adjusted error
	σ_T	$.80\ \sigma_T$	
$f(X_2)$	1.19	0.95	1.00
$f(X_3)$	0.94	0.75	0.81
$f(X_4)$	1.10	0.88	0.94

Taking 0.80 of the σ_T gives an approximate measure of the average deviations of the T values, to compare with the average deviation of the adjusted errors as calculated in Table 5. The average deviation of the T values ranges from 93 to 95 per cent of the average adjusted errors, showing the same average differences from variable to variable as were shown in Table 5 and suggesting the need of some element in the error formula to allow for the inter-correlation among the independent variables.

For the next step in the test, the term $1 - R^2_{2.34}$ was included in the error formula for $f(X_2)$ and the corresponding terms were included in the other formulas, using, in each case, the R values shown by each individual sample. Calculating the T values by comparing the actual errors with these revised estimates, and calculating their standard deviations, results were secured as follows:

Regression	σ_T, using full error formula
$f(X_2)$	0.77
$f(X_3)$	0.76
$f(X_4)$	0.90

The σ_T is calculated from 0 as origin, disregarding differences in the average error from zero. It is evident that in these sample results the errors, on the average, are somewhat less than would be expected from the formula, as σ_T falls below the unity. The distribution of the errors is also important. Figure 4 shows the distributions of the T values and compares it with the corresponding normal distribution. The extent of the agreement with the normal distribution may be judged from the following comparison:

Value of	Per cent of total frequencies in range			
	$f(X_2)$	$f(X_3)$	$f(X_4)$	Normal distribution
Over 3.00			0.5	0.14
2.00 to 2.99	0.9		2.6	2.14
1.00 to 1.99	6.5	4.4	11.9	13.59
0.00 to 0.99	46.1	40.1	31.3	34.13
0.00 to −0.99	38.3	46.2	43.2	34.13
−1.00 to −1.99	6.5	8.4	10.5	13.59
−2.00 to −2.99	1.7	0.9		2.14
−3.00 and larger				0.14

Although the distributions are not exactly normal, they agree fairly well. The different variables give slightly different distributions, however. For $f(X_3)$, in farticular, the distribution of the errors appears to be skewed, with more negative errors than positive ones. This may be due to a slight bias in the free-hand method of fitting the curve, which in this instance, for a very peculiarly-shaped regression curve, led to a slight but persistent error in the fitted curve. This possible individual bias in fitting the curve free-hand will be taken up again subsequently.

The test of the error formula described above was not a complete proof of the adequacy of the formula, since it used the same samples as those from which the original formula was constructed. For a more rigorous test the formula would have to be tried out on completely new samples secured from a different universe. Such a test was made in the next phase of the investigation.

9. Testing Tentative Formula by Samples Drawn from a New Universe

A new "universe" was constructed for testing purposes, by methods parallel to those described before. In this case only two independent variables were used. There were 328 observations in the universe and 45 samples were selected at random—15 of 10 observations, 15 of 20 and 15 of 40. (The number of observations was taken as small as 10 so as to make an extreme test of the value of the sampling formula.) Multiple curvilinear regressions were determined for $f(X_2)$

and $f(X_3)$, and the standard error of selected ordinates was computed by equation (III). The value of T was then computed by dividing the actual errors by the expected. The distribution of these errors is shown in Figure 5, as contrasted with the normal curve.

When the standard deviations of T are computed separately for each size of sample, the results are as follows:

	Size of Sample		
	10	20	40
$f(X_2)$	1.29	1.30	1.23
$f(X_3)$	1.46	1.80	1.79

Combining the distribution for both $f(X_2)$ and $f(X_3)$, the distributions of the errors for each size of sample are as follows:

Value of T	Size of sample 10	20	40	Normal Distribution
	Per cent of total	Per cent of total	Per cent of total	Per cent of total
Over 3.00	2.0	2.4	2.9	0.14
2.00 to 2.99	3.3	4.5	3.2	2.14
1.00 to 1.99	10.7	9.2	12.8	13.59
0.00 to 0.99	34.2	30.8	31.1	34.13
0.00 to -0.99	35.8	34.4	33.3	34.13
-1.00 to -1.99	8.3	11.5	7.8	13.59
-2.00 to -2.99	2.4	3.6	5.5	2.14
-3.00 and larger	3.3	3.6	3.4	0.14

There were many more wide departures—of 3.00 or larger—than would be expected if the errors had a normal distribution, with $\sigma =$ the estimated standard error. Instead of only 5 per cent of the errors exceeding twice the estimated standard errors, from 11 to 15 per cent were this large. Yet the general distribution of the errors (Figure 5) was in fair agreement with a normal distribution.

Two elements may contribute to the greater variation in the actual errors than in the estimated. With samples of the size involved—10 to

40 cases—the shape of various portions of the curve is determined by much less than 30 observations, and in some cases, by 10 or less. With such small samples, Student and Fisher have shown that for arithmetic means and other constants, the distribution of actual error $\div$ estimated error does not follow the normal curve and has a σ in excess of unity. It may be that some modification needs to be introduced into equation (III) to take account of this tendency before it can be correctly applied to small samples. From Student's table for small samples[1], 15 per cent of the errors would be expected to exceed twice the standard error if there were 3 degrees of freedom in the sample, and 10 per cent if there were 5. This indicates a reasonable number of cases, as compared with the size of the samples used in these tests. But whether $\frac{n_v \sigma}{u}$, or some other fraction of the total number of observations, would give the proper number of cases to use in entering the table, has not been determined, and more work needs to be done on this phase of the problem.

A second element of error appears to lie in using $\frac{1}{1-R^2_{2.34}}$ as one element of the error formula, instead of using the index of correlation, $\frac{1}{1-P^2_{2.34}}$. Substituting the index of correlation for the coefficient in the error formula was tried in two of the samples where the T values were the highest, and in both cases it much improved the accuracy of the estimated error—reducing values of T from 5.0 to 3.0, from 8.3 to 4.7, from 6.7 to 3.8, etc. It would appear that wherever the inter-correlation between the independent factors is markedly curvilinear, the accuracy of the estimate of the error could be much improved by measuring that curvilinear inter-correlation, and using it in computing the standard error of the function.

In view of the two sources of variation mentioned above, the fact that the variation of the actual errors ranges from 23 per cent to 79 per cent in excess of the variation of the estimated errors does not necessarily mean that the suggested formula (eq. III) is entirely inadequate, but may mean only that the necessary reservations in the use of the formula have not been applied. On the other hand, the fact that the actual results do vary as widely as this from the expected suggests that the formula can be used only as a very tentative approx-

1. This table is reproduced, in abridged form, in the author's "Methods of Correlation Analysis," on pages 19 and 392.

imation to the standard error of the regression curves until its possibilities and limitations have been more definitely determined.

10. Free-hand versus Mathematical Net Regression Curves

It was noted earlier that there appeared to be some tendency toward bias in fitting the first set of curves. The errors from the second universe, as shown in the last set of results showed a little of the same tendency, with the average error not falling exactly at 0. To test whether determination of the regression curves mathematically would eliminate this bias, mathematical partial regression curves were fitted by least squares to one set of samples from the second universe. The 15 samples of 20 observations were used, and two types of curves were fitted—the parabola and the cubic parabola. The regression equations were therefore:

$$(1)\quad X_1 = a + b_2 X_2 + b_2' X_2^2 + b_3 X_3 + b' X_3^2$$

$$(2)\quad X = a + b_2 X_2 + b_2' X_2^2 + b_2'' X_2^3 + b_3 X_3 + b' X_3^2 + b_3'' X_3^3$$

The estimated error was calculated for selected ordinates, using the same equation (III) as derived for free-hand methods, and T and σ_T computed. The values of σ_T were as follows:

	Simple parabola	Cubic parabola
$f(X_2)$	0.77	0.95
$f(X_3)$	0.90	1.13

It would appear, therefore, that equation (III) gives about as good results in estimating the reliability of net regression curves mathematically determined as it does in estimating the reliability of those secured by free-hand fitting.

Even with the curves fitted by least squares, however, there was some tendency to bias, as is illustrated in Figures 6 and 7. It is evident from these figures that neither the free-hand curve nor the mathematical curve exactly reproduced the true curve, even on the average of the fifteen samples. The average amount of bias is shcwn in the following statement:

AVERAGE BIAS IN FITTING REGRESSION CURVES

Value of independent variable	Average error[1] in $f(X_2)$			Average error[1] in $f(X_3)$		
	Parabola	Cubic parabola	Free-hand curve	parabola	Cubic parabola	Free-hand curve
2	0.16	0.09	0.62	-0.35	-0.31	-0.69
3	-0.06	-0.03	0.21	-0.10	-0.20	-0.31
4	-0.05	0.00	0.05	-0.06	-0.09	-0.15
5				-0.04	0.00	-0.02
6	0.11	0.02	-0.05			
7	0.11	-0.04	- .07	0.00	0.05	-0.05
8	0.13	-0.07	-0.10	-0.05	0.05	-0.07
9				-0.06	0.07	-0.18
10	0.23	0.43	-0.16	-0.09	0.09	-0.25
12	0.42	0.46	-0.24	-0.13	-0.09	-0.50
14				-0.27	-0.35	-0.79

In this particular, where the true curve is of such a slope as to be fairly well repiesented by a parabola or cubic parabola, the mathematical curves appear to give a slightly more accurate fit, on the aver age, than do the free-hand curves. The standard deviation of the errors, however, is only slightly greater for the free-hand curves than for those fitted by the cubic parabolae, as shown by the following tabulation:[2]

1. Taken with regard to sign.
2. At first glance it seems strange that the regressions fitted by the cubic parabola should have, on the average, larger errors than those fitted by the simple parabola. The explanation may be that the extra constant allowed the cubic parabola to follow more closely the individual characteristics of each sample; but that in fitting those (partly random) relations more closely, the regressions were distorted from the true underlying relation.

Standard deviation of errors (absolute values).

	Free-hand	Parabolic	Cubic
$f(X_2)$	0.98	0.70	0.84
$f(X_3)$	0.91	0.65	0.89

Where the true regression is of such shape that it could not be represented by any simple equation, it seems likely that the free-hand method would give a more accurate fit than would a mathematical equation which was not capable of representing the particular relation involved. Since, in practical investigations, the shape of the net regression curve is usually unknown to start with, the most satisfactory procedure would seem to be to use the free-hand method to determine the approximate shape of the curves, and then, if their shape appeared to follow any definite types by least-squares as a final check on the shape of the curves.

CONCLUSION

This article is only a progress report. The experiments reported here suggest that it may be possible to develop a formula for the standard error of net regression curves fitted free-hand. The problem has not been completely solved; the tentative formula which is developed has given only fair results in experimental tests; and several points are in need of further study. I hope at some future time to carry this investigation further, but my present plans make it necessary to lay it aside for a year or more. I am, therefore, publishing this preliminary report now, in the hope that others may be led to attack the same problem.

Mordecai Ezekiel

FIGURE A

AVERAGE ERRORS OF REGRESSION CURVES

Universe X

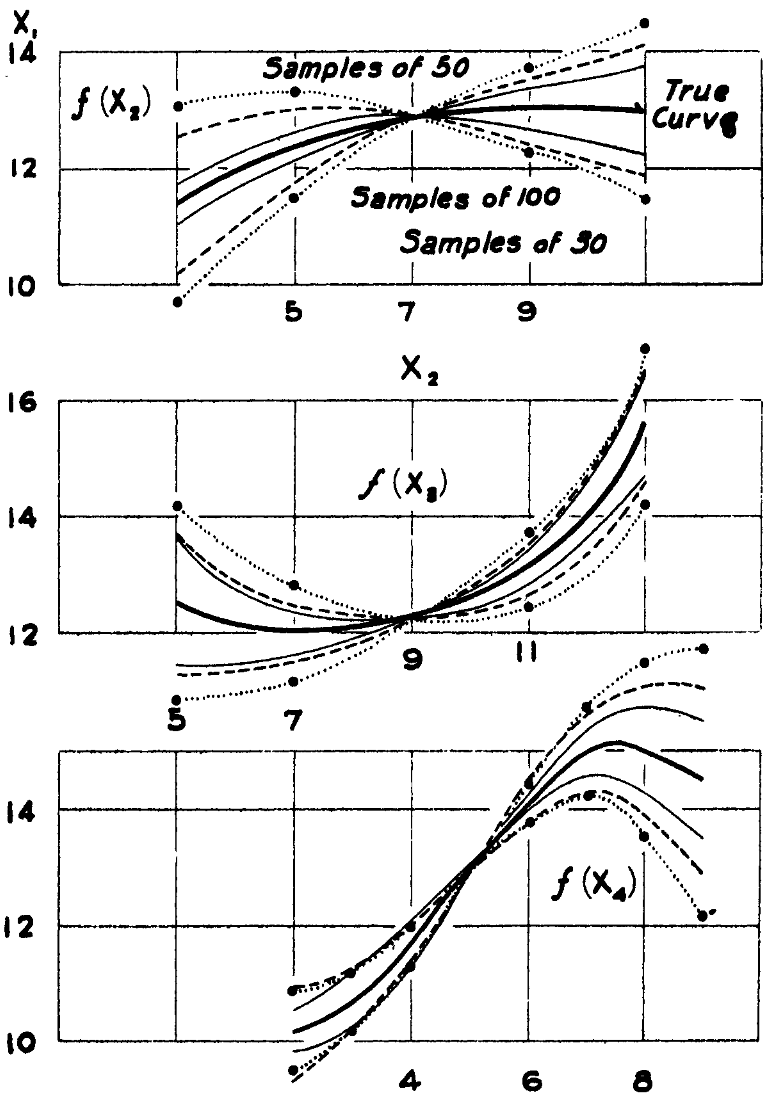

FIGURE 1

CORRELATION BETWEEN CORRESPONDING AVERAGE ERRORS IN UNIVERSES WITH DIFFERENT STANDARD ERRORS OF ESTIMATE

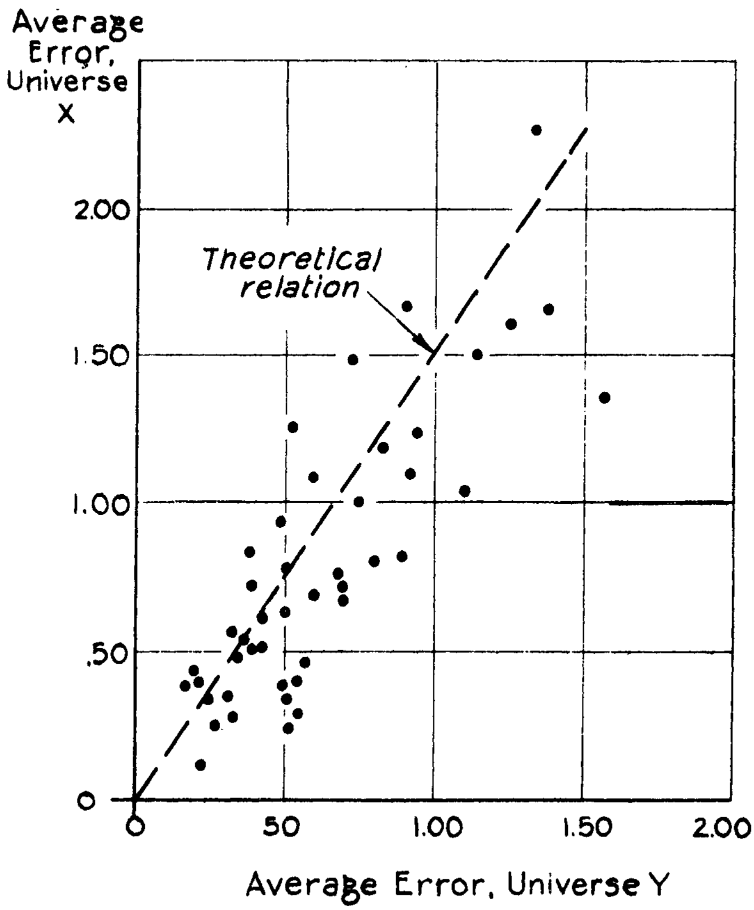

FIGURE 2

RELATION OF AVERAGE ERRORS, ADJUSTED FOR N_K, TO DEPARTURE FROM CENTER

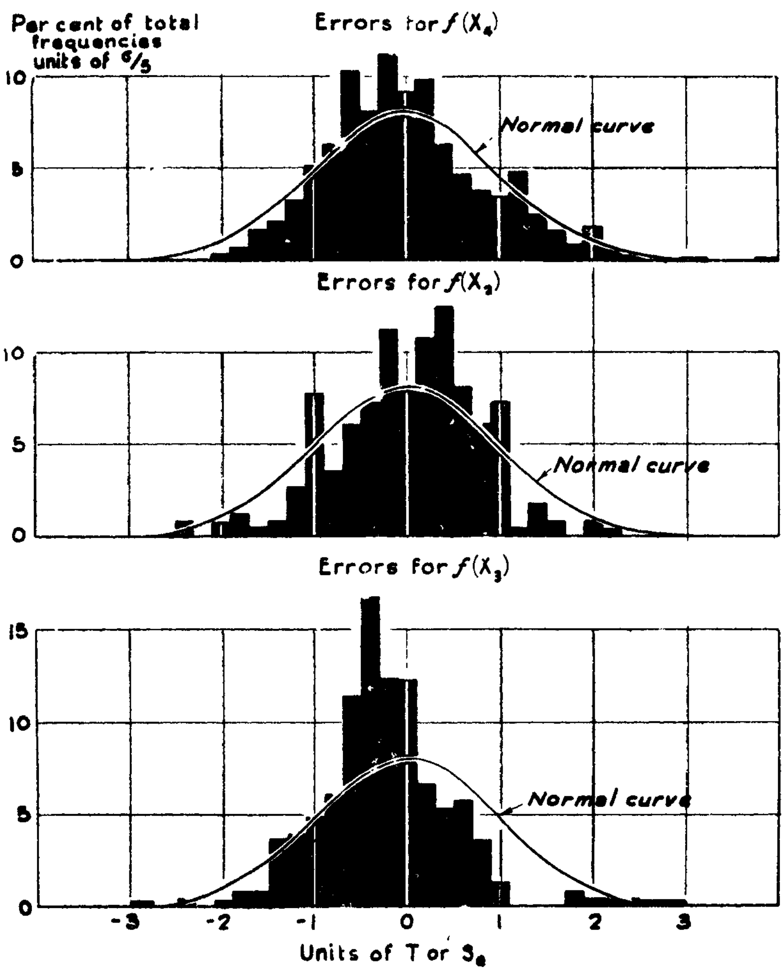

FIGURE 3

RELATION OF AVERAGE ERRORS ADJUSTED FOR N_K, TO DEPARTURE FROM CENTER

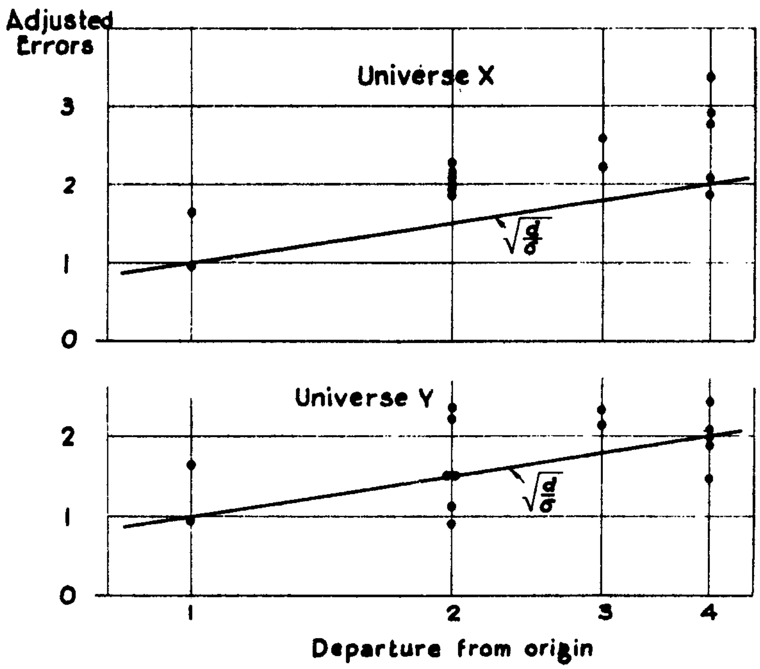

FIGURE 4

FREQUENCY DISTRIBUTIONS OF ERRORS

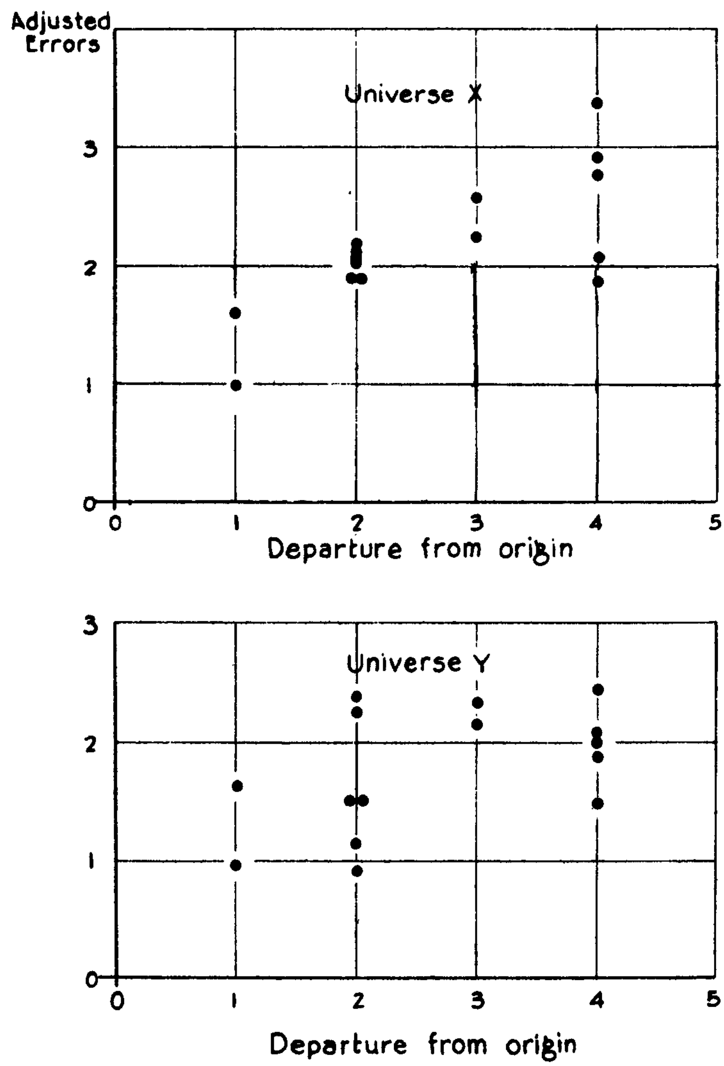

FIGURE 5

FREQUENCY DISTRIBUTIONS OF ERRORS

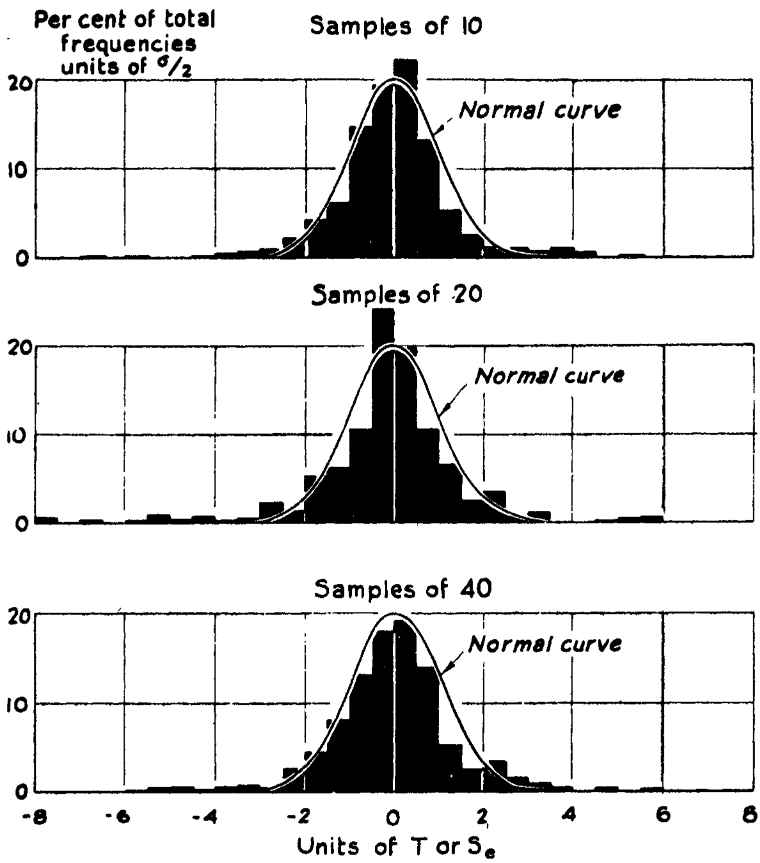

FIGURE 6

AVERAGE CURVES FITTED BY THREE METHODS $f(X_2)$

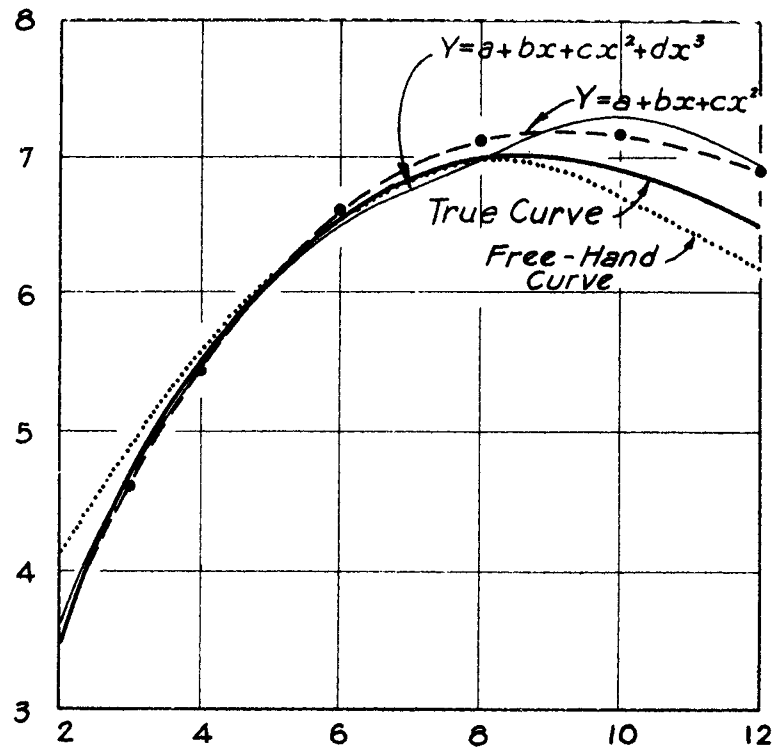

FIGURE 7

AVERAGE CURVES FITTED BY THREE METHODS $f(X_s)$

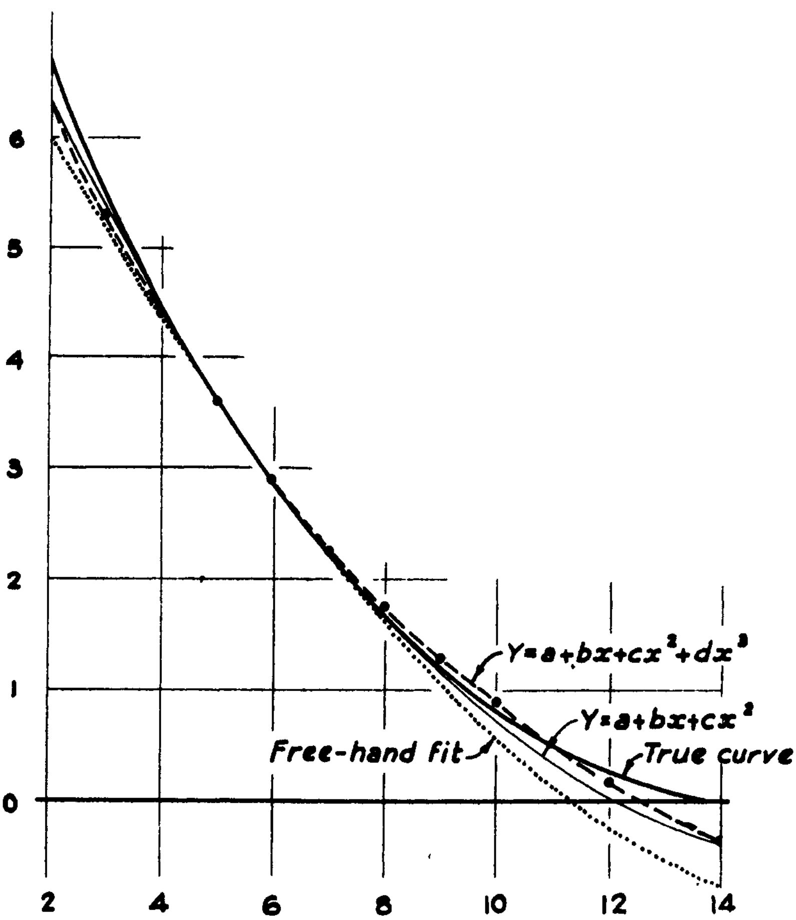

TABLE A—SYNTHETIC DATA FOR SAMPLING STUDY

No.	X_2	X_3	X_4	X_1	Y
1	5	6	3	9.4	10.4
2	6	9	2	14.2	9.2
3	8	9	4	16.0	11.0
4	7	12	5	16.8	12.8
5	10	13	2	14.8	14.8
6	6	10	6	18.5	12.5
7	8	11	4	13.8	14.8
8	2	4	10	9.9	9.9
9	10	13	2	10.8	11.8
10	11	11	3	18.8	9.8
11	4	10	8	17.7	15.7
12	4	9	7	18.3	12.3
13	7	9	3	11.0	8.0
14	7	8	4	16.7	10.7
15	10	11	2	14.3	10.3
16	11	9	5	15.3	12.3
17	11	9	6	13.4	16.4
18	10	11	2	15.3	13.3
19	10	13	3	15.3	15.3
20	6	8	3	14.4	9.4
21	8	12	4	17.5	14.5
22	6	7	3	13.4	10.4
23	10	11	5	21.1	12.1
24	11	9	4	18.0	11.0
25	7	7	3	15.7	8.7
26	7	9	2	11.5	8.5
27	6	10	2	12.6	7.6
28	9	10	3	12.4	9.4
29	6	9	5	13.0	10.0
30	12	10	2	15.9	11.9
31	6	8	3	11.4	7.4
32	11	12	4	16.5	15.5
33	4	8	9	11.5	13.5
34	9	10	6	13.8	11.8
35	7	10	7	19.7	13.7
36	11	13	1	12.6	13.6
37	6	11	7	20.8	17.8
38	6	8	3	8.4	7.4
39	5	11	4	16.2	9.2
40	7	9	6	16.4	14.4
41	5	8	4	15.1	9.1
42	6	7	2	16.9	6.9
43	7	12	6	18.9	16.9
44	10	10	2	14.9	10.9
45	4	8	4	16.7	10.7
46	3	5	5	13.9	8.9
47	11	11	3	15.8	9.8
48	7	11	7	22.1	18.1
49	10	13	4	17.3	12.3
50	5	7	4	12.1	8.1
51	8	7	4	12.7	13.7
52	9	11	3	14.8	10.8
53	7	12	3	19.5	10.5
54	9	11	2	15.3	10.3
55	8	9	4	12.0	9.0
56	9	7	3	13.7	11.7
57	10	8	1	12.0	11.0
58	11	11	5	20.1	12.1
59	4	11	8	18.1	12.1
60	7	11	7	21.1	18.1
61	9	8	2	15.2	12.2
62	8	11	2	14.3	8.3
63	3	7	7	12.4	10.4
64	8	8	7	17.0	15.0
65	3	8	9	17.9	11.9
66	9	12	2	12.0	11.0
67	4	11	9	17.6	16.6
68	6	7	2	14.9	10.9
69	7	12	6	16.9	16.9
70	2	8	10	15.9	8.9
71	5	10	4	18.8	8.8
72	8	11	5	17.1	12.1
73	9	10	2	10.9	8.9
74	11	13	5	14.6	17.6
75	5	8	8	17.4	11.4
76	5	7	7	21.4	15.4
77	7	8	2	15.2	11.2
78	6	9	2	14.2	9.2
79	6	12	4	16.2	10.2
80	3	8	6	14.5	9.5
81	7	7	3	16.7	7.7
82	8	8	3	12.7	10.7
83	6	8	4	15.4	9.4
84	8	12	7	22.8	15.8
85	12	14	3	16.1	16.1
86	7	11	4	15.8	10.8
87	5	7	6	16.5	15.5
88	11	13	3	17.3	11.3
89	7	11	2	12.3	10.3
90	7	11	4	12.8	9.8
91	7	7	6	12.1	13.1
92	7	9	6	15.4	16.4
93	7	8	5	14.0	10.0
94	6	10	7	20.4	13.4
95	11	13	5	22.6	14.6
96	8	7	6	14.1	16.1
97	6	7	6	14.8	11.8
98	2	4	6	11.7	13.7
99	7	8	4	11.7	11.7
100	10	13	6	20.7	16.7

TABLE A—SYNTHETIC DATA FOR SAMPLING STUDY (Continued)

No.	X_2	X_3	X_4	X_1	Y
101	6	12	4	17.2	15.2
102	3	5	8	17.9	15.9
103	7	10	2	13.9	11.9
104	8	7	3	14.7	7.7
105	6	7	6	19.8	11.8
106	4	10	5	12.7	11.7
107	11	14	6	20.5	16.5
108	3	9	9	16.2	11.2
109	9	9	6	18.4	15.4
110	3	5	8	17.9	12.9
111	12	14	6	20.5	14.5
112	9	8	6	12.1	16.1
113	7	7	3	10.7	10.7
114	7	8	2	14.2	7.2
115	6	9	4	13.7	10.7
116	11	9	5	19.3	13.3
117	5	8	6	17.5	10.5
118	7	11	4	14.8	14.8
119	6	9	4	11.7	13.7
120	7	8	2	15.2	11.2
121	7	7	2	14.2	7.2
122	3	7	10	13.7	9.7
123	5	6	5	15.7	9.7
124	8	8	6	18.1	16.1
125	10	8	3	14.7	9.7
126	6	8	7	19.7	13.7
127	7	12	3	13.5	12.5
128	6	10	4	16.1	9.1
129	4	6	4	16.0	10.0
130	6	12	2	9.7	9.7
131	11	9	3	13.0	13.0
132	5	11	7	15.5	12.5
133	3	8	5	15.4	8.4
134	10	12	6	21.9	16.9
135	12	10	4	18.4	11.4
136	8	11	6	19.2	15.2
137	7	11	6	19.2	13.2
138	3	8	9	17.9	12.9
139	10	9	5	17.3	13.3
140	6	8	3	15.4	8.4
141	8	9	2	10.5	12.5
142	5	9	4	16.4	10.4
143	8	11	6	23.2	13.2
144	8	11	5	17.1	12.1
145	9	12	6	23.9	14.9
146	6	11	3	17.5	13.5
147	5	9	8	17.7	13.7
148	11	14	4	18.1	17.1
149	8	11	6	21.2	13.2
150	10	12	1	16.8	9.8

No.	X_2	X_3	X_4	X_1	Y
151	7	9	5	17.3	11.3
152	9	8	5	12.0	11.0
153	7	12	3	13.5	14.5
154	7	10	6	14.8	11.8
155	8	8	7	17.0	14.0
156	3	9	6	12.8	13.8
157	12	12	1	14.8	12.8
158	12	12	3	20.5	9.5
159	7	7	3	16.7	10.7
160	2	6	9	17.4	14.4
161	8	12	3	13.5	11.5
162	9	11	7	21.1	17.1
163	6	12	4	18.2	10.2
164	7	9	3	13.0	9.0
165	6	10	5	21.4	15.4
166	9	10	5	15.7	11.7
167	6	12	7	15.5	18.5
168	9	11	2	15.3	11.3
169	7	11	6	21.2	12.2
170	5	8	8	17.4	12.4
171	4	11	9	20.6	15.6
172	8	7	6	17.1	15.1
173	6	9	3	12.7	7.7
174	12	14	6	22.5	14.5
175	7	7	2	12.2	7.2
176	7	7	2	14.2	8.2
177	4	7	6	16.1	15.1
178	7	11	4	17.8	13.8
179	7	10	6	16.8	12.8
180	10	11	5	16.1	16.1
181	11	12	5	14.8	14.8
182	9	12	6	20.9	12.9
183	10	9	1	13.3	10.3
184	7	11	4	16.8	9.8
185	9	12	3	18.5	9.5
186	8	12	4	17.5	14.5
187	11	11	4	12.8	14.8
188	5	9	4	12.4	12.4
189	6	7	4	11.4	12.4
190	2	7	11	12.1	10.1
191	5	7	6	17.5	10.5
192	5	11	5	12.5	11.5
193	7	8	7	18.0	15.0
194	9	11	5	13.1	15.1
195	11	13	6	23.7	13.7
196	9	10	4	13.4	13.4
197	7	12	4	15.5	15.5
198	6	8	4	12.4	11.4
199	4	6	4	9.0	12.0
200	3	10	10	18.4	9.4

TABLE A—SYNTHETIC DATA FOR SAMPLING STUDY (Continued)

No.	X_2	X_3	X_4	X_1	Y	No.	X_2	X_3	X_4	X_1	Y
201	4	10	7	17.7	14.7	251	11	14	5	18.4	13.4
202	2	6	11	9.4	12.4	252	5	6	8	20.7	13.7
203	6	11	6	17.9	11.9	253	4	9	5	15.3	11.3
204	7	9	4	16.0	14.0	254	3	5	7	15.9	15.9
205	3	10	6	16.2	14.2	255	9	11	4	14.8	9.8
206	8	8	5	17.0	12.0	256	11	11	1	15.1	8.1
207	5	10	8	16.1	12.1	257	9	7	6	13.1	11.1
208	4	11	8	17.1	13.1	258	7	9	3	14.0	13.0
209	8	11	5	16.1	16.1	259	6	10	7	18.4	16.4
210	7	8	6	13.1	11.1	260	6	8	4	15.4	8.4
211	11	11	1	14.1	13.1	261	5	8	4	14.1	9.1
212	8	12	6	17.9	15.9	262	9	8	3	11.7	8.7
213	7	10	6	19.8	15.8	263	7	11	4	18.8	10.8
214	6	11	2	17.0	12.0	264	2	4	6	15.7	14.7
215	8	12	2	17.0	14.0	265	9	11	3	17.8	13.8
216	10	10	1	13.7	7.7	266	7	8	4	15.7	11.7
217	10	13	3	19.3	15.3	267	11	11	5	14.1	15.1
218	6	9	4	17.7	10.7	268	5	11	8	20.5	17.5
219	8	10	6	17.8	12.8	269	6	12	2	15.7	9.7
220	6	12	5	14.5	15.5	270	6	8	2	13.9	7.9
221	8	9	5	18.3	12.3	271	8	11	4	13.8	10.8
222	8	7	6	14.1	13.1	272	8	7	6	13.1	15.1
223	7	7	7	17.0	13.0	273	5	7	3	9.1	7.1
224	10	8	2	14.2	11.2	274	5	10	5	20.1	10.1
225	7	9	7	18.3	12.3	275	11	12	1	14.8	11.8
226	10	8	1	9.0	7.0	276	7	9	7	22.3	17.3
227	6	7	5	19.7	12.7	277	5	7	8	18.4	11.4
228	5	11	4	16.2	12.2	278	6	7	6	13.8	10.8
229	10	10	3	13.4	13.4	279	5	6	6	15.8	15.8
230	7	12	6	17.9	13.9	280	8	12	3	14.5	13.5
231	6	7	4	16.4	8.4	281	5	10	3	14.8	7.8
232	2	6	8	16.9	14.9	282	10	11	4	17.8	14.8
233	3	10	6	13.2	11.2	283	7	12	6	21.9	12.9
234	9	7	5	16.0	10.0	284	5	8	3	18.1	9.1
235	10	8	6	18.1	12.1	285	12	15	5	21.2	15.2
236	3	5	5	12.9	9.9	286	4	11	4	10.8	9.8
237	8	7	7	20.0	13.0	287	3	10	5	19.1	14.1
238	9	11	6	16.2	13.2	288	8	10	6	20.8	16.8
239	7	10	3	11.4	12.4	289	11	13	5	20.6	16.6
240	3	10	7	15.1	13.1	290	10	11	2	15.3	8.3
241	4	11	5	18.1	15.1	291	4	10	9	20.2	13.2
242	7	11	2	13.3	8.3	292	7	7	5	19.0	15.0
243	9	8	3	11.7	11.7	293	8	7	5	16.0	14.0
244	4	10	5	12.7	14.7	294	3	7	8	17.4	10.4
245	5	7	3	15.1	12.1	295	6	10	6	17.5	15.5
246	7	8	5	18.0	13.0	296	4	11	4	10.8	10.8
247	4	11	6	13.2	11.2	297	5	6	3	17.4	12.4
248	3	10	9	16.6	14.6	298	8	7	7	22.0	17.0
249	8	7	2	14.2	9.2	299	12	10	2	13.9	11.9
250	5	10	8	18.1	17.1	300	4	8	9	13.5	11.5

TABLE A—SYNTHETIC DATA FOR SAMPLING STUDY (Continued)

No.	X_2	X_3	X_4	X_1	Y	No.	X_2	X_3	X_4	X_1	Y
301	6	11	6	20.9	16.9	351	4	9	7	12.3	12.3
302	6	11	2	13.0	9.0	352	7	7	4	15.7	9.7
303	8	7	6	19.1	16.1	353	7	12	3	16.5	9.5
304	12	15	6	23.3	19.3	354	9	11	3	17.8	12.8
305	7	9	3	14.0	11.0	355	6	8	5	14.7	11.7
306	3	5	5	10.9	11.9	356	9	10	5	19.7	12.7
307	8	11	7	18.1	17.1	357	9	8	4	19.7	10.7
308	9	8	7	19.0	17.0	358	7	9	2	11.5	10.5
309	5	10	4	13.8	12.8	259	5	6	3	13.4	9.4
310	7	9	2	12.5	9.5	360	5	8	6	15.5	12.5
311	7	7	5	12.0	11.0	361	12	15	2	21.4	15.4
312	3	8	8	18.4	15.4	362	6	12	2	17.7	13.7
313	4	6	6	16.4	13.4	363	8	7	7	20.0	17.0
314	7	9	4	19.0	9.0	364	6	12	3	11.2	13.2
315	3	5	7	17.9	15.9	365	6	9	6	12.1	14.1
316	6	8	7	16.7	11.7	366	8	8	6	21.1	13.1
317	2	8	7	12.6	10.6	367	10	8	3	9.7	8.7
318	9	11	3	14.8	12.8	368	7	10	7	18.7	14.7
319	11	10	3	18.4	11.4	369	5	9	3	10.4	11.4
320	7	9	7	22.3	12.3	370	12	14	4	19.1	14.1
321	7	10	7	20.7	14.7	371	7	8	6	15.1	15.1
322	5	7	4	14.1	9.1	372	4	7	8	16.0	11.0
323	7	8	2	14.2	8.2	373	3	6	7	18.7	14.7
324	9	11	7	22.1	17.1	374	7	7	3	12.7	9.7
325	5	6	4	17.4	8.4	375	9	8	6	17.1	12.1
326	8	10	7	19.7	16.7	376	6	9	4	11.7	13.7
327	12	11	1	12.1	10.1	377	7	7	3	15.7	9.7
328	5	8	6	16.5	14.5	378	8	8	4	11.7	8.7
329	8	8	5	13.0	13.0	379	5	10	5	13.1	15.1
330	6	7	2	14.9	7.9	380	5	11	3	19.2	12.2
331	8	7	5	13.0	10.0	381	11	13	2	20.8	9.8
332	3	8	6	13.5	11.5	382	9	11	5	18.1	11.1
333	8	9	6	18.4	12.4	383	8	9	7	18.3	12.3
334	9	7	3	12.7	7.7	384	3	7	8	16.4	12.4
335	7	8	7	15.0	17.0	385	4	7	9	11.5	12.5
336	7	9	3	14.0	8.0	386	5	8	7	15.4	13.4
337	7	10	3	12.4	12.4	387	5	7	3	15.1	8.1
338	11	14	2	17.6	14.6	388	9	11	6	16.2	13.2
339	4	11	9	18.6	16.6	389	10	8	6	14.1	11.1
340	7	7	3	17.7	7.7	390	8	8	4	12.7	9.7
341	4	6	5	12.3	14.3	391	8	7	6	15.1	16.1
342	2	4	8	16.6	14.6	392	8	9	3	12.0	11.0
343	3	5	6	16.0	12.0	393	7	10	7	21.7	17.7
344	7	8	6	17.1	13.1	394	10	8	2	14.2	8.2
345	11	11	6	18.2	14.2	395	9	12	2	14.0	9.0
346	8	12	5	16.8	11.8	396	4	6	4	16.0	9.0
347	9	10	2	16.9	8.9	397	4	8	5	11.0	13.0
348	7	8	4	15.7	12.7	398	10	9	3	13.0	8.0
349	11	10	4	17.4	14.4	399	7	11	2	13.3	13.3
350	7	7	3	13.7	9.7	400	9	10	5	16.7	11.7

TABLE A—SYNTHETIC DATA FOR SAMPLING STUDY (Continued)

No.	X_2	X_3	X_4	X_1	Y	No.	X_2	X_3	X_4	X_1	Y
401	7	12	6	14.9	13.9	451	6	12	7	20.5	18.5
402	2	6	8	18.9	13.9	452	7	8	5	17.0	13.0
403	5	10	4	15.8	13.8	453	5	7	8	14.4	14.4
404	8	12	2	19.0	9.0	454	9	9	6	20.4	15.4
405	7	8	7	21.0	12.0	455	3	7	10	12.7	10.7
406	6	9	6	16.1	11.1	456	9	7	7	16.0	16.0
407	6	11	4	14.5	13.5	457	12	11	1	9.1	9.1
408	6	12	6	21.6	14.6	458	8	8	3	14.7	9.7
409	9	12	7	22.8	18.8	459	11	13	1	16.6	11.6
410	8	7	3	15.7	11.7	460	6	9	2	8.2	9.2
411	4	11	8	17.1	12.1	461	6	8	5	15.7	14.7
412	10	9	3	19.0	9.0	462	3	5	9	15.4	13.4
413	7	12	6	17.9	13.9	463	6	12	2	18.7	9.7
414	11	10	1	13.7	11.7	464	7	10	2	13.9	12.9
415	8	11	7	19.1	18.1	465	11	11	2	11.3	13.3
416	4	11	9	13.6	11.6	466	9	11	5	16.1	13.1
417	6	8	5	14.7	12.7	467	8	12	7	17.8	17.8
418	3	10	5	13.1	9.1	468	9	10	7	20.7	14.7
419	8	7	5	16.0	11.0	469	7	8	5	18.0	12.0
420	5	6	5	13.7	12.7	470	9	10	3	14.4	10.4
421	4	11	8	18.1	12.1	471	10	13	3	15.3	13.3
422	7	10	3	11.4	13.4	472	7	10	6	14.8	15.8
423	5	6	6	13.8	13.8	473	11	11	2	10.3	13.3
424	3	6	5	17.7	12.7	474	7	12	5	12.8	15.8
425	11	11	1	13.1	10.1	475	5	8	7	20.4	16.4
426	7	11	6	22.2	12.2	476	12	13	6	17.7	15.7
427	4	11	8	18.1	13.1	477	4	10	5	16.7	9.7
428	7	9	4	16.0	14.0	478	9	10	4	19.4	14.4
429	6	9	3	13.7	11.7	479	9	7	6	18.1	12.1
430	8	9	2	9.5	10.5	480	9	10	7	16.7	15.7
431	11	14	3	15.1	12.1	481	6	8	6	18.8	15.8
432	6	12	5	18.5	11.5	482	5	8	4	9.1	12.1
433	5	6	6	16.8	14.8	483	11	10	4	16.4	12.4
434	9	9	7	21.3	13.3	484	8	7	4	16.7	9.7
435	9	11	3	12.8	13.8	485	6	8	6	16.8	10.8
436	7	10	4	16.4	10.4	486	12	15	5	16.2	15.2
437	5	6	7	13.7	12.7	487	4	7	9	16.5	14.5
438	5	8	8	20.4	16.4	488	5	11	6	18.6	15.6
439	7	11	2	14.3	13.3	489	4	6	5	15.3	12.3
440	3	8	10	14.7	12.7	490	7	11	5	16.1	13.1
441	12	14	3	16.1	15.1	491	12	13	1	16.6	13.6
442	8	10	2	14.9	12.9	492	7	10	2	16.9	9.9
443	7	11	6	23.2	17.2	493	11	14	4	15.1	14.1
444	8	11	3	13.8	11.8	494	6	12	4	18.2	10.2
445	4	6	9	19.8	14.8	495	7	11	4	16.8	13.8
446	9	10	7	17.7	14.7	496	7	12	4	13.5	12.5
447	5	9	7	17.7	14.7	497	10	10	4	19.4	14.4
448	5	7	6	18.5	14.5	498	9	10	2	16.9	8.9
449	7	9	2	14.5	9.5	499	5	9	8	17.7	11.7
450	11	13	5	21.6	15.6	500	9	9	4	13.0	11.0

TABLE B—COEFFICIENTS AND INDEXES OF MULTIPLE COR

(uncorrected for num

Sample No.	UNIVERSE X				
	R	ρ			
		1st curves	2nd curves	3rd curves	4th curves
		Samples of 30			
1	.620	.689	.714	.751	
2	.705	.737	.754	.756	
3	.510	.545	.703	.775	
4	.463	.487	.516	.508	548
5	.741	.614	.679	.736	
6	.486	.688	.720	.731	
7	.681	.777	.787	.801	
8	.532	.589	.659	.696	
9	.608	.649	.598	.659	
10	.469	.539	.597	.743	
11	.745	.792	.813	.818	
12	.614	.590	.628	.752	
13	.771	.741	.815	.790	
14	.586	.742	.794	.798	
15	.551	.569	.574	.578	
16	.634	.759	.809	.826	
		Samples of 50			
17	.529	.473	.545	.543	.510
18	.507	.536	.621	.622	.655
19	.418	.493	.541	.534	.536
20	.512	.686	.693	.702	.706
21	.526	.686	.721	.733	.745
22	.704	.721	.730	.727	.726
23	.666	.724	.747	.756	.753
24	.517	.650	.659	.679	.684
25	.703	.723	.721	.727	.729
26	.609	.646	.646	.677	.699
		Samples of 100			
27	.543	.629	.671	.684	
28	.679	.685	.699	.699	
29	.557	.649	.673	.673	.673
30	.565	.590	.656	.675	.678
31	.576	.650	.682	.682	.686

RELATION FOUND AT EACH SUCCESSIVE APPROXIMATION

ber of variables)

Sample No.	UNIVERSE Y					
	R	ρ				
		1st curves	2nd curves	3rd curves	4th curves	
Samples of 30						
47	.697	.698	.705			.715
48	.588	.781	.787			.794
49	.639	.800	.836			.858
50	.659	.782	.797			.801
51	.643	.812	.829			.851
52	.668	.745	.722			.746
53	.837	.877	.895			.898
54	.677	.736	.697			.737
55	.505	.679	.702			.720
56	.707	.767	.778			.782
57	.594	.639	.665			.676
58	.580	.660	.661			.669
59	.684	.762	.779			.785
60	.825	.880	.876			.881
61	.461	.621	.619			.642
62	.590	.756	.803			.819
Samples of 50						
37	.721	.786	.803	.793	.804	.804
38	.649	.686	.731	.713	.703	
39	.705	.730	.736	.738		
40	.679	.786	.797	.796		
41	.764	.773	.804			.805
42	.725	.764	.759	.723		
43	.721	.772	.798	.799	.800	.800
44	.688	.749	.781	.777		
45	.647	.676	.691	.695	.699	.699
46	.564	.672	.731	.733	.736	.736
Samples of 100						
32	.710	.764	.769	.769		
33	.482	.644	.656	.644	.650	
34	.668	.755	.760	.762		
35	.760	.794	.799	.802		
36	.555	.663	.673	.687	.684	

TABLE C—FOR UNIVERSE X: SAMPLING ERRORS IN NET

(The errors are observed or-

Ordinate	True Regression	SAMPLES OF 30															
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
X_2	$f(X_2)$																
3	11.4	.8		1.3	1.5	.3	3.0	1.1	1.5	1.8	2.1	2.4	2.1		3.0	.1	2.3
5	12.4	.6	1.4	.9	.4	.0	1.5	1.3	.2	.7	1.1	1.1	.7	3.4	.7	.1	.3
7	12.9																
9	13.0	1.0	.4	1.3	.4	.4	.4	1.3	1.5	.7	.9	.5	.1	1.0	.7	.2	.3
11	13.0	2.2	1.3	1.3	.8	1.1	.6	3.1	3.6	1.1	2.1	1.9	.4	1.0	.9	.4	1.3
X_3	$f(X_3)$																
5	12.5	2.1		1.8	.5	1.2	2.4		2.7	2.6	1.2	1.1			.2		2.3
7	12.0	.5	.2	1.3	.6	.3	1.3	1.0	1.5	1.7	.2	.7	.4	.6	.7	1.0	1.5
9	12.3																
11	13.1	.9	.2	.8	.4	.0	.0	2.1	.1	.5	.2	1.8	1.5	.0	1.1	.4	.3
13	15.6		2.2	1.8	1.6	1.4	.8	2.8		1.9	.3	2.0	1.8	.4	.4	.8	.7
X_4	$f(X_4)$																
2	10.2	.7	.4	.1	1.0	.6	1.1	.8	.5	1.4	.6	.5	.1	1.4	1.3	.1	.4
3	10.7	.0	.6	.3	.1	.7	1.2	.7	.7	1.1	.3	.5	.2	1.2	.2	.5	.1
4	11.7	.2	.3	.3	.2	.4	.7	.3	.5	.5	.0	.7	.1	.7	.1	.5	.1
5	13.0																
6	14.1	.1	.0	.1	.1	.2	.1	.3	.3	.7	.8	.4	.4	.4	.0	.4	.2
7	15.0	.8	.5	1.2	.0	.1	1.0	.5	.3	1.9	.7	.8	1.2	1.8	.1	.6	.5
8	15.0	2.8	1.9	1.8	.2	1.0	2.7	.7	.8	2.7	1.6	1.0	2.7		.9	.1	1.6
9	14.5	2.0		.6	1.3	2.9		.9	2.9	3.7	3.7	2.1	3.8		1.4	1.1	3.0

TABLE C (Continued)—FOR UNIVERSE Y

Ordinate	True Regression	SAMPLES OF 30															
		47	48	49	50	51	52	53	54	55	56	57	58	59	60	61	62
X_2	$f(X_2)$																
3	11.4	2.1	.5	.9	.2	2.6	1.0	.2	1.0	1.2	.7	.0	.5	.3	1.8	1.0	.4
5	12.4	1.0	.2	.3	.1	.8	.6	.3	.2	.4	.7	.3	.4	.1	1.1	.6	.5
7	12.9																
9	13.0	.7	.0	.0	.3	.4	.9	.3	.0	.3	.2	.1	.4	.3	1.5	.1	.6
11	13.0	1.3	.0	.1	.9	1.3	.9	.2	.0	.5		.4	.8	.1	3.1	.1	1.0
X_3	$f(X_3)$																
5	12.5				1.5		1.3				.6					1.6	1.9
7	12.0	.2	.3	.1	.0	.2	.3	.3	.5	.2	.4	.7	.6	.9	.1	.6	.6
9	12.3																
11	13.1	.6	.1	1.3	.9	1.2	.3	.4	.0	.0	.1	.3	.2	.2	.4	.3	.3
13	15.6	3.4	1.7	1.4		1.6	2.2	.4	1.7	2.0	.3	.9	2.0		2.5	.9	.9
X_4	$f(X_4)$																
2	10.2	.5	1.1	.5	.1	.0	.6	.3	.4	1.2	.0	1.2	.7	.6	.8	.4	.9
3	10.7	.4	.2	.8	.2	.1	.1	.0	.7	1.2	.4	.1	.0	.6	.6	.7	.1
4	11.7	.5	.1	.6	.1	.1	.2	.0	.5	.7	.3	.8	.1	.2	.5	.5	.1
5	13.0																
6	14.1	.2	.3	.1	.1	.5	.1	.5	.4	.0	.0	.4	.0	.9	1.3	.5	.1
7	15.0	.5	.8	.0	.0	.7	.1	1.2	.4	.6	.3	.9	.8	1.2	1.8	1.1	.5
8	15.0	.1	1.4	.6	.8	.4	.4	3.5	.2	1.2	1.3	.4	1.2	1.7	2.4	1.0	1.1
9	14.5		1.8	1.0	1.3	.1	.5		1.0	.9	3.7	.4	1.1	2.0	3.1	.5	1.4

REGRESSION CURVES, FOR SELECTED ORDINATES

dinates minus true ordinates)

SAMPLES OF 50										SAMPLES OF 100				
17	18	19	20	21	22	23	24	25	26	27	28	29	30	31
1.4	.5	.0	2.6	.8	1.3	1.5	2.7	.9	.2	.6	.5	.2	.4	.0
.1	.8	.0	1.3	1.1	.4	1.2	.9	.4	.1	.2	.3	.1	.1	.5
.9	1.1	.2	.3	.4	.6	.7	.8	.5	.1	.2	.1	.1	.4	.9
1.7	1.8	.4	.6	1.3	1.2	.1	1.5	1.2	1.1	.1	.8	.1	1.8	.9
.2	.4	.1			5.4	.6	1.0	.1	2.2	1.4	.2	1.1	2.3	.2
.5	.2	.3	.9	.3	1.0	.6	.3	.2	.1	.5	.2	.5	.5	.2
.2	.2	1.1	.4	.1	.4	.2	.4	.6	1.1	.7	.2	.2	.1	.0
1.8	1.2	.7	1.8	.0	1.0	.7	1.0	.7	.4	.1	.8	1.6	.4	1.2
1.2	1.3	1.9	.1	.8	.3	.5	.7	.8	.4	.2	.5	.4	.6	.2
1.2	1.1	1.6	.2	.1	.6	.2	.0	.3	.1	.4	.7	.3	.9	.1
.7	.5	.9	.2	.1	.4	.2	.1	.2	.3	.3	.5	.4	.6	.2
.1	.7	.3	.4	.0	.1	.1	.6	.4	.2	.0	.1	.2	.1	.2
.4	1.0	.4	.7	.3	.7	.7	1.3	.7	.5	.7	.2	.1	.7	.3
.9	.3	.3	1.7	.4	1.9	1.2	1.8	1.6	.8	.4	1.3	.3	.9	.9
.3	1.3	1.6		1.4	2.7	1.7	1.8	2.5	1.1	.1	2.7	.2	.2	1.8

SAMPLES OF 50										SAMPLES OF 100				
37	38	39	40	41	42	43	44	45	46	32	33	34	35	36
1.2	1.4	1.1	.6	.9	.1		2.2	.5	.7	.9	.5	.0	1.1	.0
.1	1.3	.7	.5	1.0	.2	.7	.6	.3	.4	.3	.1	.2	.7	.0
.6	.3	.3	.4	.4	.1	.1	.4	.0	.5	.3	.1	.3	.2	.3
1.6	.7	.5	.2	.1	.2	.3	.7	.2	.7	.6	.3	.8	.3	.5
	1.0	.5	.4		.7			.0		.4		.1	1.1	2.8
.2	.5	.1	.3	.1	.3	.1	.1	.1	.4	.2	.1	.1	.1	.3
.4	.1	.0	.2	.5	.6	1.6	.1	.0	.3	.8	.1	.9	.4	.4
1.0	2.1	1.2		.6	.5	1.2	1.5	1.1	2.3	.7	1.8	.1	.8	1.0
.5	.8	.2	1.3	1.1	1.3	1.0	.9	.1	1.3	.2	1.0	.3	.6	.4
.3	.7	.3	.6	.1	1.1	.2	.5	.2	1.0	.3	.7	.0	.1	.6
.0	.4	.3	.1	.4	.6	.0	.0	.1	.3	.3	.3	.1	.1	.3
.9	.1	.0	.2	.6	.3	.1	1.4	.2	.9	.3	.1	.3	.2	.2
.8	.0	.0	.0	1.1	.8	.3	1.0	.0	1.2	.9	.3	.2	1.2	.1
1.2	1.0	.6	.3	.8	1.9	.7	1.2	.4	1.5	.8	1.2	.6	.6	.1
2.1	2.5	1.1	.4	.2	3.2		1.7	1.0	1.7	.2	1.8	.9	.4	.4