YRB 51

GRAPHIC ANALYSIS OF MULTIPLE CORRELATIONS

by D. A. McCandliss

1951

There is nothing new in the idea of using graphic methods for analyzing various forms of mathematical relationships. Students of analytic geometry have been doing that for centuries. A method of graphic analysis of multiple correlations was described many years ago by Dr. Louis Bean, and has been called the "Bean method." This method has been criticised by some mathematical statisticians, but has proved its usefulness in many places as a convenient working tool in studying multiple correlations. It is this method that is now to be explained.

This method of graphic analysis can be used with either linear or curvilinear relationships. We shall start out with a very simple example of linear relationship, with data constructed in such a way as to be readily proved mathematically.

It should be kept in mind that this method is designed to analyze relationships which are additive, such as would be shown by an equation of the type:

$$S = ax + by \ldots + cm$$

Where S is the dependent variable correlated with a series of independent variables: x, y, m, etc. The coefficients, a, b, c, etc., might be either positive or negative in value, but the values of the various independents must be combined by addition or subtraction, rather than by multiplication or division, if the "Bean method" is used for the analysis. Another graphic method is available for analyzing relationships that are multiplicative.

To begin with, suppose we construct a very simple set of data from the equation:

$$S = ax + by$$

Suppose we let a and b each equal "1."

Take the following, which we know are true:

Observation :	S	= x	+ y
a	9	2	7
b	10	4	6
c	10	9	1
d	15	9	6
e	10	2	8
f	17	9	8
g	11	6	5

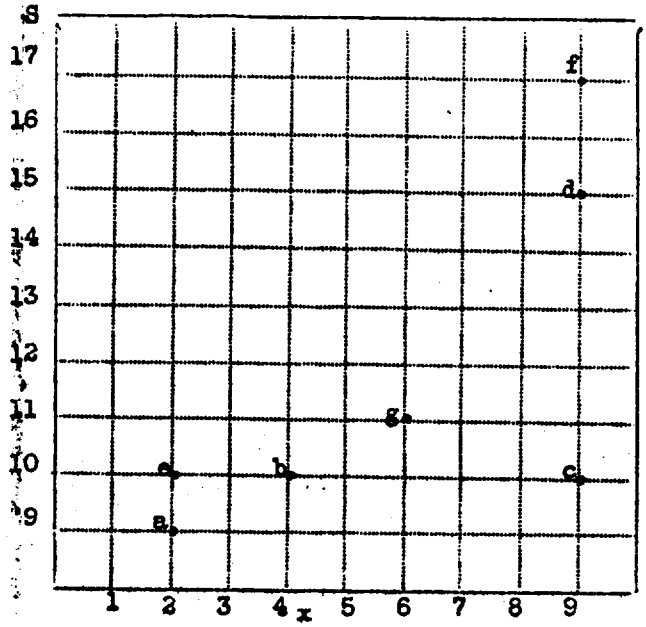
In each case it is apparent that "S" is the arithmetic sum of "x" and "y." Our problem is to compute "S" graphically when $x = 7$ and $y = 4$.

The data are to be plotted on ordinary "Cartesian" graph paper. The first step is to plot the dependent, "S", on the "Y", or vertical axis, against one of the independents on the "X", or horizontal axis. In the present case it makes no difference whether we use "x" or "y" on the first chart.

Chart 1, shows the result of plotting "S" vs "x" in this manner. Each observation is labelled so that its identity is retained.

Problem 1, in Graphic Correlation Analysis

Chart 1



Cover Sheet 1

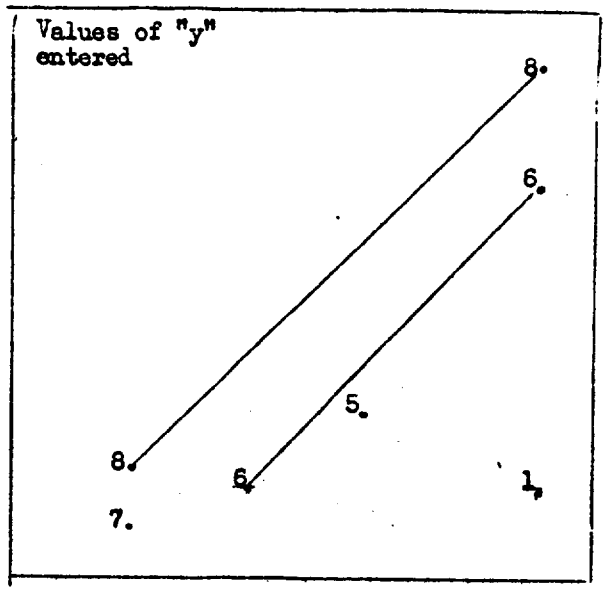


Chart 1 (Completed)

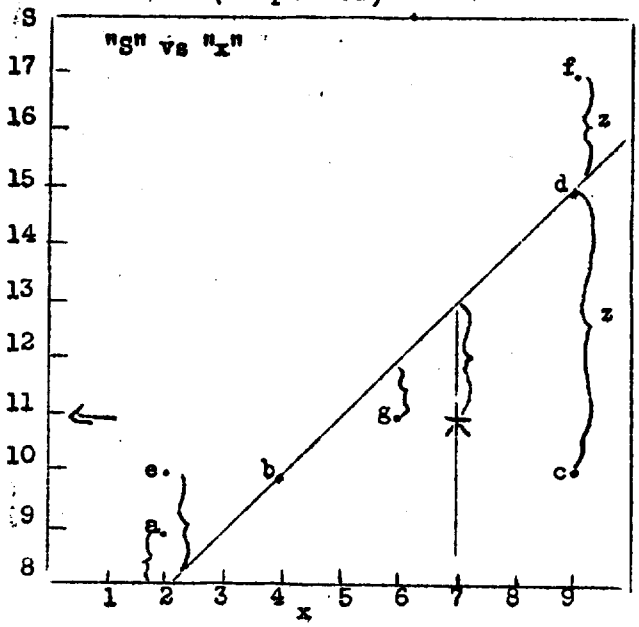
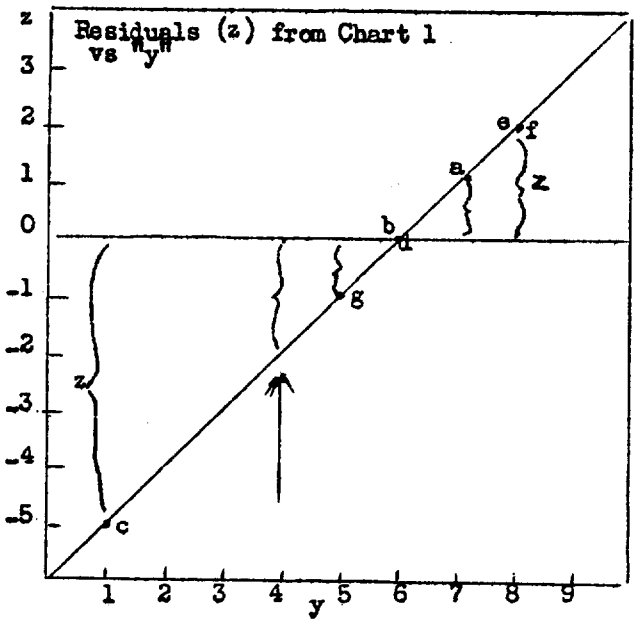


Chart 2



Data for Problem 1

Observation	S	x	y
a	9	2	7
b	10	4	6
c	10	9	1
d	15	9	6
e	10	2	8
f	17	9	8
g	11	6	5

The next step is to write in the value of "y" on the dot where each observation is located on Chart 1. A convenient way to do this is to write the "y" values on transparent paper, placed over the Chart. See "Cover sheet 1."

The next step is to locate a regression line which will take into consideration both the pattern determined by "x" and the values recorded for "y." The most important consideration in locating this line is to give it the correct angle, or "slope." It is not essential that it go through the data as a "line of fit", although when we come to curvilinear relationships we will find it helpful to do this, so far as possible. We might say that the objective is to locate a line that will divide the observations on the chart in such a way that the highest values of "y" will be on one side, the lowest values on the other side, and the intermediate values (so far as possible) are proportionately distant from the line. All these "distances" are to be measured in a vertical direction.

The first step in locating the slope of this line is to draw preliminary lines connecting identical values of "y." In the present case we have two "y" observations of "8." We connect these with a straight line. We have two more of "6" which we also connect with a line. It is apparent that these two lines have the same slope. It also is apparent that a line could be drawn through the chart, with this slope, that would put the highest values of "y" farthest away on one side, the lowest values farthest away on the other side, and the other values at proportional distances from the line. A line on the slope thru the "6" observations meets this requirement.

Accordingly we draw the line on the indicated slope. This line is first drawn on the thin cover sheet, and then transferred to the original Chart 1. Chart 1 then looks like "Chart 1, (completed)." (Except for the explanatory brackets.)

On Chart 1 we now have a regression line that explains that part of "S" which is due to "x." The unexplained part of "S" (which we know to be due to "y") is represented by distances, or residuals, marked "z", from the dots to the line. As said before, all residuals are measured in a vertical direction.

In drawing Chart 2, to measure the influence of "y", the first step is to draw a horizontal base line. With this base as "0" the vertical scale on the chart is measured "+" above the line, and "—" below the line, using the same units of measurement as were used for "S" on Chart 1. The values of "y" are measured on the horizontal axis, and these are plotted against the respective residuals from the regression line on Chart 1, as shown on Chart 2.

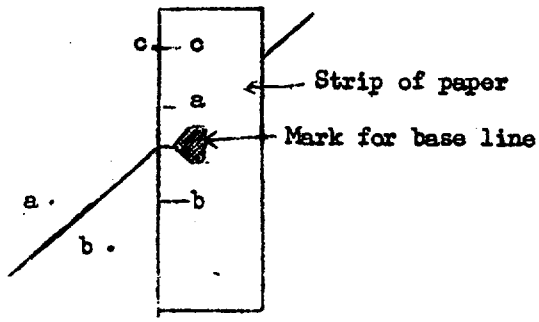
In the present case it is apparent that the dots fall on a straight line. This line is drawn, and no further residuals remain to be explained.

Our original problem was to determine "S" when "x" = 7 and "y" = 4, neither of which values occurs in our original data. On Chart 2, we find that the regression line crosses the "y" value of 4 at a level of "-2", on the scale of Z's. On Chart 1, the line crosses the "x" value of 7 at a reading of "13." Measuring two units down from the line on Chart 1 we read 11 on the "S" scale. This is the answer. Plain arithmetic tells us it is correct, because we know that 7 plus 4 does equal 11.

Suggestions for Charting

At this point it may be well to mention an idea that has proved helpful in transferring residuals from one chart to another. This can be done with dividers, but a quicker and easier way is to use a small strip of paper, or a thin card.

On the edge of this strip of paper a distinctive mark of some kind is made, to be placed at the regression line where the dot is to be measured. The measurement from the line to the dot is then marked on the edge of the strip. The strip is moved along the line and all residuals are measured in succession, before any are entered on the next chart. This can be done quite rapidly, and the method prevents accidentally plotting a residual on the wrong side of the base line, when it is plotted on the next chart. The measurements on the strip are also in convenient form to use in case the sum of the squares of the residuals is to be computed. The following sketch illustrates the procedure:



Another simple, but important, matter is that of labelling the dots on the charts. Since the curves are to be determined by observation, it is important to avoid optical illusions. When figures or letters are placed alongside dots, the eye naturally locates the position of the observation near the middle of the space occupied by both the dot and the label. The following four dots are in a straight line, but they appear to be on a curve, because of the way they are labelled:

1.1 1.2 1.3 1.4

On the other hand, the following dots are actually on a curve, but appear to be on a straight line:

1.1 1.2 1.3 1.4

This confusion can be avoided by placing the label for each dot in such a way that the dot is near the center of the area, like this:

1.1 1.2 1.3 1.4 or 1.1 1.2 1.3 1.4

When two or three observations are identical, or nearly so, the labels need to be arranged with this idea in mind, so far as possible. That can be done like this:

1.1 / 1.2 or 1.2, 1.3 / 1.4 or 1.2 / 1.2 or 1.1, 1.2 / 1.3

This piling up of observations is one reason why the present graphic method is not suitable for analyzing very large samples.

Problem 2 (Plate 2)

In our next problem we shall go one step further, and analyze a very simple linear correlation with three independent variables. Again we shall use constructed data, so that we can readily check the results. The data follow the equation: $S = ax + by + cm$, where a, b, & c, each equal 1. The data are as follows:

Observation:	S	x	y	m
10	11	2	7	2
11	16	4	6	6
12	11	9	1	1
13	16	9	6	1
14	18	2	8	8
15	18	9	8	1
16	14	6	5	3
To find:	?	7	3	5

We start out as we did with Problem 1, plotting "S" vs "x." (See Chart 3, Plate 2.) It is apparent from the scatter of the dots that no suitable "line of fit" could be plotted through the data on this chart, without considering other factors. We need to prepare a cover sheet for Chart 3, as in Problem 1, but we have two more variables to be considered, instead of only one. We have no way to be sure whether the sign between them is plus or minus, but we assume it must be one or the other if we are to use this method of analysis. Accordingly we prepare two cover sheets for Chart 3. On one we plot the sums of y+m, and on the other the differences (y - m).

Cover Sheet "3" shows the sums, and the "3-a" shows the differences between "y" and "m." It is apparent that no suitable line can be constructed on Cover Sheet "3-a." On this sheet joining the identical observations --- the fives and the zeros --- gives two lines going in opposite directions. On Sheet 3, however, we find that a suitable line is easy to draw. This line, joining Observations 10 and 15, is drawn on Chart 3. (See "Chart 3, completed.")

The residuals from the line on Chart 3 are plotted against "y" on Chart 4. Another cover sheet (not shown) is made for Chart 4, with "m" entered on it, and Chart 5 is drawn in the same manner as Chart 2 in Problem 1. It is obvious that no residuals remain.

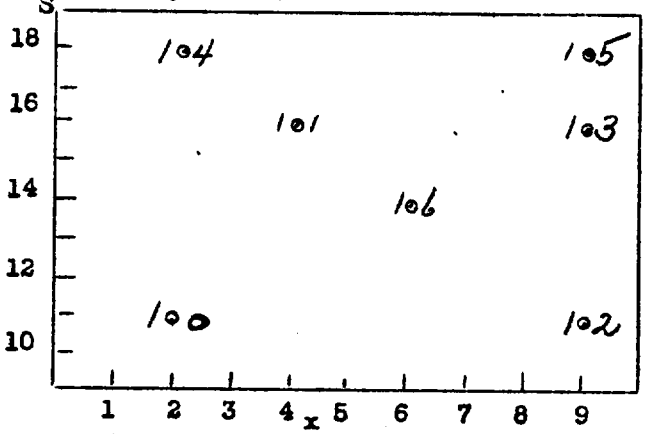
To calculate S when $X=7$; $y=3$; and $m=5$, we follow the same procedure as we did in Problem 1. On Chart 5, the value of "m" at 5 reads "2." On Chart 4, "y" at 3 reads "-3." On Chart 3, "x" at 7 reads "16." Combining them: $16+2-3=15$. This is the answer, which is easy to confirm by arithmetic: $7+3+5=15$.

Graphic Correlation Analysis

Problem 2

Plate 2

Chart 3



Cover Sheet 3: y + m entered

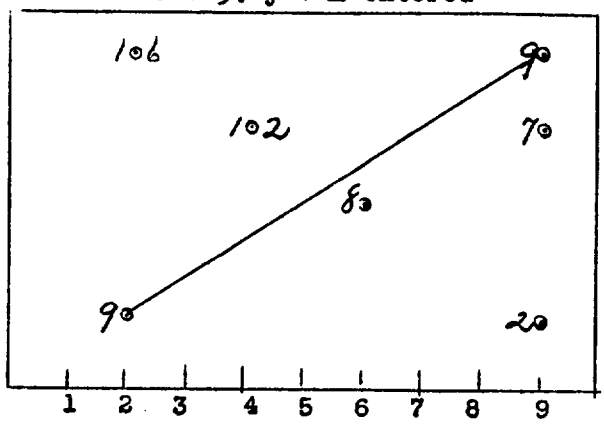
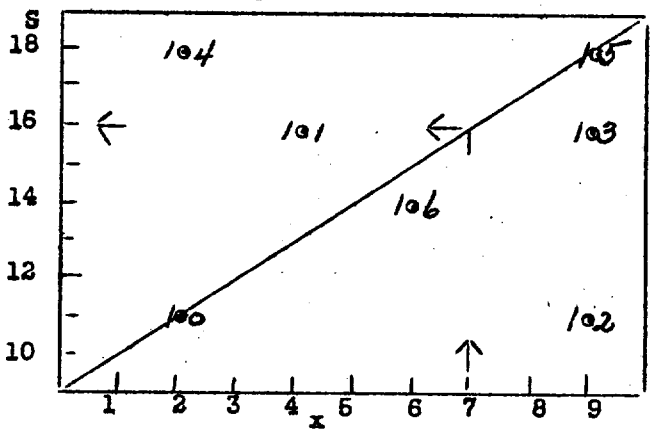


Chart 3 completed



Cover Sheet 3-a: y - m entered

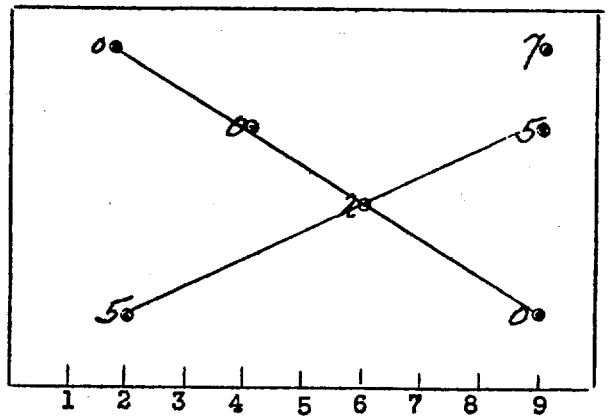


Chart 4: Zs from Chart 1 vs y

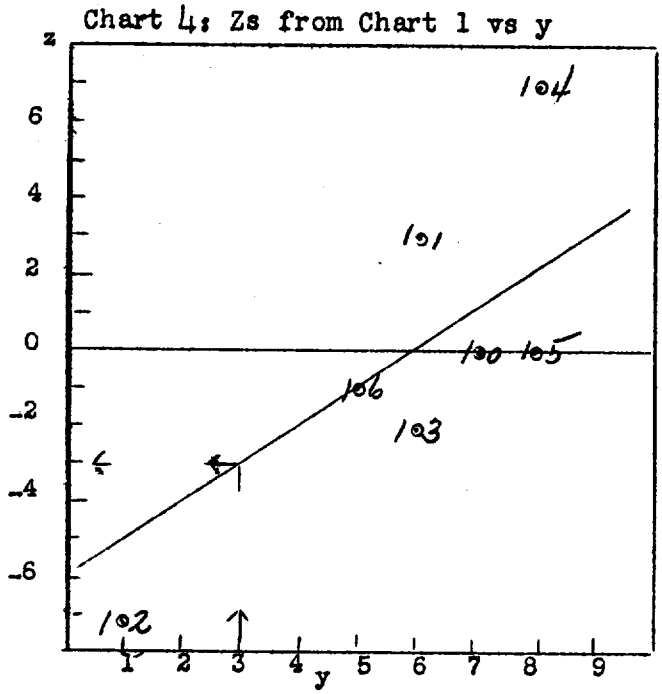
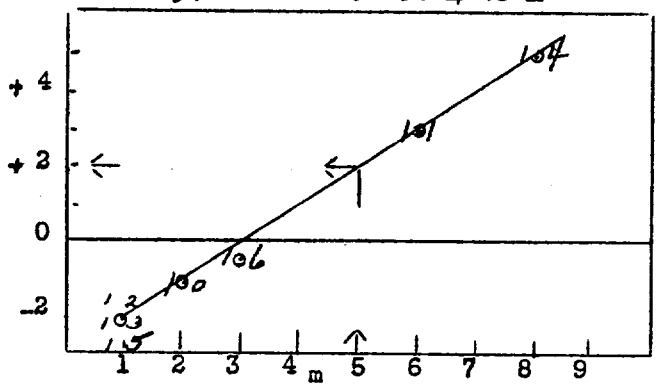


Chart 5: Za from Chart 4 vs m



Data for Problem 2

Observation:	S	x	y	m
10	11	2	7	2
11	16	4	6	6
12	11	9	1	1
13	16	9	6	1
14	18	2	8	8
15	18	9	8	1
16	14	6	5	3

Multiple Curvilinear Relationships

So far we have dealt only with very simple linear relationships. There were two reasons for doing this. First to demonstrate the procedure in a simple way, and second to prove the validity of the method of using data easy to check.

Ordinarily there might be no particular advantage in using the graphic approach with data having linear correlation, which can be analyzed readily and accurately by standard mathematical methods. The graphic approach is helpful, however, even in some cases where the relationships do prove to be linear —— especially when the number of observations is small.

As the number of observations increases, the graphic procedure gets more and more cumbersome, until it finally becomes impracticable. On the other hand, with relatively small samples —— less than about 40 or 50 —— the graphic approach has a degree of flexibility that can be very helpful. This is especially true when dealing with time series —— where occurences not explained by the data at hand may cause a few observations to be very erratic. Plotting the data graphically does not explain such cases, but it does show them up clearly, whereas they can easily be obscured and overlooked if the data are immediately thrown into a standard mathematical analysis without study of individual observations.

It is when relationships are curvilinear that graphic analysis becomes most helpful. However, when one departs from the rigidity of a straight line and assumes that a curvilinear relationship exists, the very flexibility of the graphic method may cause trouble. There is no end to the shapes that can be given to curves, and for that reason the method itself has been attacked by some very able statisticians.

In the next problem, however, it will be demonstrated that the method is adapted to analyzing curvilinear relationships even when they are determined by mathematically constructed curves —— and that reasonably simple curves can be re-constructed free-hand with fair accuracy, from a relatively small number of observations.

Problem 3 (Plates 3 to 11)

This problem of curvilinear analysis is purposely restricted to a small number of observations, to provide a very rigid test. To re-construct free-hand three mathematically-computed curves of multiple relationship, from a total of only twelve observations, would seem to be a fairly strict test for any method. This demonstration is not intended to encourage anybody to construct curves, nor draw conclusions, from an inadequate sample. It is simply intended to demonstrate how to use this method of analysis. One might well question the wisdom of drawing conclusions from a set of curves derived from a sample containing only 12 observations, using 3 independent variables. In the present case, however, these 12 might be considered as constituting the universe, without sampling error, or bias of any kind.

The data in the present problem could not possibly be analyzed properly by any method of linear analysis, because the curvilinearity is very marked.

Multiple Curvilinear Correlation Analysis

Problem 3 Plate 4

Chart 6: Y vs X_1, with 1st approximation Curve "A"

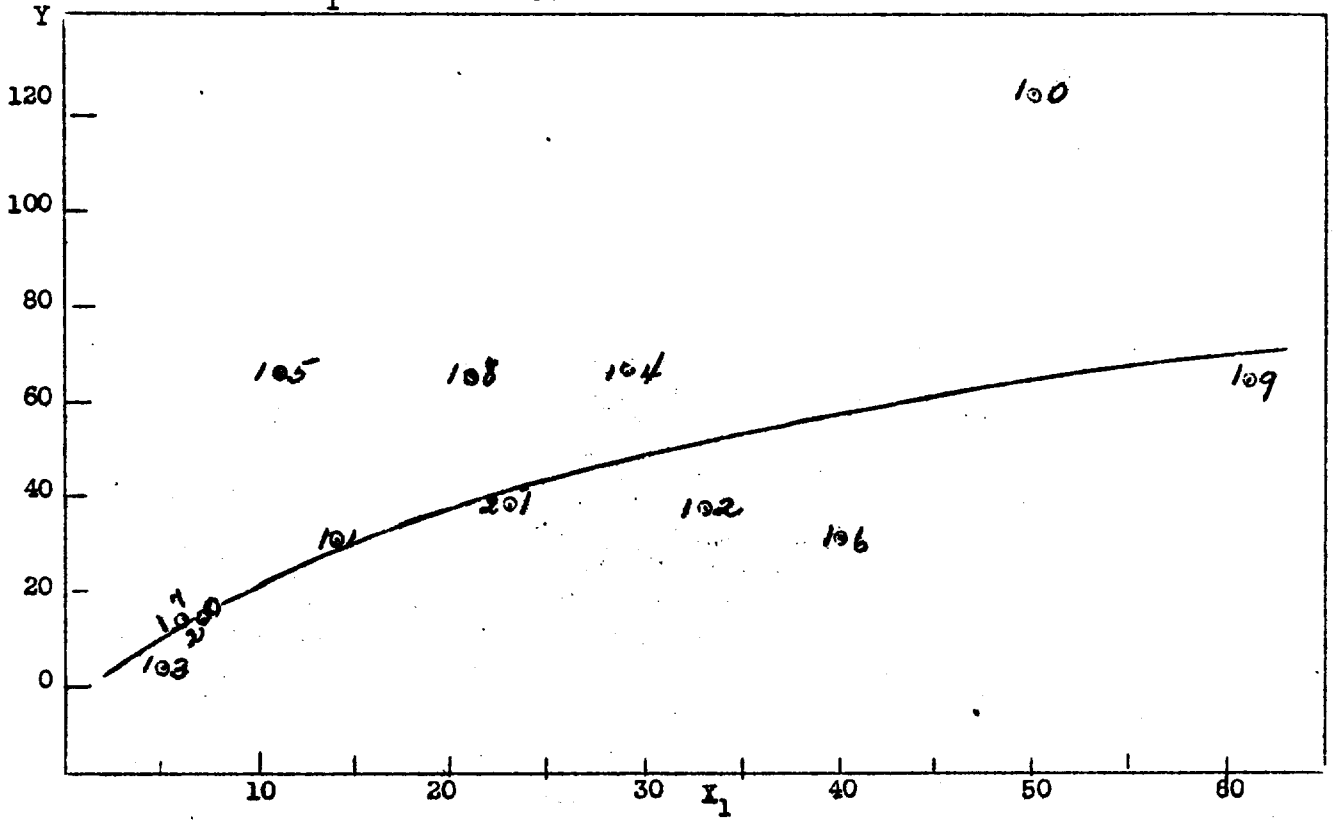
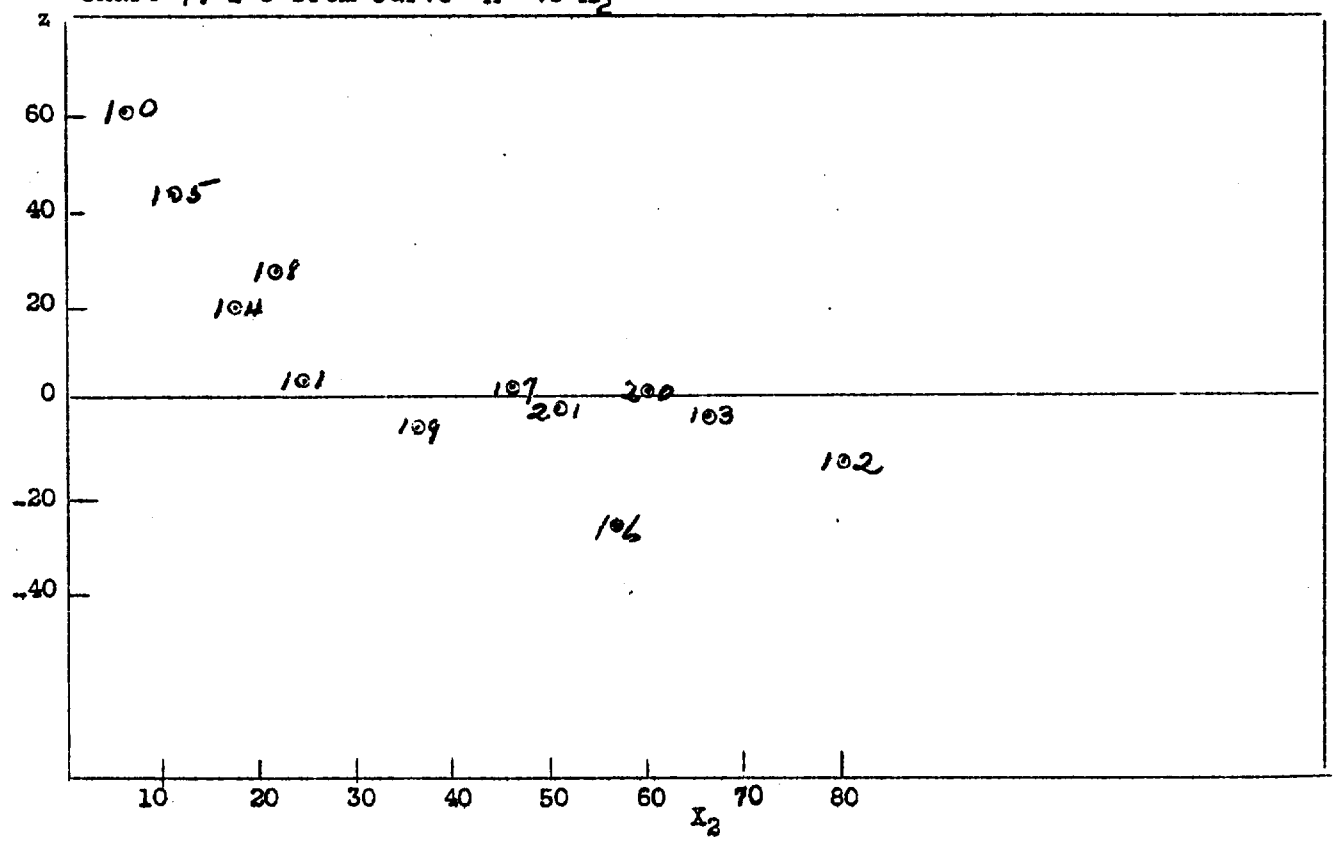


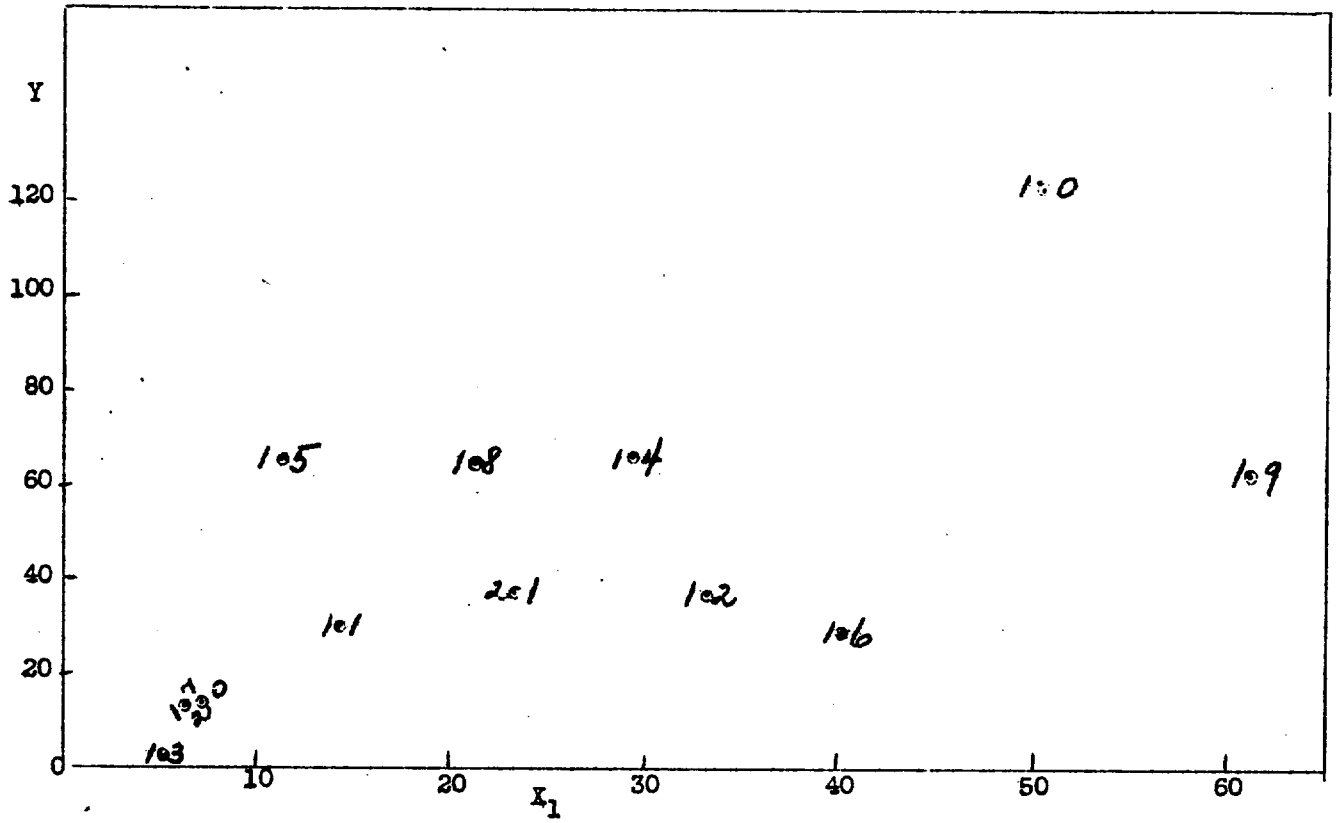
Chart 7: Z's from Curve "A" vs X_2



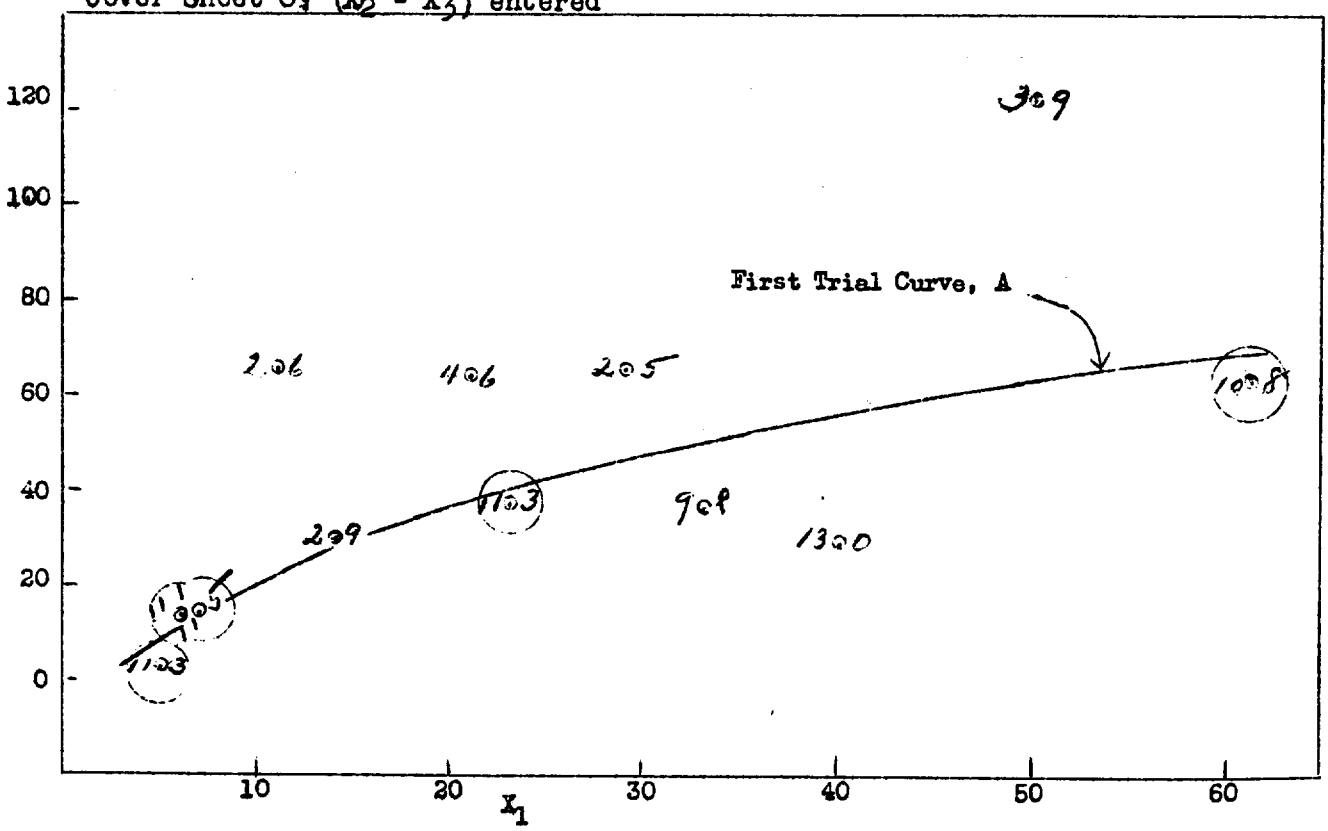
Multiple Curvilinear Correlation Analysis

Problem 3 Plate 3

Chart 6: Y vs X_1



Cover Sheet 6: (X_2 - X_3) entered



The data to be analyzed are as follows:

Observation:	Y	X_1	X_2	X_3
10	123	50	6	33
11	30	14	24	5
12	36	33	80	18
13	3	5	66	47
14	65	29	17	8
15	65	11	11	15
16	29	40	55	75
17	13	6	46	65
18	64	21	21	25
19	63	61	36	72
20	14	7	60	55
21	37	23	51	62

In analyzing curvilinear relationships by the "Bean method" the general procedure is essentially the same as we have just demonstrated in Problems 1 and 2 with linear relationships. When we begin plotting curves, however, we have to do considerable experimenting to determine the proper shape and position of each curve. This gets much harder as the number of variables increases. The operation calls for drawing successive trial curves based on a study of the individual observations, and the position of each observation as it appears on successive charts. The general idea is to so shape all the curves that the residuals from the final curve will be reduced to a minimum.

On Plate 3, Chart 6 shows the values of "Y" plotted vs X_1. As in Problem 2 a Cover Sheet (6) was prepared, with the sums of X_2 and X_3 entered over the dots on Chart 6. (The Cover Sheet with values of $X_2 - X_3$ disclosed little relationship, and is not shown.)

On Cover Sheet 6 the slope of relationship is not so clear as that which showed up in Problem 2, on Cover Sheet 3. In dealing with curved relationships we may expect them to be rather obscure on this first trial sheet --- because simple addition of X_2 and X_3 implies that they have linear relationships.

In the present case we find five observations on Cover Sheet 6 that have about the same values. These have been circled lightly, and are, respectively, 113, 111, 115, 113, and 108. A smooth curve ("A") is plotted through these five points as a first trial. This curve puts the lowest observations on one side and the highest ones on the other.

We need not be disturbed because all the observations are not proportionally distant from the curve, because that often happens when dealing with curved relationships.

Plate 4 shows Curve A transferred to Chart 6, and Chart 7 shows the residuals plotted vs X_2.

Multiple Curvilinear Currelation Analysis

Problem 3

Plate 5

Cover Sheet 7, with X_3 entered

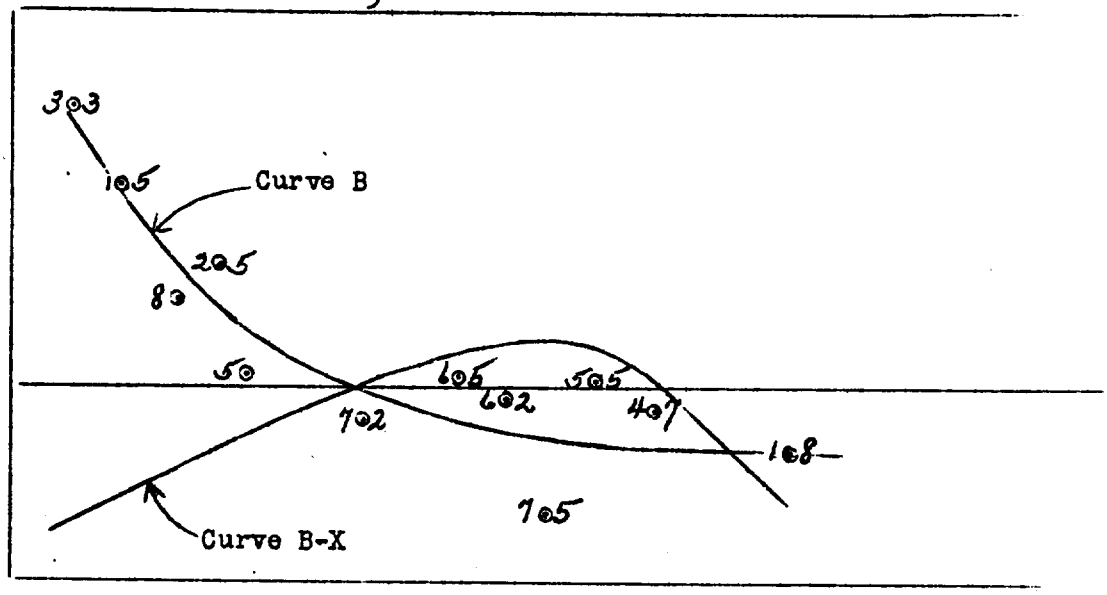
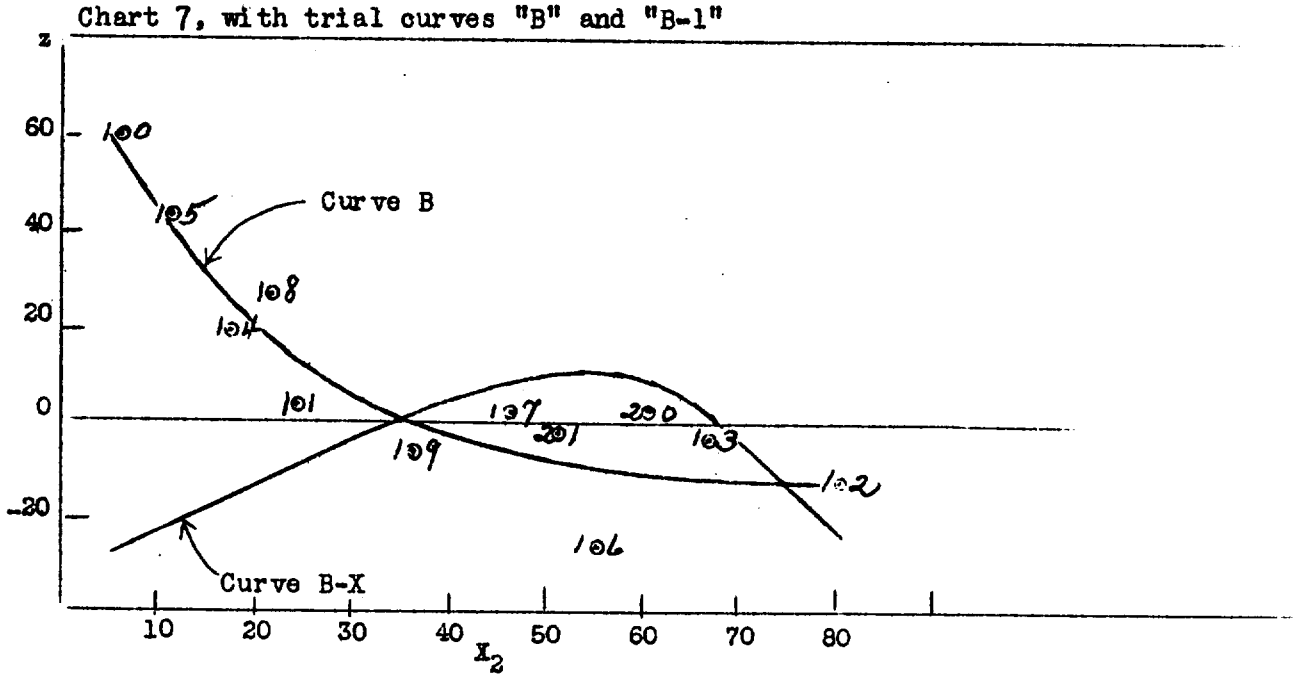


Chart 7, with trial curves "B" and "B-1"



Graphic Multiple Correlation Analysis

Problem 3 Plate 6

Chart 8: Z's from Curve "B" (Chart 7) vs X_3 - With Trial Curve "C" fitted.

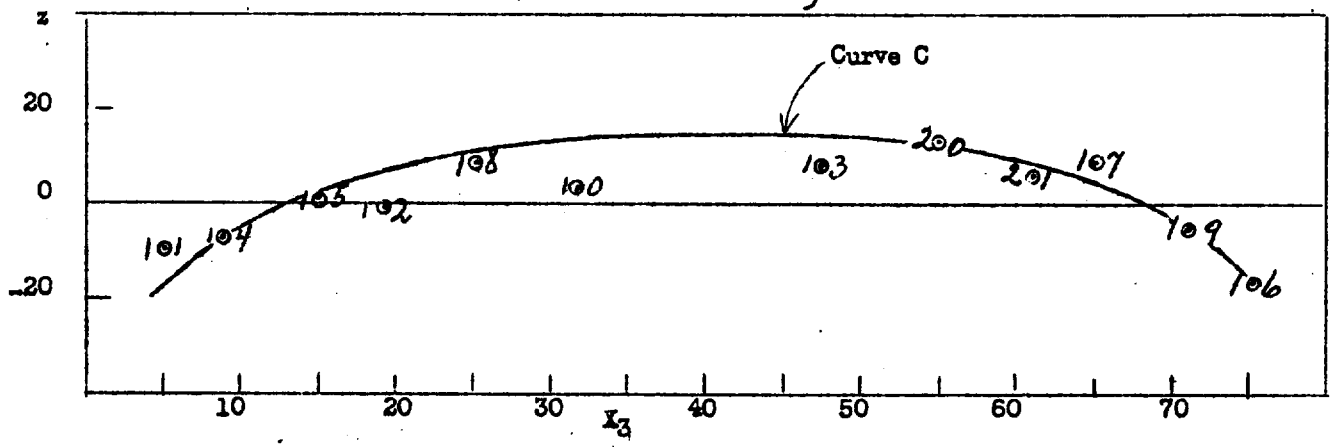
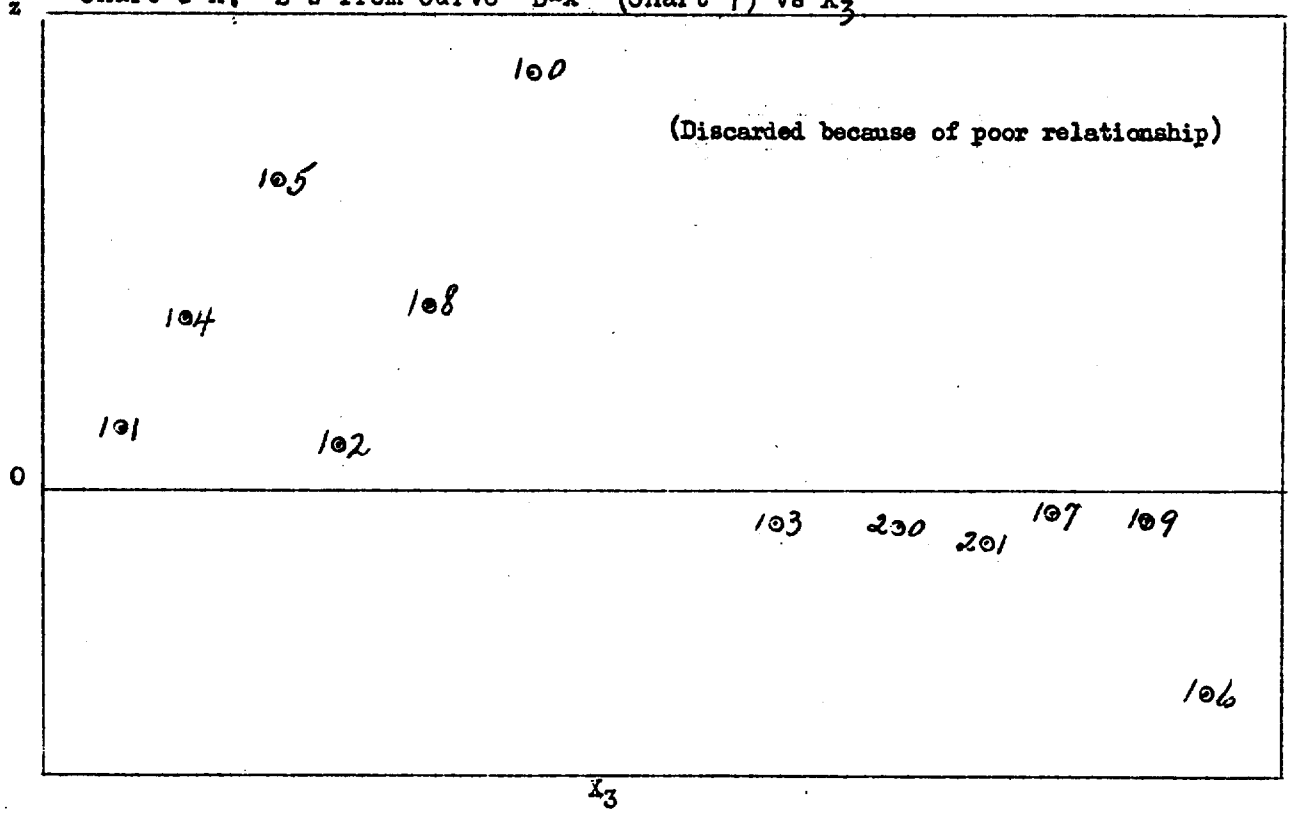


Chart 8-x: Z's from Curve "B-x" (Chart 7) vs X_3



Problem 3 Plate 7

Cover Sheet 6-A: Z's from Curve "C" (Chart 8) measured from Curve "A" (Chart 6)

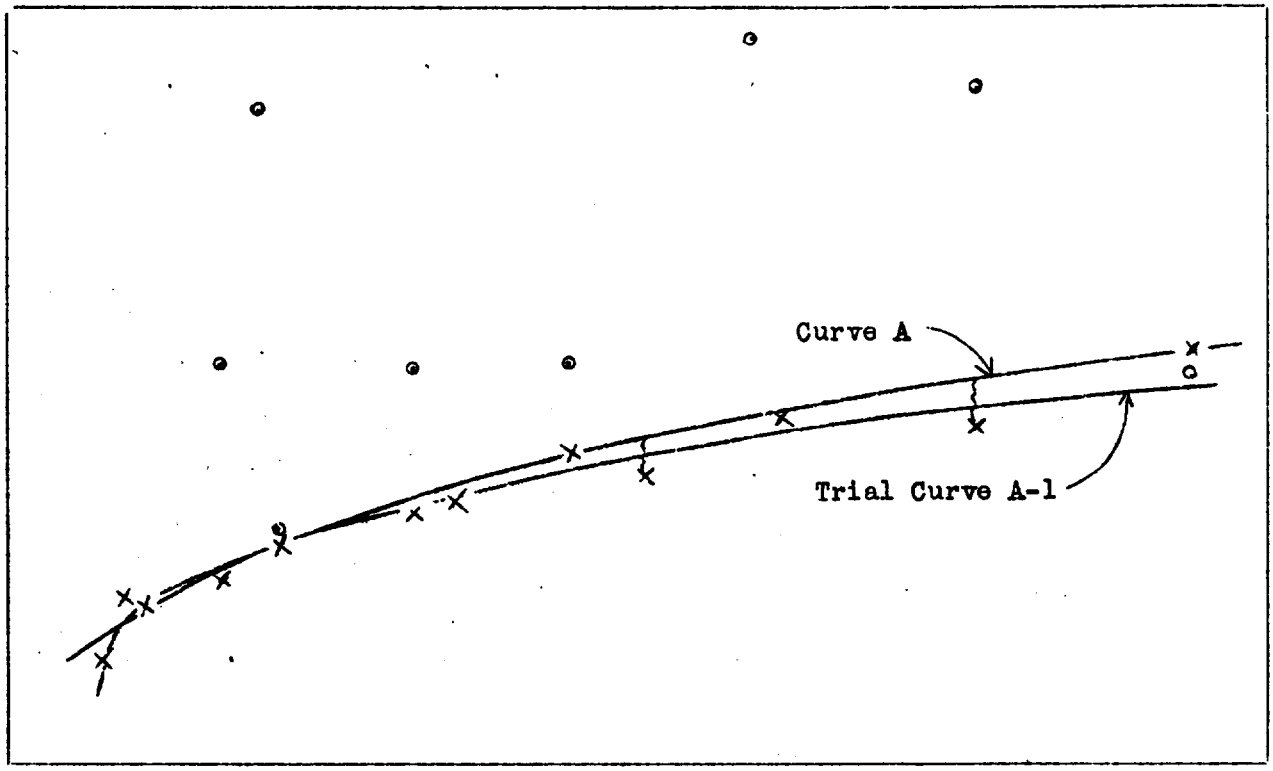


Chart 6-A: Same as Chart 6, with 2nd Trial curve "A-1"

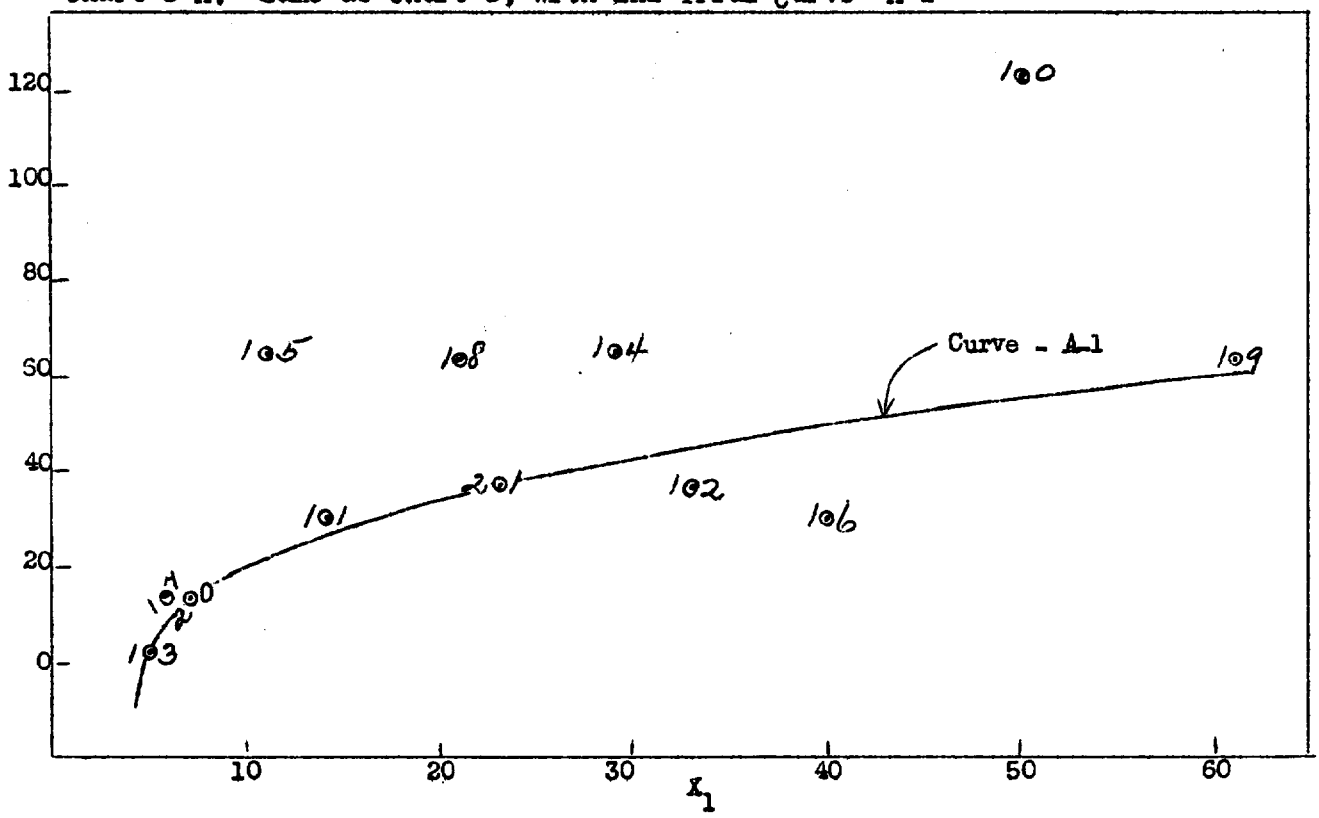


Plate 5 shows the Cover Sheet for Chart 7, with values of X_3 entered. Here we are faced with something of a dilemma. We have no identical values of X_3, and in order to draw a "line of fit" through the general area occupied by all the dots, we go through the low and high values rather indiscriminately. Such a curve, however, does reduce the residuals very materially. This is illustrated by Curve "B". On the other hand, we could draw a curve with the general shape of Curve "B-X" that puts most of the low values on one side and the high values on the other side. Such a curve obviously, would increase the residuals, rather than reduce them.

In dealing with a tangible statistical problem, rather than with [redacted] curves that have no meaning, we should know enough about the problem itself to be able to deduce by logic the general direction that the respective curves might be expected to take, even though we don't know what shape the curve might have. In the present case we have no such logic to help us, so we shall experiment with both curves and see what happens.

On Plate 6, Chart 8 shows the residuals from Chart 7 vs X_3, using Curve B. Chart 8-X shows the results from using Curve B-X. It is readily apparent that Chart 8 gives a fairly close curve of relationship, while the scatter on Chart 8-X does not. Accordingly we discard Curve B-X from further consideration.

So far, the only information we have about the 3 curves of X_1, X_2 and X_3 is that they have been computed mathematically. Mathematical curves are naturally rather "smooth", or symmetrical, without sharp kinks and irregularities --- unless they are plotted from very complex equations. For that reason we make no attempt to follow every slight irregularity in the arrangement of the dots, but try to shape our curves symmetrically and without sharp breaks.

"Polishing" the Curves.

Having now plotted 3 "trial" curves "A", "B" and "C", for X_1, X_2 and X_3, the next step is to experiment further to adjust their shapes and positions in such a way that the final residuals from the X_3 curve will be minimized. We might call this a "polishing" process. This was not necessary in Problems 1 and 2, but does become necessary with most problems.

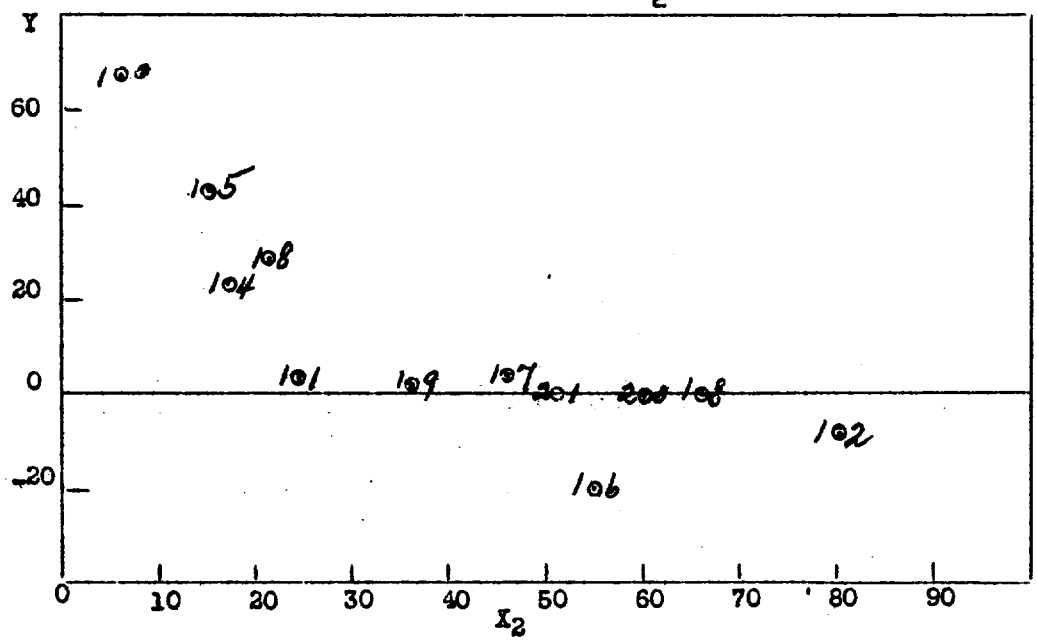
The first step in "polishing" the curves is to assume that Curve "C" is correct, and find out where the other two curves would have to be in order to eliminate the residuals left on Chart 8. Accordingly we measure the residuals from Curve C (Chart 8), and plot them from Curve "A" on Chart 6. This is shown on Cover Sheet 6-A, Plate 7, the "X" marks being the points where the respective residuals land when measured from Curve "A".

Then we plot a new curve through the "X" marks. This is the second trial curve "A-1". This new curve is transferred to Chart 6, on Chart "6-A", and residuals are plotted against X_2 on Chart "7-A", Plate 8.

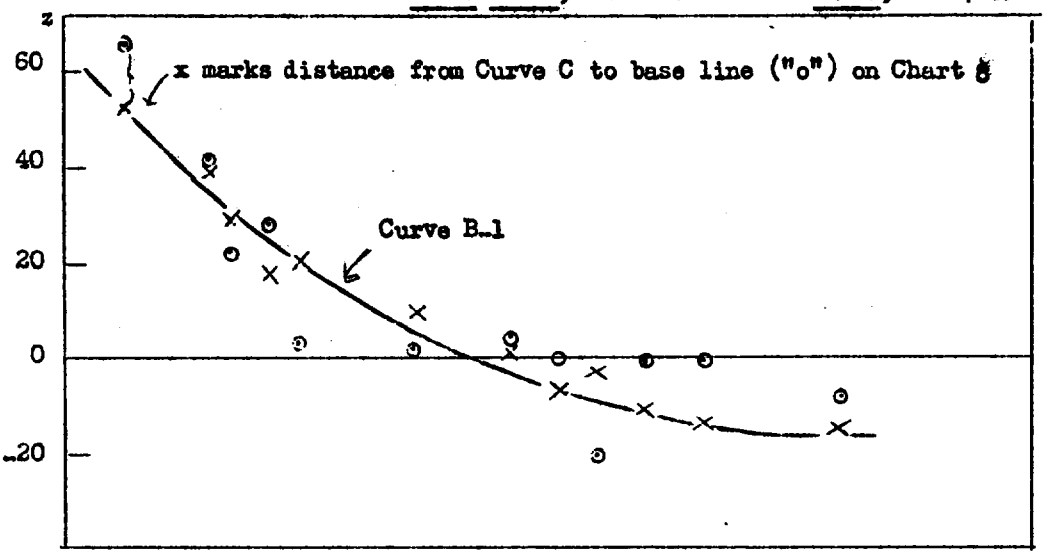
Since part of the residuals from Curve C have now been corrected by Curve A-1, we have no further use for Curve B (Plate 5). To find where the curve should be plotted on Chart 7-A we go first to Chart 8 (Plate 6) and measure the distances from the Base Line ("0") to Curve C --- note that we are not measuring the residuals from the curve to the dots this time, but from the curve to the base line.

Problem 3 Plate 8

Chart 7-A: Z's from Curve "A-1" vs X_2



Cover Sheet 7-A: With distances from Curve "C" (Chart 8) to base line, measured from dots, on 7-A



Problem 3

Plate 9

Chart 7-A, with Curve "B-1" entered

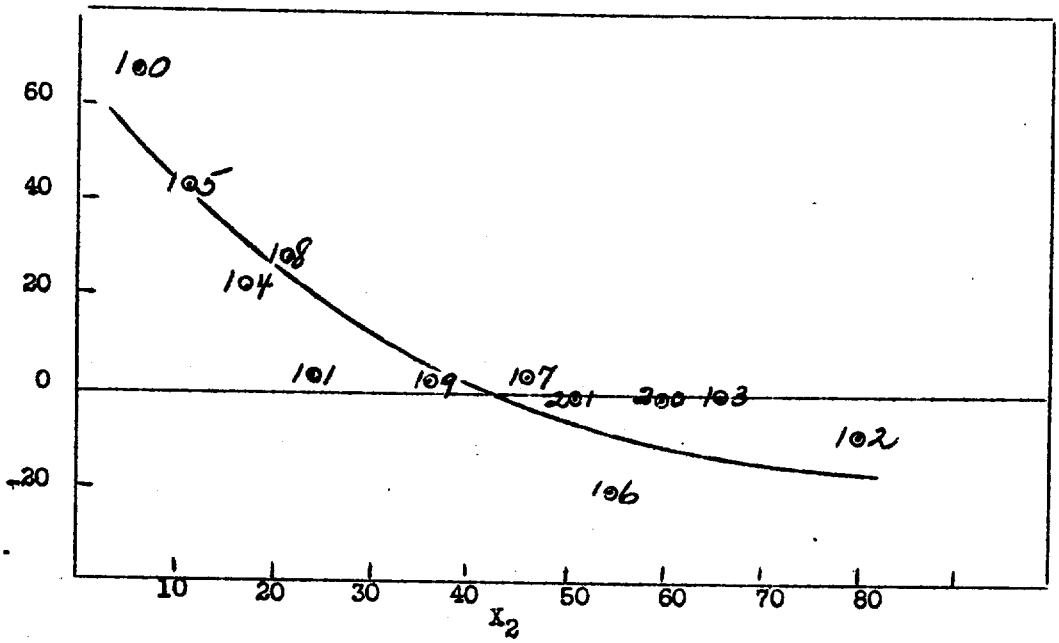
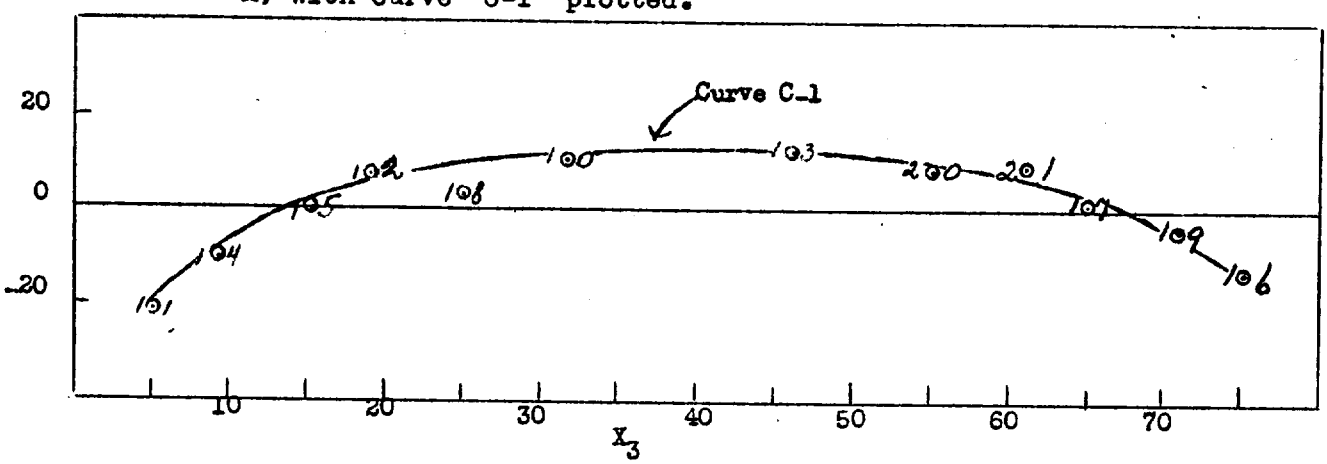


Chart 8-A, with Curve "C-1" plotted.



Multiple Curvilinear Correlation Analysis

Problem 3 Plate 10

Chart 6-B: Third Trial Curve "A-2"

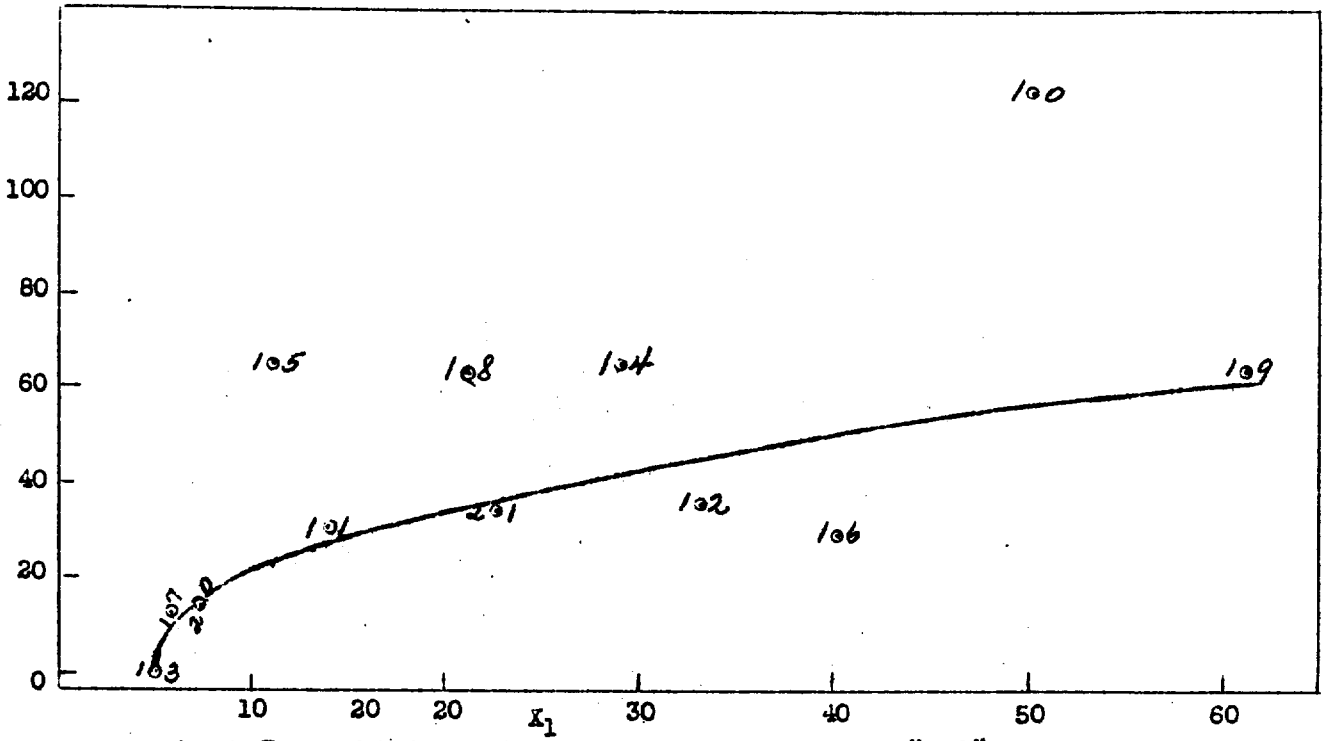


Chart 7-B: Third Trial Curve "B-2"

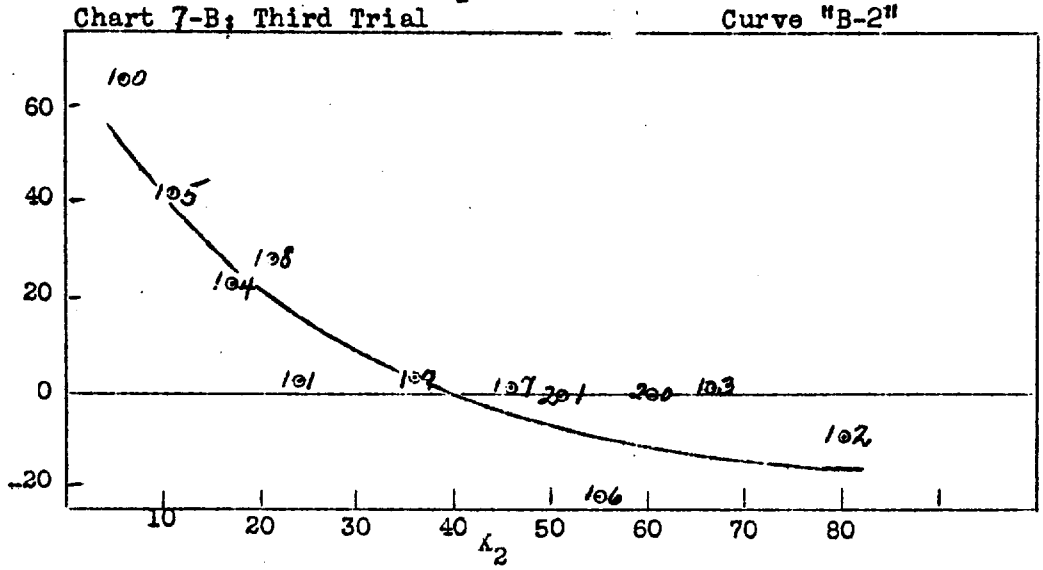
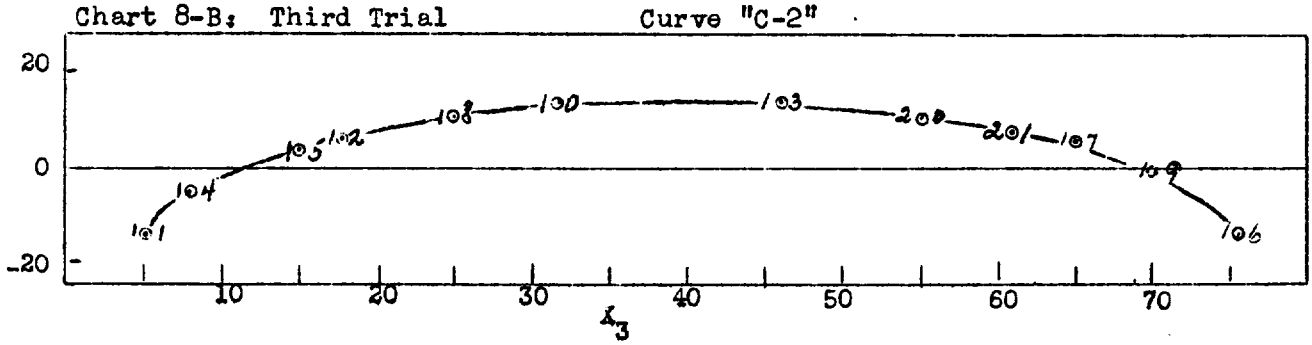


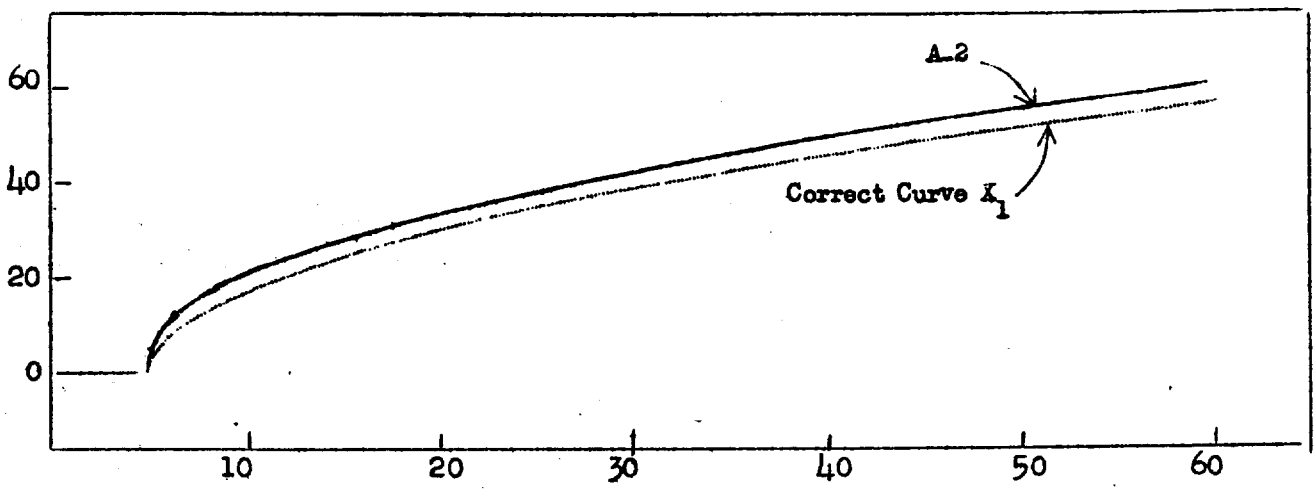
Chart 8-B: Third Trial Curve "C-2"

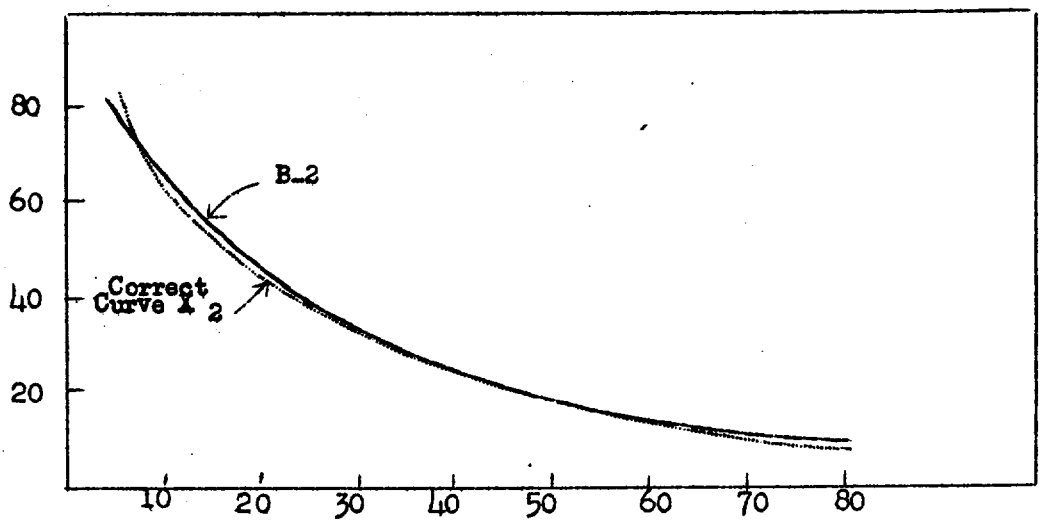


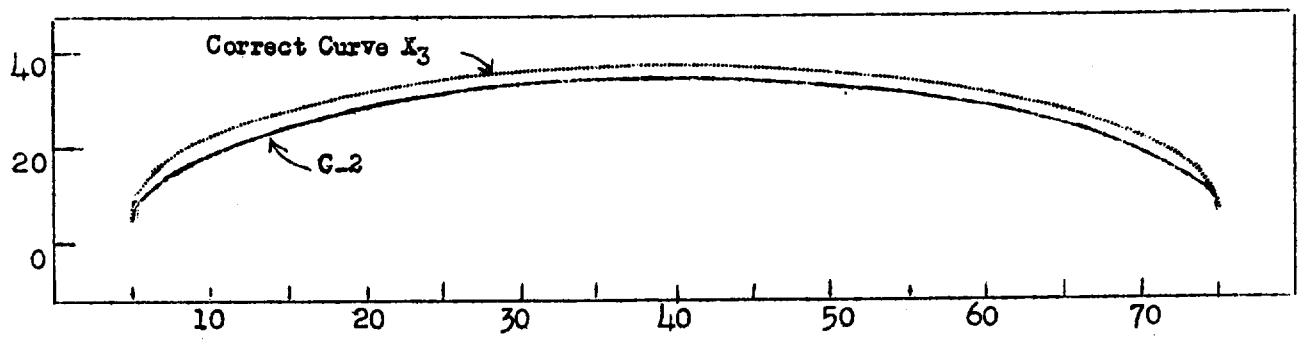
Multiple Curvilinear Correlation Analysis

Problem 3 Plate 11

Comparisons of free-hand curves "A-2", "B-2" and "C-2" with the correct computed curves which determined the data, X_1, X_2, and X_3.







These measurements are then transferred to Cover Sheet 7-A, measuring them from the respective dots on Chart 7-A. They are marked "X" on Cover Sheet 7-A to distinguish them from the dots. These "X" marks show where the curve for X_2 should be, in order for the residuals to fall on Curve C when carried forward to Chart 8 (Still assuming that Curve C is correct).

We now plot a new curve "B-1" to fit the "X" marks on Cover Sheet 7-A. This is transferred to Chart 7-A on Plate 9. Residuals from Curve B-1 are plotted against X_3 on Chart 8-A.

The dots on Chart 8-A follow a somewhat different pattern than on Chart 8, and indicate that Curve C should be altered a little. Accordingly, Curve "C-1" is plotted on this chart.

Using this new curve "C-1" the polishing process is repeated as before, and slight modifications are made in Curves "A-1", "B-1" and "C-1". Without showing the cover sheets, the resulting curves "A-2", "B-2" and "C-2" are shown on Plate 10. With these curves, practically no residuals are left on the final chart for X_3, so we can go no further with the data we have.

To measure the accuracy of the analysis, Plate 11 shows these final curves in comparison with the original mathematical curves from which the data for Y, X_1, X_2, and X_3 were read. It will be seen that the curves all have nearly the correct shape, but "A-2" is a little too high, while "C-2" is too low by about the same amount. These compensating errors are due to the peculiarities of the sample, and would be disclosed when a larger number of observations were read from the original curves. Although we assumed no sampling error to begin with, we find that there actually is some --- which is not surprising, under the circumstances.

SUMMARY

The various steps in this process of making a graphic analysis of curvilinear miltiple correlation with three independent variables, X_1, X_2 and X_3, are:

1. Plot the dependent variable on the vertical axis, vs X_1 on the horizontal axis (Chart 1).
2. On a cover sheet write-in the sum of $X_2 + X_3$, (or $X_2 - X_3$) for each observation, over the respective dots on the first chart.
3. Plot a trial curve (Curve A) through the data on this cover sheet and transfer this curve to Chart 1.
4. Plot the residuals from the trial curve vs X_2 (Chart 2).
5. On a cover sheet insert the X_3 values over the respective dots on Chart 2.
6. Plot a trial curve (Curve B) through the data on this cover sheet, and transfer this to Chart 2.
7. Plot the residuals from Chart 2 vs X_3 (Chart 3).
8. Draw a curve-of-fit through the data on Chart 3.

To improve, or "polish" the trial curves;

9. On a cover sheet over Chart 1, measure the residuals on Chart 3 from the curve on Chart 1.

10. Plot a new curve (if necessary) through the spots indicated by the residuals on this cover sheet. Transfer this new curve to Chart 1.

11. Plot the residuals from the new curve vs X_2 (Chart 2-A).

12. On a cover sheet over Chart 2-A measure from each dot the distance on Chart 3 from the curve to the base line.

13. Plot a new curve through the data on this cover sheet, and transfer the curve to Chart 2-A.

14. Plot the residuals from Chart 2-A vs X_3 and draw a new curve to fit.

Repeat the polishing process (steps 9 to 14) until no further reduction can be made in the residuals on the X_3 chart.

It might be observed that one readily convenient way to measure progress in reducing the residuals is to measure each one (using the original Y scale) and compute the sum of the squares of the residuals. The objective is to reduce this sum of squares to a minimum.

In reading the value of "Y" from any given values of X_1, X_2, and X_3, the same procedure is used as is described in connection with Problem 2.