UMAP

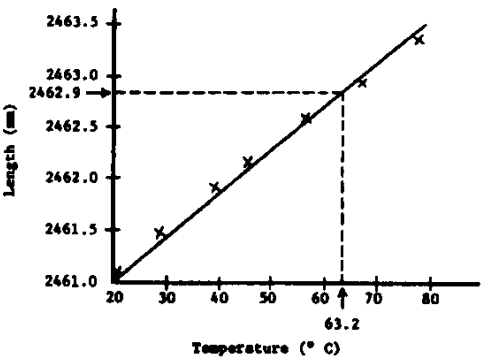
Modules in Undergraduate Mathematics and its Applications

Published in cooperation with the Society for Industrial and Applied Mathematics, the Mathematical Association of America, the National Council of Teachers of Mathematics, the American Mathematical Association of Two-Year Colleges, The Institute of Management Sciences, and the American Statistical Association.

Module 321

Curve Fitting via the Criterion of Least Squares

John W. Alexander, Jr.



Applications of Algebra and Elementary Calculus to Curve Fitting

COMAP, Inc., Suite 210, 57 Bedford Street, Lexington, MA 02173 (617) 862–7878

Intermodular Description Sheet: UMAP Unit 321

Title: CURVE FITTING VIA THE CRITERION OF LEAST SQUARES

Author: John W. Alexander, Jr.
Corporate Actuarial Department
Connecticut Mutual Life Insurance Company
Hartford, CT 06115

Review Stage/Date: III 9/3/79

Classification: APPL ALG & ELEM CALC/CURVE FITTING

Suggested Support Materials: A computer terminal on line to a system with BASIC compiler (to be used for the appendix).

Prerequisite Skills:

1. Be able to do partial differentiation.
2. Be able to maximize functions.
3. Know how to solve simultaneous equations by elimination or substitution for 2 x 2 cases.
4. Know how to graph elementary, exponential, and logarithmic functions.

Output Skills:

1. Be able to construct scatter diagrams.
2. Be able to choose an appropriate function to fit specific data.
3. To understand the underlying theory of the method of least squares.
4. To be able to use a computer program to do desired curve-fitting.
5. Be able to use augmented matrix approach to solve simultaneous equations.

CURVE FITTING VIA THE CRITERION OF LEAST SQUARES

by

John W. Alexander, Jr.
Corporate Actuarial Department
Connecticut Mutual Life Insurance Company
Hartford, Connecticut 06115

TABLE OF CONTENTS

1. INTRODUCTION 1

2. SCATTER DIAGRAMS 1

3. THE LINE OF REGRESSION 4

4. COEFFICIENT OF CORRELATION 9

5. REGRESSION FOR LOGARITHMIC SCATTERS 13

6. REGRESSION FOR EXPONENTIAL SCATTERS 19

7. POLYNOMIAL SCATTERS 24

8. MODEL EXAM 30

9. ANSWERS TO MODEL EXAM 31

10. ANSWERS TO EXERCISES 35

APPENDIX 42

MODULES AND MONOGRAPHS IN UNDERGRADUATE
MATHEMATICS AND ITS APPLICATIONS (UMAP) PROJECT

The goal of UMAP was to develop, through a community of users and developers, a system of instructional modules in undergraduate mathematics and its applications to be used to supplement existing courses and from which complete courses may eventually be built.

The Project was guided by a National Advisory Board of mathematicians, scientists, and educators. UMAP was funded by a grant from the National Science Foundation and is now supported by the Consortium for Mathematics and Its Applications (COMAP), Inc., a non-profit corporation engaged in research and development in mathematics education.

COMAP STAFF

Paul J. Campbell	Editor
Solomon A. Garfunkel	Executive Director, COMAP
Laurie Aragón	Development Director
Joseph Owumi	Business Manager
Roland D. Cheyney	Project Manager
Philip A. McGaw	Production Manager
Laurie M. Holbrook	Copy Editor
Dale Horn	Design Assistant
Robin Altomonte	Distribution

The Project would like to thank Thomas R. Knapp and Roger Carlson, members of the UMAP Statistics Panel, and Lee H. Minor, Nathan Simms, Jr., and Charles Votaw for their reviews, and all others who assisted in the production of this unit.

This material was prepared with the support of National Science Foundation Grant No. SED76–19615 A02. Recommendations expressed are those of the author and do not necessarily reflect the views of the NSF or of COMAP.

CURVE FITTING VIA THE CRITERION OF LEAST SQUARES

1. INTRODUCTION

In many instances, we wish to be able to predict the outcome of certain phenomena. For example, we may want to know which students in a graduating high-school class will do well in their first year of college.

One way to get a measure or at least an indication would be to observe the high-school grades in English of 20 or so students who have gone to college. If we match the students' English grades with their grade point average after one semester, we would be able to see if good grades in English matched with high grade point averages.

If the "correlation" is high, then, we might wish to assert that students who do well in high-school English do well in college. There may be exceptions of course. We may want to look at other indicators (e.g., math grades) but, the point is, we wish to look at two or more statistics on the same individual, and we are interested to know how these statistics relate.

Ideas of the sort alluded to above are the subject of this module.

2. SCATTER DIAGRAMS

Many statistical problems are concerned with more than a single characteristic of an individual. For instance, the weight and height of a number of people could be recorded so that an examination of the relationship between the two measurements could be made. As a further example, consider how the length of a copper rod relates to its temperature.

TABLE 1

Temperature (° C)	Length (mm)
x	y
20.1	2461.16
28.2	2461.49
38.5	2461.88
44.6	2462.10
57.4	2462.62
66.2	2462.93
78.1	2463.38

When we draw a scatter diagram letting the horizontal axis be the scale for the temperature and the vertical axis the scale for the length, we note that the plotted points lie very close to a straight line. It is, therefore, reasonable to make a quick and accurate estimate of the length of the rod for any temperature between 20.1° and

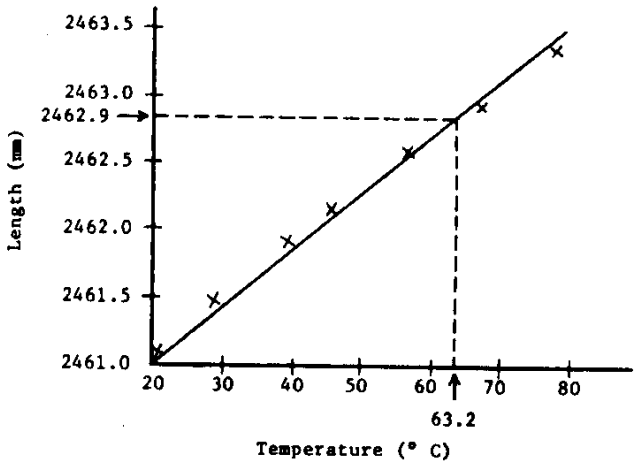


Figure 1.

78.1°.* For example, if the temperature was 63.2° the dotted lines in Figure 1 indicate that the corresponding point on the line gives a length of approximately 2462.9 mm.

Let us explore another example that gives us a scatter diagram where the points are more scattered. Table 2 gives us the weight in grams, x, and the length of the right hind foot in millimeters, y, of a sample of 14 adult field mice.

TABLE 2

Weight (g) x	Length (mm) y
22.3	23.0
16.0	22.6
18.8	23.2
18.2	22.5
16.0	22.2
20.4	23.3
17.9	22.8
19.4	22.4
16.9	21.8
17.6	22.4
16.5	22.4
18.8	21.5
17.2	21.9
20.4	23.3

The point in Figure 2 that is circled indicates where two points of the data coincide. The points here are much more scattered than those of the previous set. It would be extremely difficult to determine which straight line best fits this set of points. In fact, if a number of people were to attempt to fit a line to these points, there is little doubt that each person would come up with a different line. What we need is a mathematical method for determining the line that comes "closest" to all of the points.

*We are not in a position to speculate about values outside of this range.

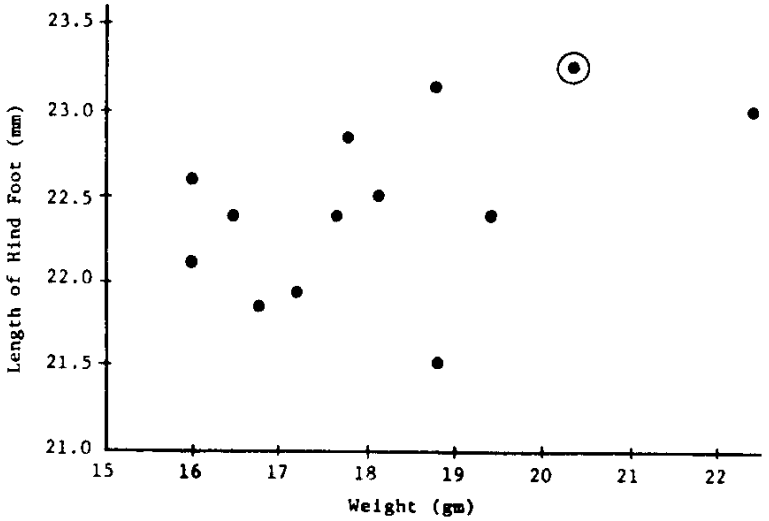


Figure 2.

3. THE LINE OF REGRESSION

The criterion traditionally used to define a "best" fit dates back to the nineteenth century French mathematician Adrien Legendre. It is called the *criterion*, or *method, of least squares*. This criterion requires the line of regression which we fit to our data to minimize *the sum of the squares of the vertical deviations (distances) from the points to the line.* In other words, the method requires the sum of the squares of the distances represented by the solid line segments of Figure 3 to be as small as possible.

From the figure, we see that the actual grade received for a student who studied 11 hours was 79. Reading from the line of regression we predict a grade of about 71.

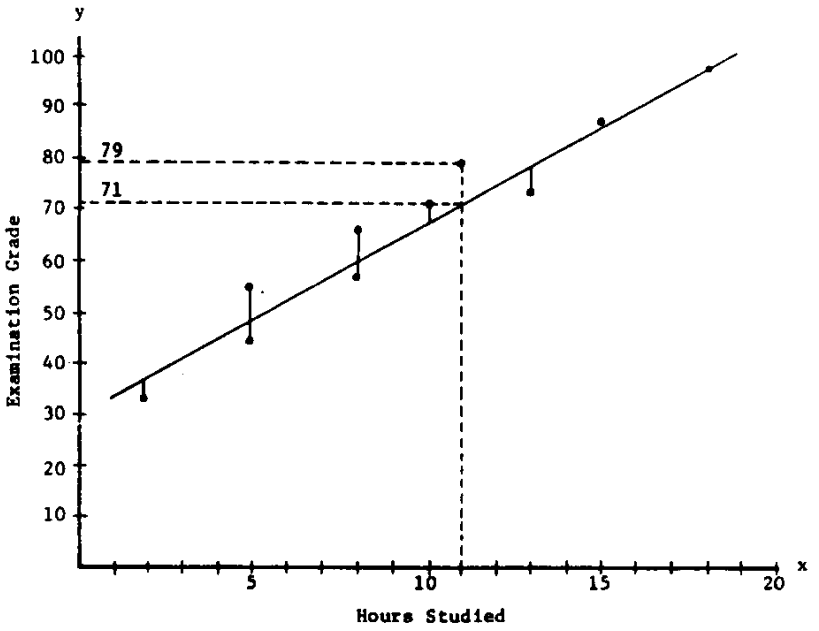


Figure 3. Line of regression fitted to data on hours studied and examination grades.

Observe that any line can be expressed:

(1) $y = bx + c$

or

(2) $x = b'y + c'$,

where the b's represent the slope of the line and the c's are interpreted as the intercept of the axis.

If we consider Equation (1), knowing the values of b and c will allow us to compare the actual values in the y column with bx + c. We take the difference in each case and square the result. Consider the values of x and y in Table 3.

TABLE 3

x	y	bx + c	[Difference]^2
25	30	b25 + c	[30 - (25b + c)]^2
30	46	b30 + c	[46 - (30b + c)]^2
50	51	b50 + c	[51 - (50b + c)]^2
20	28	b20 + c	[28 - (20b + c)]^2
70	48	b70 + c	[48 - (70b + c)]^2
80	88	b80 + c	[88 - (80b + c)]^2
91	75	b91 + c	[75 - (91b + c)]^2
46	52	b46 + c	[52 - (46b + c)]^2
35	35	b35 + c	[35 - (35b + c)]^2
25	28	b25 + c	[28 - (25b + c)]^2
80	95	b80 + c	[95 - (80b + c)]^2

We add up all of these squared differences. It must then be determined what values of b and c must be used in order to have a line such that the sum of the vertical distances from the line to the data points is at a minimum.

The Problem: Find the values of b and c such that the sum indicated below is a minimum.

$$\sum D^2 = (30-25b-c)^2 + (46-30b-c)^2 + (51-50b-c)^2 + (28-20b-c)^2$$
$$+ (48-70b-c)^2 + (88-80b-c)^2 + (75-91b-c)^2 + (52-46b-c)$$
$$+ (35-35b-c)^2 + (28-25b-c)^2 + (95-80b-c)^2.$$

The symbol sigma can be employed on both sides of the equation above (i.e., $\Sigma D^2 = \sum_{i=1}^{n} (y_i - bx_i - c)^2$). Since $\sum D^2$ is a function of b and c we can write

$$\sum D^2 = f(b,c) = \sum_{i=1}^{n} (y_i - bx_i - c).$$

To find our desired minimum we find the partial derivatives with respect to b and c and set the results equal to zero. We obtain two equations in two unknowns

which we solve simultaneously. This gives us the desired values of b and c and thus our line of best fit (the line of regression).

Trace through the actual development given below.

$$f(b,c) = \sum_{i=1}^{n} (y_i - bx_i - c)^2$$

$$\frac{\partial f}{\partial b} = \sum_{i=1}^{n} 2(y_i - bx_i - c)(-x_i) = 0$$

$$= \sum_{i=1}^{n} (-2y_i x_i + 2bx_i^2 + 2cx_i) = 0,$$

finally,

$$\sum_{i=1}^{n} bx_i^2 + \sum_{i=1}^{n} cx_i = \sum_{i=1}^{n} x_i y_i.$$

To continue with the other derivatives:

$$\frac{\partial f}{\partial c} = \sum_{i=1}^{n} 2(y_i - bx_i - c)(-1) = 0$$

$$= 2\sum_{i=1}^{n} (-y_i + bx_i + c) = 0$$

and

$$\sum_{i=1}^{n} bx_i + \sum_{i=1}^{n} c = \sum_{i=1}^{n} y_i.$$

Thus our two equations which are traditionally called *normal equations* are:

$$(3) \qquad b\sum_{i=1}^{n} x_i^2 + c\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} x_i y_i.$$

$$(4) \qquad b\sum_{i=1}^{n} x_i + nc = \sum_{i=1}^{n} y_i.$$

In order to solve these equations, we must calculate the indicated sums as is done in Table 4. We have also included the table of y_i^2's because we can use the sum $\sum_{i=1}^{n} y_i^2$ to find the line of regression $x = b'y + c'$.

The normal equations for this line are obtained merely by interchanging x and y in the original two equations (3) and (4).

TABLE 4

x	y	x^2	y^2	xy
25	30	625	900	750
30	46	900	2116	1380
50	51	2500	2601	2550
20	28	400	784	560
70	48	4900	2304	3360
80	88	6400	7744	7040
91	75	8281	5625	6825
46	52	2116	2704	2392
35	35	1225	1225	1225
25	28	625	784	700
80	95	6400	9025	7600
552	576	34372	35812	34382

As an example, for $x = b'y + c'$ we have:

(3') $$b' \sum_{i=1}^{n} y_i^2 + c' \sum_{i=1}^{n} y_i = \sum_{i=1}^{n} y_i x_i .$$

(4') $$b' \sum_{i=1}^{n} y_i + nc' = \sum_{i=1}^{n} x_i .$$

From Table 4, our equations become:

(3) $$34372b + 552c = 34382.$$

and

(4) $$552b + 11c = 576.$$

Thus,

$$11c = 576 - 552b$$

$$c = \frac{576 - 552b}{11}.$$

Substituting the value of c into (3) we get

$$34372b + 552\left(\frac{576 - 552b}{11}\right) = 34382$$

$$378092b + 317952 - 304704b = 378202$$

$$73388b = 60250$$

$$\therefore \quad b = 0.8209789$$

$$\text{and} \quad c = 11.165422.$$

Hence, $y = 0.8209789x + 11.165422$. Similarly,

$$(3') \qquad 35812b' + 576c' = 34382$$

$$(4') \qquad 576b' + 11c' = 552$$

$$\therefore \quad 11c' = 552 - 576b'$$

$$c' = \frac{552 - 576b'}{11}.$$

Substituting in (3') we get

$$35812b' + 576\left(\frac{552 - 576b'}{11}\right) = 34382$$

$$393932b' + 317952 - 331776b' = 378202$$

$$62156b' = 60250$$

$$\therefore \quad b' = 0.9693352$$

and

$$c' = -0.5760972$$

$$\therefore \quad x = 0.9693352y - 0.5760972.$$

So, we now have the lines of best fit with respect to y and with respect to x. (See Figures 4a and 4b.) We can use either one, depending on our needs. Further than that, having the two lines allows us to calculate what is called the coefficient of correlation.

4. COEFFICIENT OF CORRELATION

In order to get a numerical indicator of how well the two sets of scores compare, we take the geometric mean of the slopes of the two lines of regression (i.e., $r = \pm\sqrt{bb'}$).

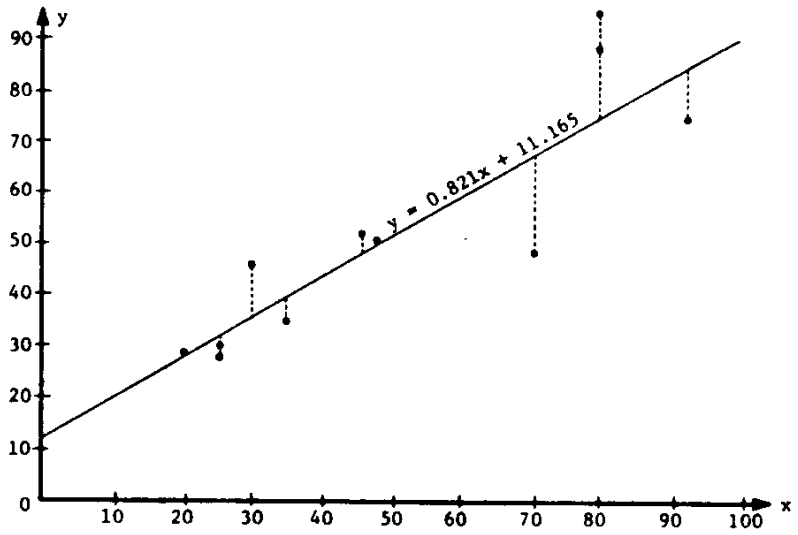


Figure 4a. Regression of y on x.

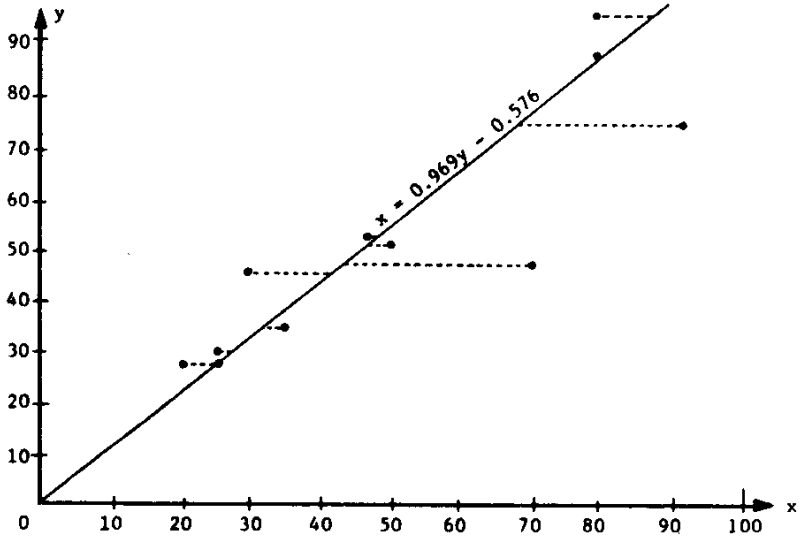


Figure 4b. Regression of x on y.

The sign is chosen to be negative if both slopes are negative, and positive if both slopes are positive. This value, which ranges from -1 to 1 is called the coefficient of correlation. If we have good correlation the value r is close to 1. Poor correlation is indicated by a value near 0. If high values of one characteristic are associated with low values of the other, the correlation is considered negative. Observe the distribution of points in the graphs of Figure 5.

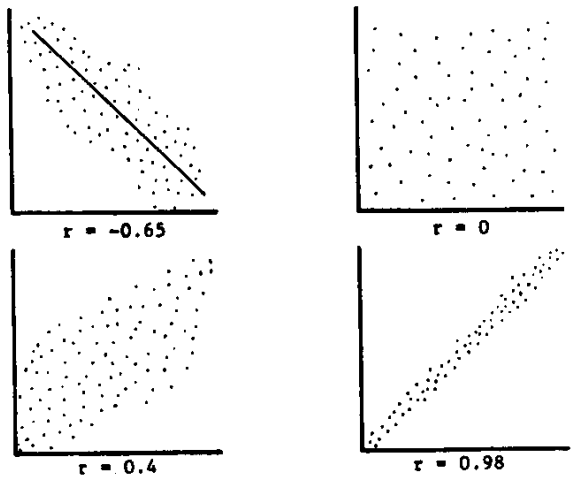


Figure 5.

Using the data from the example in the previous section we have:

$$r = \sqrt{bb'} \approx \sqrt{(0.8209789)(0.9693352)}$$
$$\approx \sqrt{0.7958037}$$
$$\approx 0.8921.$$

The value of r indicates a reasonably good correlation.

We can determine b and b' directly from the two normal equations. This allows us to calculate r without

the trouble of finding the lines of regression. With a little algebra we can write:

$$b = \frac{n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$

and

$$b' = \frac{n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} y_i\right)\left(\sum_{i=1}^{n} x_i\right)}{n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2}.$$

Exercise 1.

Given the normal equations (3), (4), (3'), and (4'), use algebra to obtain b and b' above.

Since $r = \sqrt{bb'}$ we can write:

$$r = \sqrt{\frac{\left(n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)\right)^2}{\left(n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2\right)\left(n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2\right)}}.$$

More simply:

$$r = \frac{n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{\sqrt{\left(n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2\right)\left(n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2\right)}}.$$

As a check we substitute the indicated sums in this new formula:

$$r = \frac{11(34382) - 552(576)}{\sqrt{(11(34372) - (552)^2)(11(35812) - (576)^2)}}$$

$$= \frac{378202 - 317952}{\sqrt{(73388)(62156)}} = \frac{60250}{67540}$$

$\therefore$ $r \doteq 0.8920639 \doteq 0.8921.$

This agrees with the results obtained by employing the explicit slopes, b and b' of the two lines of regression.

5. REGRESSION FOR LOGARITHMIC SCATTERS

Consider the graph in Figure 6 of a man's growth measured every three years after birth. Notice that there is a great deal of growth between birth and 15 years. After that time, growth tapers off. Table 5 gives the data used in plotting the graph.

TABLE 5

Age in Yrs.	Birth	3	6	9	12	15	18	21	24	27	...
Height in Ft.	1.5	3	3.75	4.5	5	5.8	6.1	6.15	6.17	6.18	...

Using the techniques developed in Section 3, we can easily fit a line to the data. See the calculations below in Table 6.

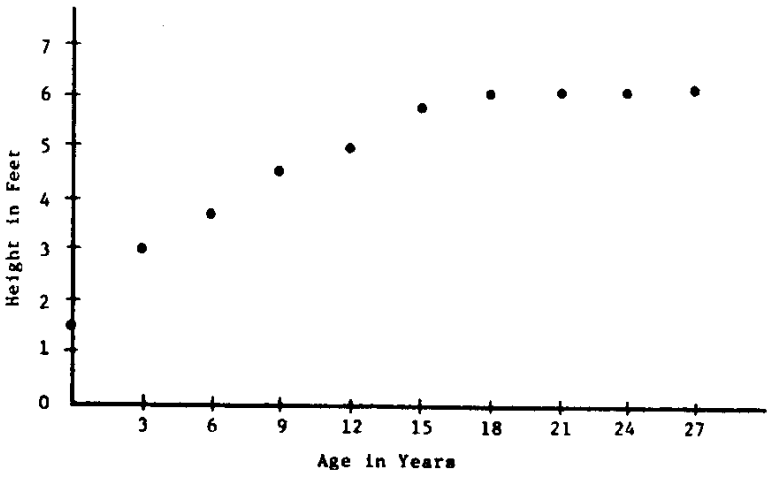


Figure 6.

TABLE 6

x	x^2	y	xy
0	0	1.60	0
3	9	2.90	8.70
6	36	3.75	22.50
9	81	4.50	40.50
12	144	5.00	60.00
15	225	5.80	87.00
18	324	6.10	109.80
21	441	6.15	129.15
24	576	6.17	148.08
27	729	6.18	166.86
135	2565	46.55	772.59

Using $b\Sigma x^2 + c\Sigma x = \Sigma xy$

$b\Sigma x + cn = \Sigma y.$

Therefore we can write:

$$2565b + 135c = 772.59$$

$$135b + 10c = 46.55$$

$$c = \frac{46.55 - 135b}{10}$$

and we can further write

$$2565b + 135\left(\frac{46.55 - 135b}{10}\right) = 772.59$$

$$25650b + 135(46.55) - (135)^2 b = 7725.9$$

$$25650b + 6284.25 - 18225b = 7725.9$$

$$7425b = 1441.65$$

$$b = 1441.65/7425$$

$$b = 0.1942$$

$$\therefore \quad c = 2.0333$$

and we have our line of regression:

$$y = 0.1942x + 2.0333.$$

In order to draw the line we need only locate two points.

For $x = 0$, $y = 2.0333$,

for $x = 3$, $y = 0.1942(3) + 2.0333 = 2.6159$.

While this is not a bad fit, we can do better. It turns out that the data will fit a logarithmic curve much better than a straight line. In general, logarithmic curves look like the one shown in Figure 8.

We can take the $\log_e$* of each of the x values (age in this example). We then use the same technique of least squares to find a log line of best fit. The calculations are given below. Notice how much closer this curve is to the actual data.

---

*$\log_e$ is also written ln. We are using $\log_e$ here to emphasize the general nature of logarithms. We can arbitrarily use any base.

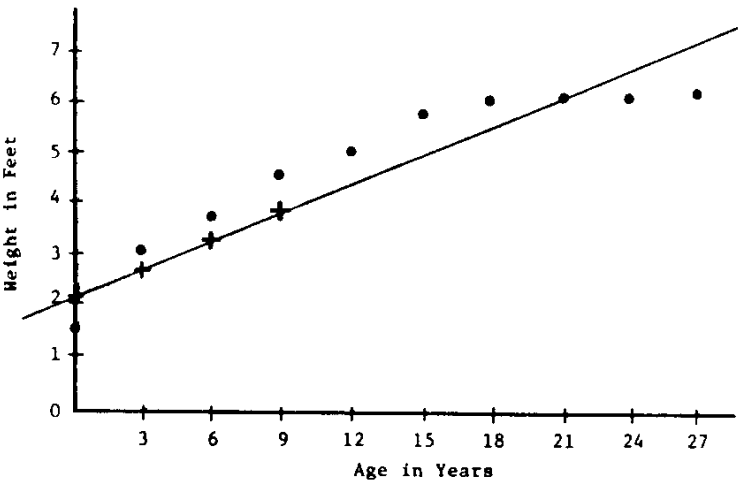


Figure 7.

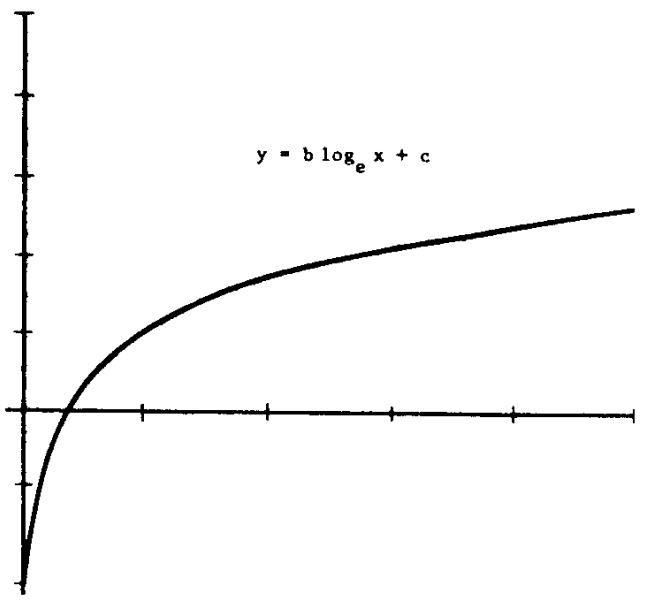


Figure 8.

TABLE 7

x	$\log_e x$	$(\log_e x)^2$	y	$(\log_e x)y$
0	$-\infty$			
3	1.0986	1.2069	2.90	3.1859
6	1.7918	3.2105	3.75	6.7193
9	2.1972	4.8277	4.50	9.8874
12	2.4849	6.1743	5.00	12.4245
15	2.7081	7.3338	5.80	15.5707
18	2.8904	8.3544	6.10	17.6314
21	3.0445	9.2690	6.15	18.7237
24	3.1781	10.1003	6.17	19.6089
27	3.2958	10.8623	6.18	20.3680
	22.6894	61.3392	46.55	124.1198

We use

$$b\Sigma(\log_e x)^2 + c\Sigma\log_e x = \Sigma(\log_e x)y$$

$$b\Sigma\log_e x + cn = \Sigma y.$$

Therefore we can write:

$$61.3392b + 22.6894c = 124.1198$$

$$22.6894b + 9c = 46.55$$

$$c = \frac{46.55 - 22.6894b}{9}$$

therefore

$$61.3392b + 22.6895\left(\frac{46.55-22.6894b}{9}\right) = 124.1198$$

$$9(61.3392b) + 22.6894(46.55) - (22.6984)^2 b = 9(124.1198)$$

$$552.0528b + 1056.1915 - 515.2174b = 1117.0782$$

$$36.8354b = 60.8867$$

$$b = 1.6529$$

$$c = \frac{46.55 - 22.6894(1.6529)}{9} = \frac{46.55 - 37.5033}{9}$$

$$= \frac{9.0467}{9} = 1.0052$$

$$y = 1.6529 \log_e x + 1.0052.$$

Let $x = 3$, then $y = 1.6529(1.0986) + 1.0052 = 2.821.$
$x = 6$, then $y = 1.6529(1.7918) + 1.0052 = 3.9669.$
$x = 9$, then $y = 1.6529(2.1972) + 1.0052 = 4.637.$
$x = 12$, then $y = 1.6529(2.4849) + 1.0052 = 5.1125.$
$x = 15$, then $y = 1.6529(2.7081) + 1.0052 = 5.4814.$
$x = 18$, then $y = 1.6529(2.8904) + 1.0052 = 5.7827.$
$x = 21$, then $y = 1.6529(3.0445) + 1.0052 = 6.0375.$
$x = 24$, then $y = 1.6529(3.1781) + 1.0052 = 6.2583.$
$x = 27$, then $y = 1.6529(3.2958) + 1.0052 = 6.4528.$

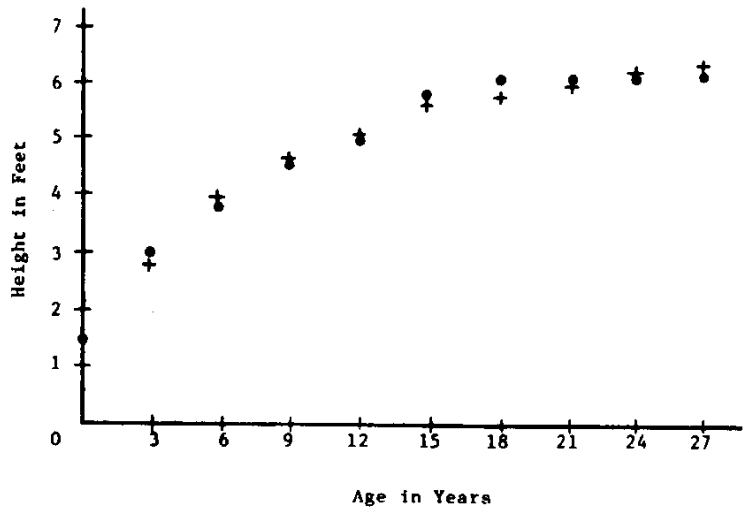


Figure 9.

Exercise 2.

Fit a logarithmic curve to the data given in the table below.

x	1	6	11	16	21	...
y	12	42	53	71	76	...

6. REGRESSION FOR EXPONENTIAL SCATTERS

Consider now, an experiment where a large number of corn seedlings were grown under favorable conditions. Every two weeks a few plants were weighed, and the average of their weights was recorded. (See Table 8.) We also give a graph in Figure 10. It would be difficult to find a straight line that would fit very well. The logarithmic curve does not fit so well either.

TABLE 8

Age in Weeks	2	4	6	8	10	12	14	16	18	20	...
Average Weight in Grams	21	28	58	76	170	422	706	853	924	966	...

This set of data is probably best fit to an exponential curve. The general shape of such curves ($y = e^x$) is given in Figure 11. Algebraically $y = e^x$ can be written $\log_e y = x$. For a general exponential we can write:

$$y = ce^{bx}.$$

With a little algebra, we can get a form that will allow us to use the least squares method. Analyze the development below.

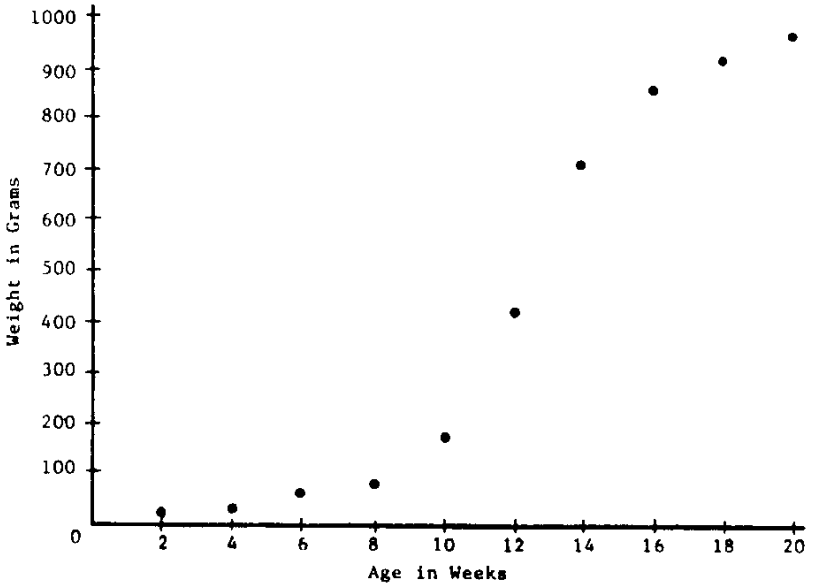


Figure 10.

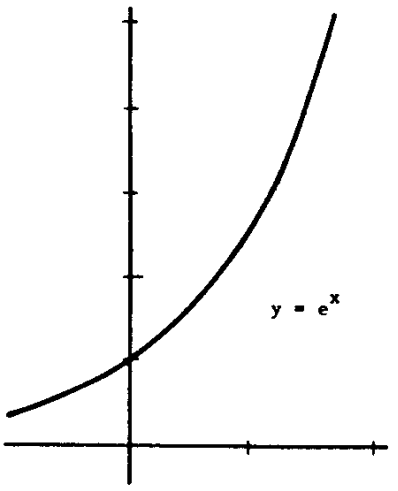


Figure 11.

$$y = ce^{bx} \rightarrow \frac{y}{c} = e^{bx} \rightarrow \log_e \frac{y}{c} = bx.$$

And further we have:

$$\log_e y - \log_e c = bx \quad \text{or} \quad \log_e y = bx + \log_e c.$$

We can find the "line" of best exponential fit by taking the $\log_e$ of the y values and then proceeding with the least squares technique. (See Table 9.)

TABLE 9

x	y	$\log_e y$	x^2	$x(\log_e y)$
2	21	3.0445	4	6.0890
4	28	3.3322	16	13.3288
6	58	4.0604	36	24.3624
8	76	4.3307	64	34.6456
10	170	5.1358	100	51.3580
12	422	6.0450	144	72.5400
14	706	6.5596	196	91.8344
16	853	6.7488	256	107.9808
18	924	6.8287	324	122.9166
20	966	6.8732	400	137.4640
110		52.9597	1540	662.5196

Keeping the equation

$$\log_e y = bx + \log_e c$$

in mind, we calculate

$$1540b + 110 \log_e c = 662.5196$$

$$110b + 10 \log_e c = 52.9597$$

$$b = \frac{52.9597 - 10 \log_e c}{110}$$

$$\log_e c = \frac{52.9597 - 110b}{10}$$

$$1540b + 110\left(\frac{52.9597 - 110b}{10}\right) = 662.5196$$

$$1540b + 5825.567 - 12100b = 6625.196$$

$$3300b = 799.629$$

$$b = 0.2423$$

$$\log_e c = \frac{52.9597 - 26.653}{10} = 2.63067$$

$$\log_e y = 0.2423x + 2.63067.$$

For $x = 2$, $\log_e y = 0.2423(2) + 2.63067$

$= 3.11527$

therefore

$$y = e^{3.11527} \doteq 2.71828^{3.11527}$$

$= 22.5395.$

For $x = 4$, $\log_e y = 0.2423(4) + 2.6307$

therefore

$y = 36.5946.$

For $x = 6$, $y = 59.4122$
for $x = 8$, $y = 96.4573$
for $x = 10$, $y = 156.6008$
for $x = 12$, $y = 254.2454$
for $x = 14$, $y = 412.7739$
for $x = 16$, $y = 670.1489$
for $x = 18$, $y = 1088.0038$
for $x = 20$, $y = 1766.4019.$

We could have solved for c when we obtained $\log_e c = 2.63067$.

If $\log_e c = 2.63067$, then $c \doteq 13.8831$,

therefore

$$y = 13.8831e^{0.2423x}.$$

If we substitute 2 for x we get a value which is virtually the same as we got using the other form. That is,

$$y = 13.8831e^{0.2423(2)} = 22.5395.$$

Observe the fitted curve in Figure 12.

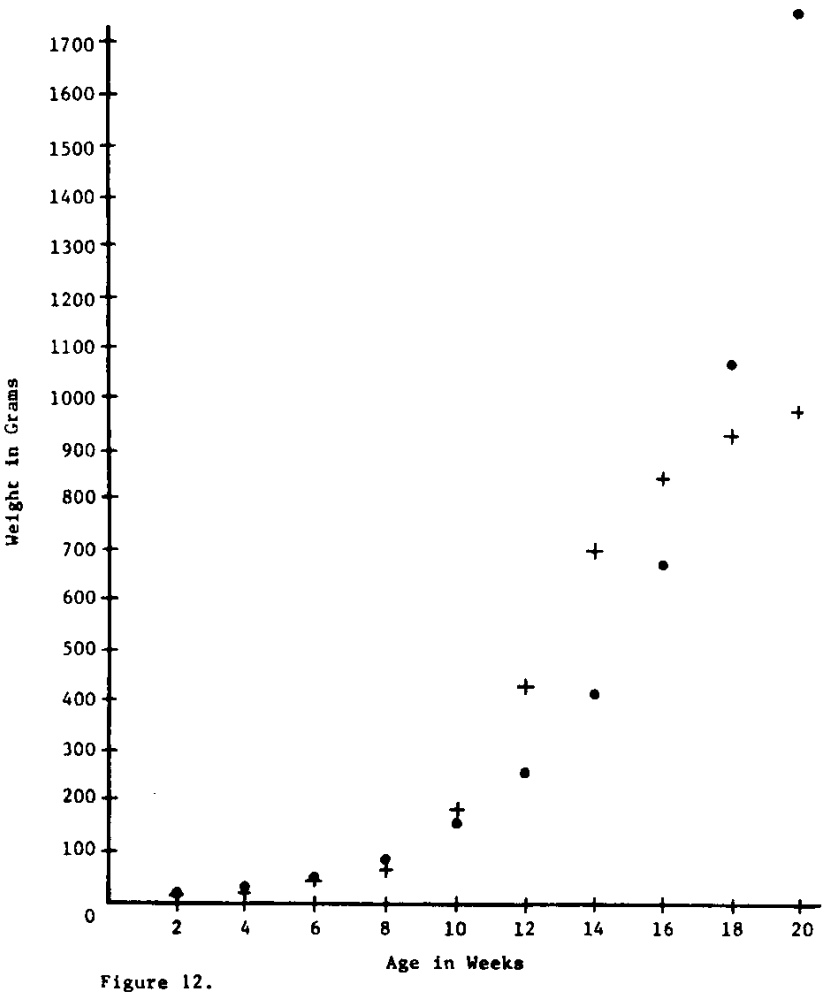


Figure 12.

Exercise 3.

Try to fit the data for the growth of the corn seedlings using 15 as a base instead of 10.

---

7. POLYNOMIAL SCATTERS

A disc was rolled down an inclined plane and the distance it travelled was measured after 0, 2, 4, ..., seconds. The results are organized in Table 10.

TABLE 10

Time (x)	0	2	4	6	8	10	12	14	16
Distance (y)	0	1	3	5	8	12	17	23	29

We give a graph of the data in Figure 13. Notice that it looks as if it could be fitted to an exponential. However, this data fits closer to a second degree polynomial or a parabola, $y = ax^2 + bx + c$.

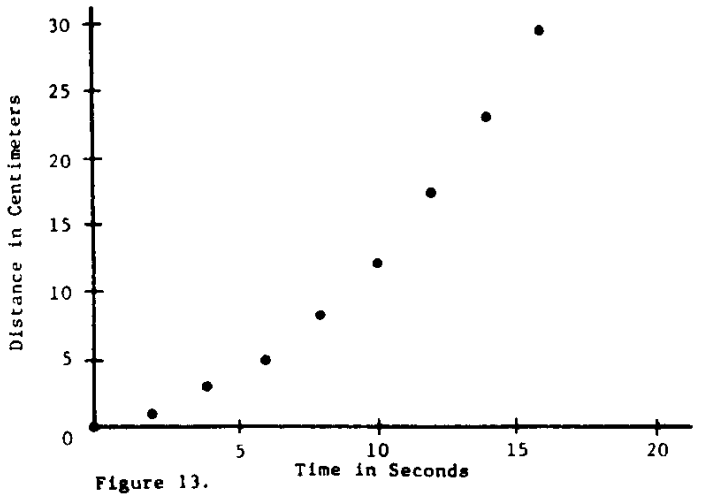


Figure 13.

In order to fit a polynomial we must do a little more mathematics. Notice, that we now have three constants to identify, namely a, b, and c.

We must consider minimizing the sum

$$\sum_{i=1}^{n} [y_i - (ax_i^2 + bx_i + c)]^2.$$

This means that we must calculate the partial derivatives of this sum with respect to a, b, and c. We set these derivatives equal to 0 and come up with three equations in three unknowns a, b, and c. That development is given below:

$$\frac{\partial}{\partial a} = \sum_{i=1}^{n} -2[y_i - ax_i^2 - bx_i - c]x_i^2.$$

$$= \sum_{i=1}^{n} (-2x_i^2 y_i + 2ax_i^4 + 2bx_i^3 + 2cx_i^2) = 0.$$

$$a\Sigma x_i^4 + b\Sigma x_i^3 + c\Sigma x_i^2 = \Sigma x_i^2 y_i.$$

$$\frac{\partial}{\partial b} = \sum_{i=1}^{n} -2[y_i - ax_i^2 - bx_i - c]x_i = 0.$$

$$a\Sigma x_i^3 + b\Sigma x_i^2 + c\Sigma x_i = \Sigma y_i x_i.$$

$$\frac{\partial}{\partial c} = \sum_{i=1}^{n} -2[y_i - ax_i^2 - bx_i - c](1) = 0.$$

$$a\Sigma x_i^2 + b\Sigma x_i + c\Sigma(1) = \Sigma y_i.$$

$$a\Sigma x_i^2 + b\Sigma x_i + nc = \Sigma y_i.$$

From these *normal* equations we can obtain a best parabolic fit. We must find the indicated sums Σx_i^4, Σx_i^3, Σx_i^2, Σx_i, $\Sigma x_i^2 y$, $\Sigma x_i y_i$, and Σy_i. We give these calculations below using the data from Table 11.

$$a(140352) + b(10368) + c(816) = 16324.$$

$$a(10368) + b(816) + c(72) = 1218.$$

$$a(816) + b(72) + c(9) = 98.$$

TABLE II

x_i	x_i^2	x_i^3	x_i^4	y_i	x_iy_i	x_i^2y
0	0	0	0	0	0	0
2	4	8	16	1	2	4
4	16	64	256	3	12	48
6	36	216	1296	5	30	180
8	64	512	4096	8	64	512
10	100	1000	10000	12	120	1200
12	144	1728	20736	17	204	2448
14	196	2744	38416	23	322	4508
16	256	4096	65536	29	464	7424
72	816	10368	140352	98	1218	16324

To solve this system, we can use an augmented matrix.*

$$\begin{bmatrix} 140352 & 10368 & 816 & 16324 \\ 10368 & 816 & 72 & 1218 \\ 816 & 72 & 9 & 98 \end{bmatrix}.$$

We first divide the top row through by 140352 to obtain 1 in the first row and first column:

$$\begin{bmatrix} 1 & 0.0739 & 0.0058 & 0.1163 \\ 10368 & 816 & 72 & 1218 \\ 816 & 72 & 9 & 98 \end{bmatrix}.$$

Next, we multiply the top row by -10368 and add it to the second row. Then, we multiply the top row by -816 and add it to the bottom row. The resulting matrix is given below:

*For a more detailed discussion on matrix manipulations see *Elementary Differential Equations with Linear Algebra* by Ross. L. Finney and Donald R. Ostberg.

$$\begin{bmatrix} 1 & 0.0739 & 0.0058 & 0.1163 \\ 0 & 423.0528 & 11.8656 & 12.2016 \\ 0 & 41.0736 & 4.2672 & 3.0992 \end{bmatrix}.$$

To continue we divide the second row by 423.0528 to obtain 1 in the second row, second column.

$$\begin{bmatrix} 1 & 0.0739 & 0.0058 & 0.1163 \\ 0 & 1 & 0.0280 & 0.0288 \\ 0 & 41.0736 & 4.2672 & 3.0992 \end{bmatrix}.$$

We now multiply the second row by -41.0736 and add it to the third row.

$$\begin{bmatrix} 1 & 0.0739 & 0.0058 & 0.1163 \\ 0 & 1 & 0.0280 & 0.0288 \\ 0 & 0 & 3.1172 & 1.9163 \end{bmatrix}.$$

We divide the last row by 3.1172 and obtain c from the system above:

$$\begin{bmatrix} 1 & 0.0739 & 0.0058 & 0.1163 \\ 0 & 1 & 0.0280 & 0.0288 \\ 0 & 0 & 1 & 0.6148 \end{bmatrix}.$$

Therefore,

$$c = 0.6148.$$

$$b + 0.028(0.6148) = 0.0288$$

$$b + 0.0172 = 0.0288$$

$$b = 0.0116.$$

$$a + 0.0739(0.0116) + 0.0058(0.6148) = 0.1163$$

$$a + 0.0008572 + 0.003566 = 0.1163$$

$$a = 0.1119.$$

Therefore the parabola of best fit is:

$$y = 0.1119x^2 + 0.0116x + 0.6148.$$

We obtain the y values below:

$x = 0$, $y = 0.6148$.

$x = 2$, $y = 0.1119(4) + 0.0116(2) + 0.6148$

$= 0.4475 + 0.0232 + 0.6148$

$= 1.0855$.

$x = 4$, $y = 1.6683$.

$x = 6$, $y = 4.9684$.

$x = 8$, $y = 7.8692$.

$x = 10$, $y = 11.9208$.

$x = 12$, $y = 16.8676$.

$x = 14$, $y = 22.1104$.

$x = 16$, $y = 29.4468$.

When this data is graphed on the original set of axes, we see that we have a very close fit. (See Figure 14.)

Exercise 4.

Fit the data to an exponential. It should be convincing that the exponential does not fit as well as the parabola.

There are sets of data that produce scatters that fit higher order polynomials than 2. For example, the

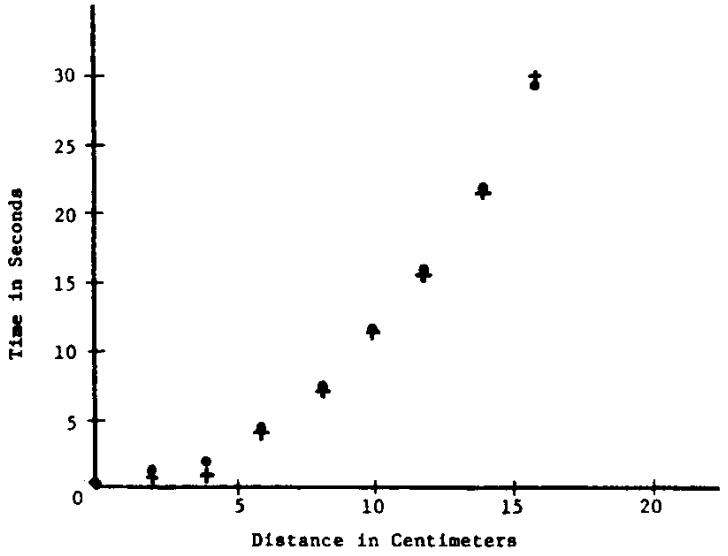


Figure 14.

corn seedling example in Section 6 might be fit with a cubic (i.e., $y = ax^3 + bx^2 + cx + d$).

However, this means that we would have to solve four equations in four unknowns (a, b, c, d). This is no small task. There are methods for finding the coefficients without going through all the work of the partial derivations, namely, the square root method and Gauss's method.

There is still a great deal of calculation to do even with these methods. In fact, all curve fitting requires a good deal of calculation. Now that we have computers, we can write programs to deal with any type of scatter.

We present, as an appendix, a BASIC program called "Super Fit." After going through this unit the reader should be comfortable with using the program. The

program uses the same procedures, but you are spared the calculations.

It should also be pointed out that in practice, the amount of data collected would more than likely be more extensive. We have also kept the numbers reasonably small.

With a computer program to do the work, we can enter a large number of data and the numbers can be either very large, or very small.

There are other functions such as powers and powers raised to powers that can be employed, and data fitted to them (i.e., $y = cx^n$, $y = cx^{nb}$, etc.). Appropriate manipulation of the data can be employed to handle these situations. The basic mathematics of the least square method can still be used. Hopefully, this material has given enough background so that virtually any type of scatter can be fitted.

8. MODEL EXAM

1. Given the data in the table below construct a scatter diagram:

x	2	8	7	10	15	18	15	18	23	23	25	25	30	32	35
y	2	2	8	15	10	5	20	16	10	15	20	25	23	27	25

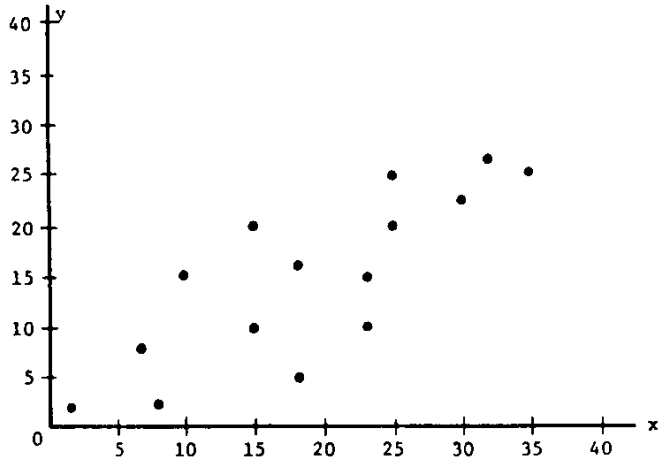
2. For the data given in Question 1, is the coefficient of correlation positive, negative, or zero? Fit a line by eye through the points of the scatter diagram that was constructed for Question 1. Fit a line through the data using the least square technique.

3. Given the parabola $y = 2x^2 + 3$, let x take on the values 1, 2, 3, 4, 5, 6, and 7. Find the corresponding y values. Which type of function—logarithmic or exponential—will best fit the given parabola?

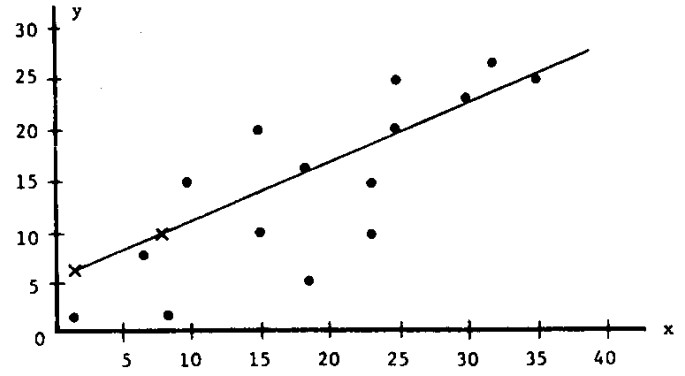
4. Fit the data in Question 3 to either a logarithmic or exponential curve depending on your choice from Question 3.

9. ANSWERS TO MODEL EXAM

1.



2. Positive.



x	y	x^2	xy
2	2	4	4
8	2	16	16
7	8	49	56
10	15	100	150
15	10	225	150
18	5	324	90
15	20	225	300
18	16	324	288
23	10	529	230
23	15	529	345
25	20	625	500
25	25	625	625
30	23	900	690
32	27	1024	864
35	25	1225	875
253	223	6724	5183

$$6724b + 253c = 5183$$
$$253b + 15c = 223$$

$$c = \frac{223 - 253b}{15}$$

$$6724b + 253\left(\frac{223 - 253b}{15}\right) = 5183$$
$$6724(15)b + 253(223) - 253^2b = 5183(15)$$
$$100860b + 56419 - 64009b = 77745$$
$$36851b = 21316$$
$$b = 0.5787.$$

Since $c = \frac{223 - 253b}{15}$ we have $c = \frac{223 - 253(0.5787)}{15}$

therefore

$$c = 5.1059.$$

The line of regression is

$$y = 0.5787x + 5.1059.$$

3. $y = \{5,11,21,35,53,75,101\}$

An exponential would fit best.

4.

x	y	$\log_e y$	x^2	$x \log_e y$
1	5	1.6094	1	1.6094
2	11	2.3979	4	4.7958
3	21	3.0445	9	9.1335
4	35	3.5553	16	14.2212
5	53	3.9703	25	19.8515
6	75	4.3175	36	25.9050
7	101	4.6151	49	32.3057
28		23.5100	140	107.8221

$$140b + 28\log_e c = 107.8221$$

$$28b + 7\log_e c = 23.5100$$

$$\log_e c = \frac{23.51 - 28b}{7}$$

$$140b + 28\left(\frac{23.51 - 28b}{7}\right) = 107.8221$$

$$7(140)b + 28(23.51) - 28^2 b = 7(107.8221)$$

$$980b + 658.28 - 784b = 754.7547$$

$$196b = 96.4747$$

$$b = 0.4922.$$

$$\log_e c = \frac{23.51 - 28(0.4922)}{7} = \frac{9.7284}{7} = 1.3898$$

therefore

$$c \doteq 4.0139.$$

We have $y = ce^{bx}$ which yields $y = 4.0139e^{0.4922x}$
so if

$x = 1$
$y = 4.0139e^{0.4922(1)} = 4.0139(1.6359) = 6.5664$

$x = 2$
$y = 4.0139e^{0.4922(2)} = 4.0139(2.6762) = 10.742$

$x = 3$
$y = 4.0139e^{0.4922(3)} = 4.0139(4.378) = 17.573$

$x = 4$
$y = 4.0139e^{0.4922(4)} = 4.0139(7.162) = 28.7479$

$x = 5$

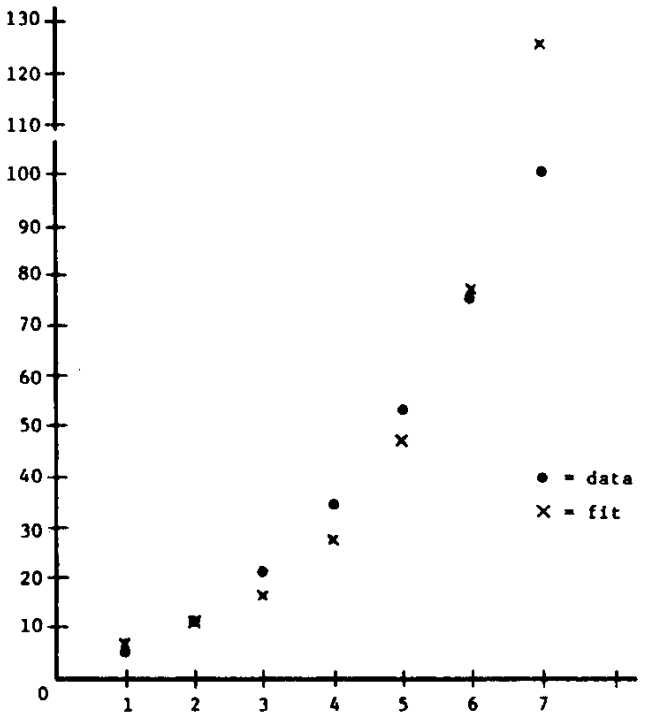
$y = 4.0139e^{0.4922(5)} = 4.0139(11.7165) = 47.0289$

$x = 6$

$y = 4.0139e^{0.4922(6)} = 4.0139(19.1672) = 76.9352$

$x = 7$

$y = 4.0139e^{0.4922(7)} = 4.0139(31.3558) = 125.8491.$



10. ANSWERS TO EXERCISES

1. Given

$$(3) \quad b\sum_{i=1}^{n} x_i^2 + c\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} x_i y_i$$

$$(4) \quad b\sum_{i=1}^{n} x_i + nc = \sum_{i=1}^{n} y_i,$$

using (4) we can write:

$$nc = \sum_{i=1}^{n} y_i - b\sum_{i=1}^{n} x_i$$

$$\therefore \quad c = \frac{\sum_{i=1}^{n} y_i - b\sum_{i=1}^{n} x_i}{n} .$$

Substitute this value of c into (3):

$$b\sum_{i=1}^{n} x_i^2 + \left(\frac{\sum_{i=1}^{n} y_i - b\sum_{i=1}^{n} x_i}{n}\right)\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} x_i y_i .$$

Multiply the equation obtained by n and remove the parentheses using the distributive law:

$$nb\sum_{i=1}^{n} x_i^2 + \sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i - b\sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i = n\sum_{i=1}^{n} x_i y_i .$$

Further we can write:

$$b\left(n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2\right) = n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i$$

$$b = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i}{n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)^2} .$$

To solve for b' we do a similar procedure:

Given

$$(3') \quad b'\sum_{i=1}^{n} y_i^2 + c\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} y_i x_i$$

$$(4') \quad b'\sum_{i=1}^{n} y_i + nc = \sum_{i=1}^{n} x_i,$$

using (4') we can write:

$$nc = \sum_{i=1}^{n} x_i - b'\sum_{i=1}^{n} y_i$$

$$\therefore \quad c = \frac{\sum_{i=1}^{n} x_i - b'\sum_{i=1}^{n} y_i}{n} .$$

Substitute the value of c into (3'):

$$b'\sum_{i=1}^{n} y_i^2 + \left(\frac{\sum_{i=1}^{n} x_i - b'\sum_{i=1}^{n} y_i}{n}\right)\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} y_i x_i$$

$$nb'\sum_{i=1}^{n} y_i^2 + \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i - b'\sum_{i=1}^{n} y_i \sum_{i=1}^{n} y_i = n\sum_{i=1}^{n} y_i x_i$$

$$b'\left(n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2\right) = n\sum_{i=1}^{n} y_i x_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i$$

$$b' = \frac{n\sum_{i=1}^{n} y_i x_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2} .$$

2.

x	$\log_e x$	$(\log_e x)^2$	y	$(\log_e x)y$
1	0	0	12	0
6	1.7918	3.2104	42	75.2556
11	2.3979	5.7499	53	127.0887
16	2.7726	7.6872	71	196.8546
21	3.0445	9.2691	76	231.3820
55	10.0068	25.9166	254	630.5809

Use

$$b\sum(\log_e x)^2 + c\sum\log_e x = \sum(\log_e x)y$$

$$b\sum\log_e x \quad + \quad cn = \sum y.$$

Therefore we can write:

$$25.9166b + 10.0068c = 630.5809$$

$$10.0068b + 5c = 254$$

$$c = \frac{254 - 10.0068b}{5}$$

$$\therefore \quad 25.9166b + 10.0068\left(\frac{254 - 10.0068b}{5}\right) = 630.5809$$

$$5(25.9166b) + 10.0068(254 - 10.0068b) = 5(630.5809)$$

$$129.583b + 2541.7272 + 100.136 = 3152.9045$$

$$129.583b = 511.0413$$

$$b = 3.9437.$$

Hence,

$$c = \frac{254 - 10.0068(3.9437)}{5}$$

$$= 42.9072$$

and

$$y = 3.9437\log_e x + 42.9072.$$

Let

$x = 1$, then $y = 3.9437(0) + 42.9072$

$= 42.9072$

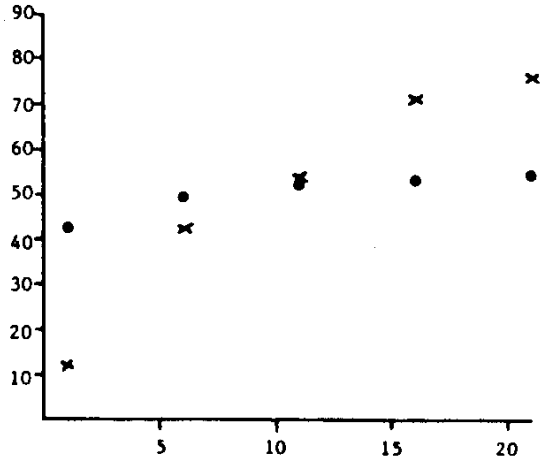
$x = 6$, $y = 3.9437(1.7918) + 42.9072$

$= 49.9735$

$x = 11$, $y = 52.3638$

$x = 16$, $y = 53.8415$

$x = 21$, $y = 54.9138.$



Graph for solution to Exercise 2.

3. We could use any base and obtain the same fit. The calculations are given for a base 10 (common logs) fit. Notice that the y values are virtually the same as those obtained for $\log_e$.

x	y	$\log_{10} y$	x^2	$x(\log_{10} y)$
2	21	1.3222	4	2.6444
4	28	1.4472	16	5.7888
6	58	1.7634	36	10.0580
8	76	1.8808	64	15.0464
10	170	2.2304	100	22.3040
12	422	2.6253	144	31.5036
14	706	2.8488	196	39.8832
16	853	2.9309	256	46.8944
18	924	2.9657	324	53.3826
20	966	2.9850	400	59.7000
110		22.9997	1540	287.2054

$$1540b + 110 \log c = 287.2054$$

$$110b + 10 \log c = 22.9997$$

$$b = \frac{22.9997 - 10 \log c}{110} = \frac{22.9997 - 11.5968}{110} = 0.10366$$

therefore

$$1540\left(\frac{22.9997 - 10 \log c}{110}\right) + 110 \log c = 287.2054$$

$$35419.538 - 15400 \log c + 12100 \log c = 31592.594$$

$$3300 \log c = 3826.944$$

$$\log c = 1.15968$$

$$c \doteq 14.444.$$

$x = 20$

$y = 14.444(10)^{0.10366(20)} = 14.444(10)^{2.0732}$

$= 14.444(118.3586) = 1709.5716.$

$x = 18$

$y = 1060.6879$

$x = 16$

$y = 658.0932$

$x = 14$

$y = 408.2134$

$x = 12$

$y = 253.2727$

$x = 10$

$y = 157.1406$

$x = 8$

$y = 97.4956$

$x = 6$

$y = 60.49$

$x = 4$

$y = 37.5226$

$x = 2$

$y = 23.2808$

4.

x	y	$\log_e y$	x^2	$x(\log_e y)$
0	0	$-\infty$		
2	1	0	4	0
4	3	1.0986	16	4.3944
6	5	1.6094	36	9.6564
8	8	2.0794	64	16.6352
10	12	2.4849	100	24.8490
12	17	2.8332	144	33.9984
14	23	3.1354	196	43.8956
16	29	3.3673	256	53.8768
72	98	16.6082	816	187.3058

Keeping the equation

$$\log_e y = bx + \log_e c$$

$$816b + 72 \log_e c = 187.3058$$

$$72b + 8 \log_e c = 16.6082$$

$$\log_e c = \frac{16.6082 - 72b}{8}$$

$$816b + 72\left(\frac{16.6082 - 72b}{8}\right) = 187.3058$$

$$6528b + 1195.7904 - 5184b = 1498.4464$$

$$1344b = 302.656$$

$$b = .2252$$

$$\log_e c = \frac{16.6082 - 72(.2252)}{8}$$

$$= .0492$$

$\therefore$ the fitted curve is:

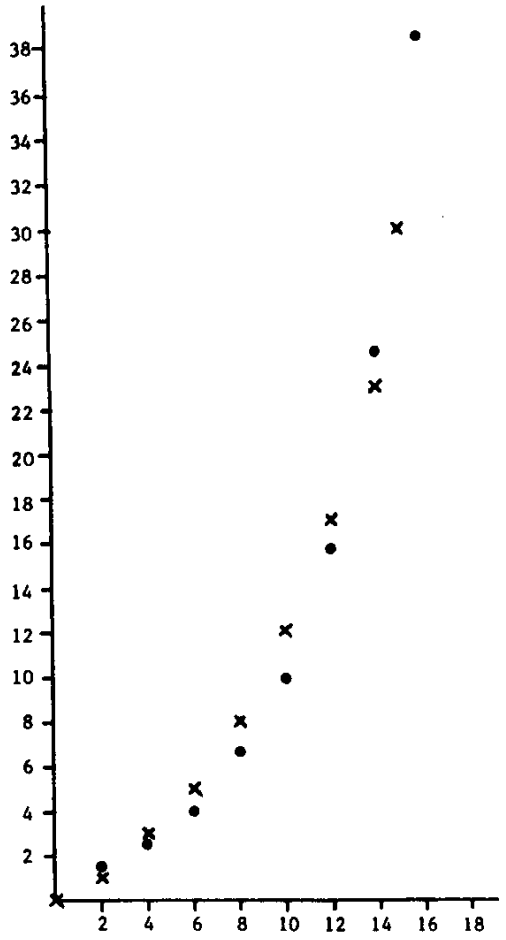
$$\log_e y = .2252x + .0492.$$

for $x = 2$, $\log_e y = .2252(2) + .0492$

$$= .4996$$

$$y = e^{.4996} \doteq 1.6481$$

x = 4,	y = 2.5857	x = 12,	y = 15.6677
x = 6,	y = 4.0568	x = 14,	y = 24.5818
x = 8,	y = 6.6859	x = 16,	y = 38.5673
x = 10,	y = 9.9742		



Graph for solution to Exercise 4.