UMAP

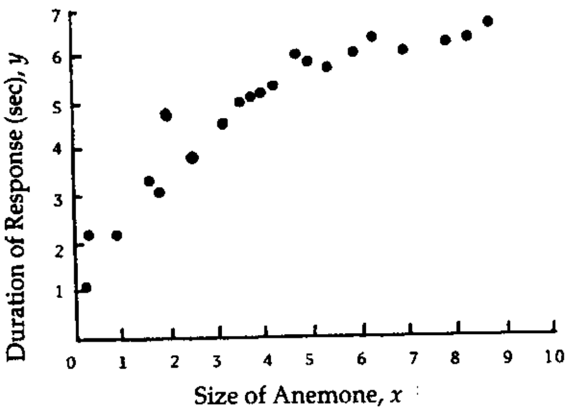
Modules in
Undergraduate
Mathematics
and its
Applications

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 433

Fitting Equations to Data: Part I

Thomas M. Lamm



Applications of Statistics

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

FITTING EQUATIONS TO DATA: PART I

by

Thomas M. Lamm
Department of Mathematical Sciences
Eastern Kentucky University
Richmond, KY 40475

TABLE OF CONTENTS

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

The Project would like to thank Roger Carlson of the University of Missouri and member of the UMAP Statistics Panel, Carol O'Connor of University of Louisville, and one anonymous reviewer 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 and No. SED80-07731. Recommendations expressed are those of the author and do not necessarily reflect the views of the NSF or of COMAP.

Intermodular Description Sheet: UMAP Unit 433

Title: Fitting Equations to Data: Part I

Author: Thomas M. Lamm
Department of Mathematical Sciences
Eastern Kentucky University
Richmond, KY 40475

Review Stage/Date: III 4/15/81

Classification: STATISTICS

Prerequisite Skills:

1. (Functions) Be familiar with the concepts of constant, variable, and function, including independent and dependent variable and the notation $y = f(x)$. Previous exposure to quadratic and logarithmic functions is desirable but not necessary.
2. (Graphing) Given a table of x and y values, choose an appropriate scale and axis limits, and plot the points on standard (square-grid) graph paper.
3. (Lines) Know that the slope-intercept equation for a line is $y = mx + b$ where m = slope and b = y-intercept. Know that the equation of a line passing through the point (x_0, y_0) is $y - y_0 = m(x-x_0)$, where m = slope.
4. Know that:

a) $$\sum_{i=1}^{N} x_1 = x_1 + x_2 + \dots + x_N$$

b) $$\sum_{i=1}^{N} (x_i + y_i) = \sum_{i=1}^{N} x_i + \sum_{i=1}^{N} y_i$$

c) $$\sum_{i=1}^{N} cx_i = c \sum_{i=1}^{N} x_i$$

d) $$\sum_{i=1}^{N} c = Nc.$$

5. (Use of calculator) Given a table of x and y values, compute sums such as Σx_i, $\Sigma x_i y_i$, and Σx_i^2 using a hand-held calculator.

Output Skills:

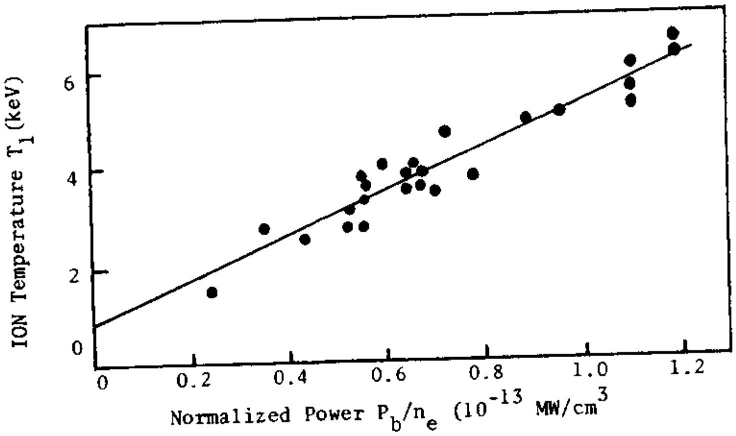
1. State four purposes for fitting an equation to data, and give an example of each.
2. Know that experimental error is the difference between the observed value of a variable and its hypothetical true value. Know that it is present in all real experiments.
3. Given a table of x and y values, make a scatter diagram.
4. Given a scatter diagram (showing small variance in y for given x), sketch a free-hand curve through the data points.
5. Given a list of y values, compute the mean and variance. Know that the variance is a measure of the scatter of the y values about the mean.
6. Know that the least-squares (LS) fit of a given type of equation $y = f(x)$ to data points (x_i,y_i), $i = 1,2,3,\dots,N$, is defined to be the particular equation of that type that minimizes the sum of the squares of the deviations

$$S = \sum_{i=1}^{N} (y_i - f(x_i))^2.$$

1. THE BASIC PROBLEM

1.1 Introduction

Have you ever read an article in Scientific American, Physics Today, Science or other periodical, and come across a picture like this one:



(Source: Murakami, M., and Eubank, H.P., "Recent progress in tokamak experiments," Physics Today 32 (1979), #5, p.30.)

Some experiment has been performed to determine how one quantity, ion temperature, depends on another quantity, normalized power. The results of the experiment are indicated by the dots in the picture. Then someone has drawn in the curve which "best fits" these data points. Perhaps the equation for the curve has been determined.

Did you wonder how they knew where to draw the curve? How can the equation for the curve be determined? Sometimes it is not obvious what type of curve to draw. Should it be a straight line, a parabola, a log curve, or some other type? Among all curves of the same type, which one gives the best fit? What is meant by "best fit" anyway?

In this unit and its sequel, Unit 434, we shall consider the problem of fitting a curve to data. We shall take for our measure of goodness of fit the most common one--the least-squares criterion. In this unit, you will learn how to fit a line to data and how to tell if the fit is a good one. In the sequel, you will learn how to fit other common curves, using the computer.

The four examples that follow illustrate the purposes and process of curve fitting. You need not understand all the technical language.

1.2 Four Examples

Example 1. The chemical engineers of a pilot plant need a way to measure the organic acid content of a certain chemical. Two methods are available: a relatively inexpensive titration method, and a more expensive method of extraction and weighing.

An experiment is carried out to determine how well the acid number (x) obtained by titration can serve to estimate the organic acid content (y) determined by extraction and weighing. Twenty samples of the chemical are selected to cover the range in organic acid content of interest. Each sample is divided and analyzed by both methods. The results are shown in Table 1.

TABLE 1

ORGANIC ACID CONTENT

Sample number	x Acid no. by titration (mg)	y Acid content by extraction (ml)
1	123	76
2	109	70
3	62	55
4	104	71
5	57	55
6	37	48
7	44	50
8	100	66
9	16	41
10	28	43
11	138	82
12	105	68
13	159	88
14	75	58
15	88	64
16	164	88
17	169	89
18	167	88
19	149	84
20	167	88

A computer analysis of the data is performed, which determines that the linear equation that best fits the data is $y = 0.3216x + 35.46$. Ninety-nine and one half percent of the total variation in y is accounted for by this equation.

On the basis of this analysis, the engineers decide that titration does give a reliable estimate of the organic acid content; they will use this less expensive method.

Example 2. An introductory physics class performed an experiment to determine the latent heat of fusion of water. (This is the number of calories of heat absorbed by one gram of ice as it melts.) Each of the 30 students started with 55cc of water in an insulated container. Each student added a different amount of ice (x) to the water and noted the fall in temperature (y) of the water as the ice melted. Their results are shown in Table 2.

TABLE 2

HEAT OF FUSION OF WATER

Weight of Ice (x), grams	13.4	12.8	15.1	19.7	17.5
Fall in Temperature (y), °C	19	19	22	30	27

Weight of Ice (x), grams	19.3	12.5	11.5	14.2	11.2	10.5	11.8	10.2	13.6	18.1
Fall in Temperature (y), °C	29	18	18	21	17	16	17	16	20	29

Weight of Ice (x), grams	17.8	18.1	14.2	15.9	18.3	19.3	13.9	15.6	10.7	18.5
Fall in Temperature (y), °C	27	28	23	25	27	29	21	24	16	28

Weight of Ice (x), grams	17.2	15.0	14.2	19.0	17.2
Fall in Temperature (y), °C	25	22	20	28	26

According to theory, the fall in temperature is related to the amount of ice by the equation $y = \beta x/55$, where β is the latent heat of fusion of water. The class estimated β by finding the value of the slope m that made the line $y = mx$ fit the data best.

Example 3. A biologist wants to formulate a theory about the nervous system of the sea anemone. In particular, she is interested in how the response of an anemone to a standard stimulus is related to its size. She has gathered the following data on 21 specimens.

TABLE 3

RESPONSE OF SEA AENMONE TO STANDARD STIMULUS

Size of Anemone (x), conventional units	0.2	3.9	5.8	7.8
Duration of Response (y), seconds	1.1	5.2	6.0	6.2

Size of Anemone (x), conventional units	4.9	8.6	3.1	6.2	0.3	4.2	0.9	3.5	1.8	3.7	6.9	4.7
Duration of Response (y), seconds	5.9	6.6	4.5	6.3	2.2	5.3	2.2	5.0	3.1	5.1	6.0	6.0

Size of Anemone (x), conventional units	2.5	1.6	8.2	5.3	1.9
Duration of Response (y), seconds	3.8	3.3	6.3	5.7	4.8

Studying these data, she concludes that the duration of the response (y) is related to the size of the anemone (x) by an equation of the form $y = m \log x + b$. Her theory will try to explain this in terms of the physiology of the anemone's nervous sytem.

Example 4. In 1886, Sir Francis Galton published his now classic experiments on the heredity of exceptional traits. For a wide range of species, from yeast to man, he showed that, with respect to specific measurable traits, offspring tend to be closer to the population mean (average) than their parents. As he put it, the offspring "regressed" toward the mean.

For instance, with humans Galton noticed that children of exceptionally tall parents tended to be tall, but not so tall as their parents. Similarly, children of very short parents tended to be short, but not so short as their parents.

To support his theory, Galton measured 938 fully grown children and their parents, and related each child's height (y) to the mean height (x) of his parents. A sample of his data is presented in Table 4.

Galton plotted the data and found the line that best represented the relation of y to x. The line had slope less than one. If no regression toward the mean had been observed, the best line would have been $y = x$. So his theory was supported by his data.

TABLE 4

HEIGHTS IN INCHES OF 46 CHILDREN (y) AND THEIR PARENTS (x)

x	y	x	y	x	y	x	y
64.5	63.2	67.5	69.2	68.5	70.2	71.5	70.2
64.5	64.2	67.5	69.2	68.5	70.2		
65.5	72.2	67.5	70.2	69.5	67.2		
66.5	65.2	67.5	71.2	69.5	67.2		
66.5	66.2	67.5	72.2	69.5	68.2		
66.5	67.2	68.5	64.2	69.5	69.2		
66.5	70.2	68.5	65.2	69.5	69.2		
67.5	65.2	68.5	66.2	69.5	70.2		
67.5	66.2	68.5	67.2	69.5	70.2		
67.5	66.2	68.5	67.2	69.5	71.2		
67.5	66.2	68.5	67.2	69.5	72.2		
67.5	67.2	68.5	68.2	69.5	73.2		
67.5	67.2	68.5	68.2	70.5	69.2		
67.5	69.2	68.5	69.2	71.5	67.2		
67.5	69.2	68.5	69.2	71.5	70.2		

1.3 The Purposes of Curve Fitting

As examples show, fitting equations to data is used in many different fields for various purposes. Some of the purposes are:

a) To summarize masses of data to obtain formulas or curves for calibration, interpolation, or prediction (Example 1).

b) To determine a meaningful constant, or to compare sets of data by means of the constants in the equation (Example 2).

c) To suggest the type of theoretical model to construct (Example 3).

d) To confirm or refute an hypothesis or model (Example 4).

In each of these examples, some experiment was performed to determine how a quantity y depends on a quantity x. The object was to find a function $y = f(x)$ that fit the data well. First, the experimenter decided on the type of function to use. In Examples 1, 2, and 4, this was a line $y = mx + b$, and in Example 3, it was a log curve $y = m \log x + b$. The type of function chosen might be indicated by theory (Examples 1 and 2), by the data itself (Example 3), by the purpose of the experiment (Example 4), or by past experience, convenience, or other means. Then some method was used to calculate the particular function of that type that fit the data best. In Example 1, an indication was also given as to how well the chosen function fit the data. (It was said to account for 99.5% of the total variation in y.)

The type of curve chosen is often a line. There are several reasons for this. Many relations, like the one in Example 2, are really linear. Other non-linear relations can be made linear by a transformation of the variables, as you will see if you read Unit 434. Sometimes one is interested only in a limited range of the variable x, and over this range the relation may be approximately linear. (The technique of interpolation, which you may have used to compute logarithms, makes use of this.) Also a line may be used simply for convenience in summarizing the results (as in Example 4). The line is the easiest type of curve to fit, but the method we shall use will generalize to non-linear cases.

Finding the function that best fits the data is different from the problems usually encountered in an algebra course. There, one starts with a given function and finds its values at certain points, or its zeros, or where it is equal to another function. Here, we are in effect given the (approximate) values of the function at certain points, and required to find the function.

In fact, we do not even know the exact values of the function. Example 2 illustrates this. Students with the same measured amount of ice observed different changes in thermometer readings. One function cannot, of course, have two y-values for the same x-value. We assume that there is one "correct" or "true" value of y for each x, but due to the limitations of the experimental technique, the observed values of y differ from this by an unknown (small) amount.

1.4 Experimental Error

The difference between the observed values of y and the unknown "true" value is called experimental error (error in the sense of uncertainty, not in the sense of "mistake"). The sources of experimental error are numerous and sometimes hard to determine, let alone eliminate. Some general sources are:

a) Measurement error. Any instrument of measurement has a limited degree of precision. In Example 2, for instance, the thermometer read only to the nearest degree, introducing a possible error of $\pm 0.5^{\circ}$ in each value of y.

b) Uncontrolled variables. The background conditions may vary from one measurement to another. For example, some containers in Example 2 may have been better insulated than others. The value of y may depend on other variables besides x, and these may not remain constant throughout the experiment.

c) Sampling error. Results can be misleading because of the small number of data on which they are based. For example, the measurements in Table 4 may not represent the population as a whole. Sampling error can be controlled, but not usually eliminated, by taking large and wisely chosen samples.

There are also human blunders, such as misreading a thermometer or miscopying a number, and systematic errors--errors that consistently bias the results in one direction. In Example 3, a faulty thermometer that consistently gives low readings would lead to systematic errors. Or, in Example 1, a systematic error would result if the titrations had been done incorrectly, giving values of x that were all too large or too small. The best one can hope for is that, by good experimental design and careful procedure, all systematic errors and human blunders are eliminated and that the remaining errors are small.

1.5 True Values

In some problems, the true value of y is meaningful only as an average. In Example 4, the height of the child depends on so many uncontrolled factors (hereditary and environmental) that it cannot really be considered a function of the parent's height alone. Nevertheless, among all children whose parents have the same height x, there will be a definite average height. This will be considered the "true" value of y for that x, and the observed y's will differ from it by apparently random amounts.

Thus, the data we must work with necessarily shows some amount of variation or scatter from the "true" values. The ever-present variation makes it impossible for us to determine the true values, and hence the true function, exactly. The best we can do is to come up with a good estimate of the "true" function. The curve joining the true values of y for each x is called the regression curve, after the work of Galton. The process of estimating the regression curve is called regression analysis. One reason regression analysis is valuable is that when large amounts of data are used, some experimental errors cancel out so that the value of y given by the fitted equation is a more reliable estimate of the true value of y than the observed value is.

Exercises

1.1 The equation in Example 1 will be used to estimate the organic acid content from the titration results. What is the estimated acid content for samples with these acid numbers by titration: 62, 100, 159? Do the estimated values agree exactly with the experimental results? What are some possible sources of experimental error in this example?

1.2 In Example 2 (heat of fusion), the equation $y = mx$ can be rearranged to give $m = y/x$. Estimate m by computing the mean (average) value of y/x for the data in Table 2. What value for β do you get by this method?

1.3 What sources of experimental error might be present in Example 3?

1.4 In Example 4, what is the average (mean) value of y for $x = 66.5$? For $x = 71.5$? Do these values support Galton's theory?

1.5 State four purposes of fitting an equation to data. Give an example illustrating each.

1.6 Find an example of curve fitting in Science, Scientific American, or other periodical of your choice. Make an oral report on the example, answering the following questions:

a) What was the purpose for fitting the curve?
b) What type of curve was used? (Free-hand, line, parabola, log, power curve, etc.)
c) Were the variables related by cause and effect, or more by circumstance or common cause?
d) What were some of the possible sources of experimental error?
e) Was an indication given as to how well the curve fit the data?

2. SCATTER DIAGRAMS

2.1 Scatter Diagrams and Free-hand Curves

The first step toward estimating the "true" function is to get an idea of what its graph looks like. We do this by plotting the data in a two-dimensional coordinate system, with x-coordinates on the horizontal axis and y-coordinates on the vertical. The resulting picture is called a scatter diagram. (For ease of interpretation, the diagram should be carefully labeled.)

Figures 1 and 2 show the scatter diagrams for the data in Examples 1 and 3.

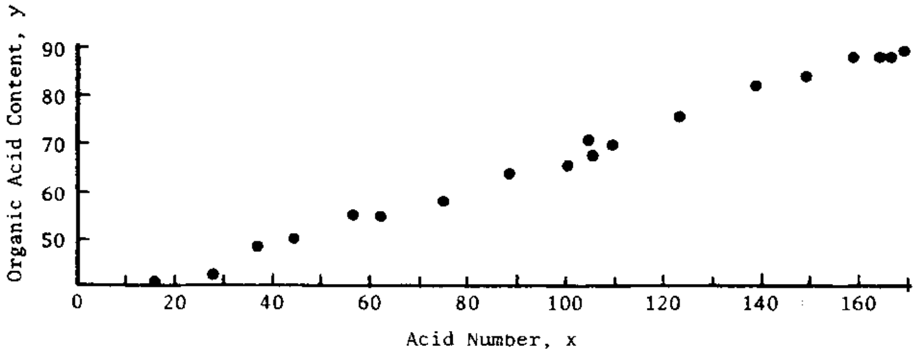


Figure 1. Plot of Acid Number vs. Organic Acid Content

These diagrams give you at a glance a much better idea of the relation between x and y than you get from looking at Tables 1 and 3. In both examples, y increases with x. In Fig. 1, the points appear to lie on a straight line, while in Fig. 2, the points show downward curvature--the slope decreases with increasing x. You could sketch a curve through the data points. This would be your free-hand estimate of the regression curve. (In some applications, this may be all you need.)

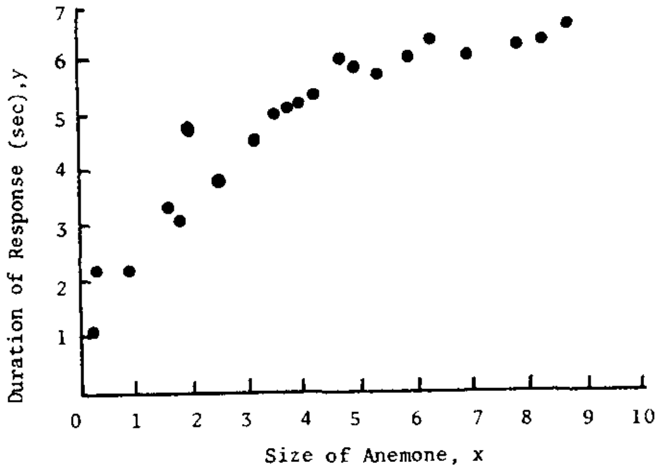


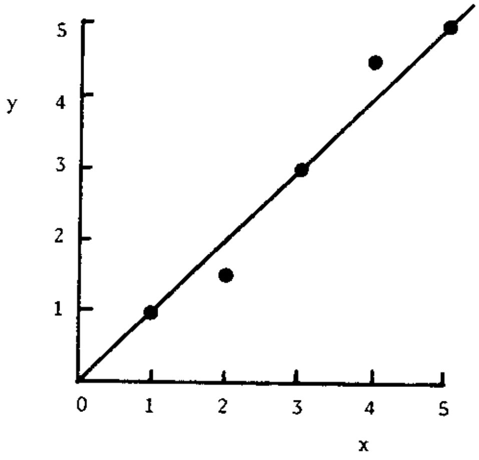
Figure 2. Response of Sea Anemones to Standard Stimulus

In addition to indicating the general shape of the curve, the scatter diagram shows something about the amounts of variation or scatter in y. In Fig. 1, all the points are close to a line; the scatter is small. In Fig. 2, some points are rather far from the free-hand curve (wherever you draw it), so the scatter is larger. If the scatter is too great, it is impossible to draw a good free-hand curve. (See Exercise 2.4.)

In Section 3, we shall see how to measure the scatter of the data points about a curve drawn through them. For now, let's consider the vertical scatter, that is, the variation in the y values alone, ignoring any dependence on x. This will be useful later on in Section 5 when we try to decide if the curve fits the data well. (If the scatter about the curve is as great as the scatter in the y values alone, then we have gained nothing by fitting the curve.)

Consider the five-point example in Fig. 3. The points seem to lie along the line y = x, which is one person's free-hand estimate of the regression line. If we ignore the x values and just plot the y's, we get the one-dimensional vertical scatter diagram in Fig. 4(a). Figures 4(b) and 4(c) show what one might get in other examples.

Scatter refers to the dispersion of the points about some central position. For this, we use the <u>mean</u> value of y, denoted $\bar{y}$. If the y values are y_1, y_2, ... , y_N, then, by definition, the mean is



x	y
1	1.0
2	1.5
3	3.0
4	4.5
5	5.0

Figure 3. A 5 Point Example

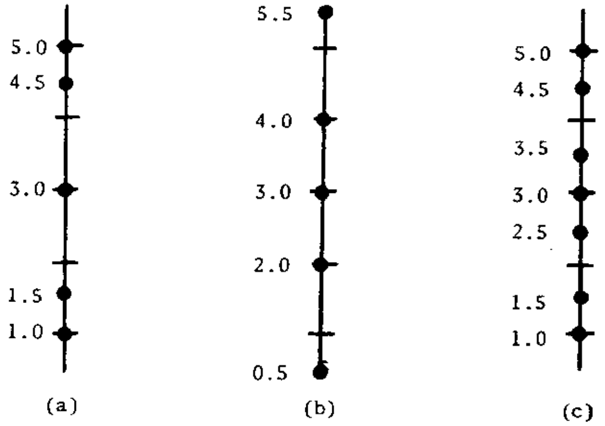


Figure 4. Vertical Scatter Diagrams

(2.1) $$\bar{y} = \sum_{i=1}^{N} y_i/N.$$

In a way, $\bar{y}$ is the "center of mass" of the y values. In Fig. 4(a),

$$\bar{y} = (1 + 1.5 + 3 + 4.5 + 5)/5 = 15/5 = 3.$$

In Figs. 4(b) and 4(c), $\bar{y} = 3$ also.

The difference between a particular y value y_i and the mean $\bar{y}$ is called its deviation from the mean. We might try to measure the scatter by the sum of the deviations. If we did this for the data in Fig. 4(a) we would get

$$\sum_{i=1}^{N} (y_i-\overline{y}) = (1-3) + (1.5-3) + (3-3) + (4.5-3) + (5-3)$$

$$= -2 + -1.5 + 0 + 1.5 + 2 = 0.$$

In fact, the sum of the deviations from the mean is <u>always</u> zero, because

$$\sum_{i=1}^{N} (y_i-\overline{y}) = \sum_{i=1}^{N} y_i - \sum_{i=1}^{N} \overline{y} = N\overline{y} - N\overline{y} = 0.$$

The sum of the deviations is not a good measure of scatter. Some of the points are above the mean, some are below, and the positive and negative deviations cancel out.

One way to correct this is to use the absolute values of the deviations. For the data in Fig. 4(a) we get

$$\sum_{i=1}^{N} |y_i-\overline{y}| = |-2| + |-1.5| + |0| + |1.5| + |2| = 7.$$

For the data in 4(b) and 4(c), the sums are 7 and 8. The sum in 4(c) is larger simply because there are more points. Actually, the scatter in (c) is less than that in (a) or (b), because the points are clustered more tightly about the mean. This suggests that we should calculate mean absolute deviation by dividing the sum by the number of points. The mean absolute deviations for (a), (b), and (c) are 7/5 = 1.4, 7/5 = 1.4 and 8/7 = 1.14. This does provide a reasonable way to measure scatter, and it is sometimes used. However, it is difficult to work with, both in practice and in statistical theory, because absolute values do not combine easily. Notice also that according to this measure, Figs. 4(a) and 4(b) have the same amount of scatter. Since the points in (b) are more spread out, we might want to say that (b) really shows more scatter than (a).

<u>2.2 Variance</u>

Another way to make the deviations positive is to square them. This leads to the <u>mean squared deviation</u>, or <u>variance</u>, denoted by var(y). By definition,

$$\text{(2.2)} \qquad \text{var}(y) = \frac{\sum_{i=1}^{N} (y_i-\overline{y})^2}{N} .$$

For example, in Fig. 4(a),

$$\mathrm{var}(y) = \frac{(-2)^2 + (-1.5)^2 + (0)^2 + (1.5)^2 + (2)^2}{5}$$

$$= \frac{12.5}{5} = 2.5.$$

In 4(b),

$$\mathrm{var}(y) = \frac{(-2.5)^2 + (1)^2 + (0)^2 + (1)^2 + (2.5)^2}{5}$$

$$= \frac{14.5}{5} = 2.9.$$

By this measure, (b) shows more scatter than (a). In (c), var(y) = 1.85, so (c) shows less scatter than (a) or (b). The variance is the measure of scatter we shall use.

For more realistic examples, which normally have many more data points, another formula for the variance is useful:

$$\text{(2.3)} \qquad \mathrm{var}(y) = \sum_{i=1}^{N} y_i^2/N - \bar{y}^2.$$

We shall derive this in the next section.

The square root of the variance is called the *standard error*. It indicates how far a typical measurement is from the mean. The standard error gives an idea of the reliability of the estimate, or its usefulness in prediction.

Example 4 continued (Galton)

In Table 4, there are 10 values of y for which x = 69.5. Let's compute the mean, variance, and standard error for these. A hand-held calculator is useful!

	y_i	y_i^2
	67.2	4516
	67.2	4516
	68.2	4651
	69.2	4789
	69.2	4789
	70.2	4928
	70.2	4928
	71.2	5069
	72.2	5213
	73.2	5358
TOTALS:	698.0	48757

Mean:

$$\bar{y} = 698/10 = 69.8$$

Variance:

$$\mathrm{var}(y) = 48757/10 - (69.8)^2 = 3.6$$

Standard error:

$$\sqrt{3.6} = 1.9 \text{ inches}$$

Note: In computing the variance and in the other computations in this unit you do not round off too soon. For var(y) to be correct to the nearest tenth when there are 10 measurements, Σy_i^2 must be correct to the nearest unit. In this example, being correct to the nearest unit requires five significant digits.

We have been discussing the vertical scatter. The horizontal scatter, the scatter in the x values alone, is measured similarly. The mean and variance in x are

$$\bar{x} = \sum_{i=1}^{N} x_i/N \tag{2.4}$$

and

$$\text{var}(x) = \sum_{i=1}^{N} (x_i-\bar{x})^2/N$$

$$= \sum_{i=1}^{N} x_i^2/N - \bar{x}^2. \tag{2.5}$$

Exercises

2.1 Sketch a line through the points in Fig. 1. (Use a transparent ruler or thread, or put your eye at table level and sight through the points.) Find the slope of your line, and write an equation for it.

2.2 Make a scatter diagram for the data in Example 2. Sketch a line through the data points. Make sure your line goes through the origin. (Why?) Find the slope of your line. If m is your slope, then 50m is your estimate of β by the "eye-ball" method. Compare this estimate with the one you got in Exercise 1.2.

2.3 Sketch a curve through the data points in Fig. 2. What types of functions have graphs like this?

2.4 The following data were collected on 35 heifers. x = a certain body measurement at birth (cm); y = same body measurement at one year (cm).

x	22.2	19.0	21.3	20.5	18.5	19.0	24.8	19.1	25.1
y	28.5	22.4	27.1	29.0	27.7	29.1	31.5	23.6	38.3

x	24.7	22.1	20.8	19.3	22.3	25.7	23.2	21.5	25.2
y	39.3	34.3	35.0	28.3	34.2	37.2	31.6	33.0	36.3

x	25.0	19.7	25.0	20.3	19.3	24.0	24.0	22.9	18.2
y	37.6	31.7	36.6	30.1	29.4	39.9	38.5	30.0	23.7

x	21.2	24.7	18.9	23.2	19.9	23.1	18.9	19.5
y	28.4	39.8	28.6	34.9	34.1	33.2	28.7	32.5

Make a scatter diagram from these data. Notice that the scatter is considerable, making it impossible to decide where to sketch a curve.

2.5 In a course on computer programming, a large group of business administration students were given instructional material and allowed to proceed at their own pace. At the end of the course, ten students were selected at random to take an achievement test. The following data were collected:

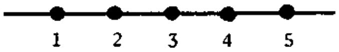
x = hours spent in completing the course	30	25	50	38	20	70	35	24	60	45
y = achievement test score	80	80	45	70	95	20	50	90	25	50

Make a scatter diagram. Sketch a line through the data points. Estimate the test score of a student who required 40 hours to complete the course.

2.6 Compute the mean and variance in y for the data in Table 1.

2.7 Compute the mean and variance for y in Exercise 2.4.

2.8 In Fig. 3, if we ignore the y's and just plot the x's, we get this one-dimensonal horizontal scatter diagram:



Compute the mean and variance for this.

2.9 Compute the mean and variance for y and for x in Exercise 2.5.

3. THE LEAST-SQUARES METHOD

3.1 The Sum of the Squares of the Deviations

The free-hand method of curve fitting may be adequate in some situations, but it is subjective in that different people will get different results. And when the scatter is considerable, no one can decide, just by "eye-balling" it, where to draw the curve. We need an analytic method that is reproducible and workable even when there is considerable scatter.

So, suppose we have N data points, (x_i, y_i) for i = 1, 2, ..., N, and suppose we have decided on the kind of function y = f(x) we want to use. (How to decide this will be discussed more in Unit 434.) That is, suppose we know f(x) except for some undetermined coefficients a, b, c, m, etc. For instance, if the curve is a line, then f(x) has the form mx + b. For a parabola, $f(x) = ax^2 + bx + c$. Or, we might wish to fit a curve of the form y = m log x + b or $y = ax^m$, and so on. The problem now is to find the values of the coefficients that make the function fit the data best.

Obviously, we need some measure of "goodness of fit." Since we are trying to predict y from x, the function will fit well if the distances between the function values, $f(x_i)$, and the observed values of y, the y_i, are small, taken as a whole. The difference $y_i - f(x_i)$ is called a deviation. (See Fig. 5.) The most frequently used criterion for fitting the function is the "least-squares" criterion, which says: "Choose the coefficients to minimize the sum of the squares of the deviations." In symbols, we want the values of a, b, c, etc. that minimize

$$S = \sum_{i=1}^{N} [y_i - f(x_i)]^2. \tag{3.1}$$

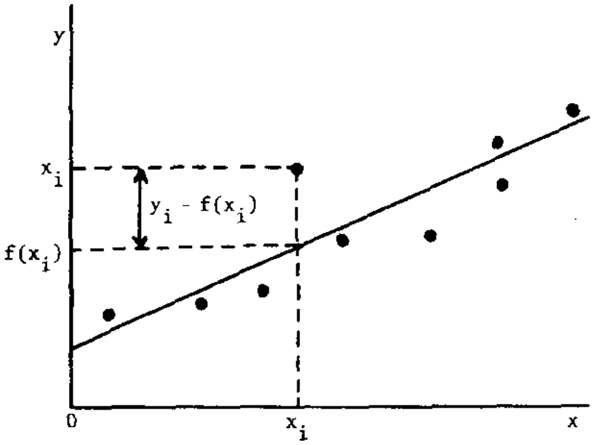


Figure 5. Deviations

As in the previous section, we take the mean squared deviation, S/N, as a measure of the amount of scatter of the data points about the fitted curve. In the five-point example of Fig. 3, the fitted line was y = x. The sum of squares for this line is

$$\begin{aligned} S &= \sum_{i=1}^{N} (y_i - x_i)^2 \\ &= (1-1)^2 + (2-1.5)^2 + (3-3)^2 + (4-4.5)^2 + (5-5)^2 \\ &= 0.5. \end{aligned}$$

Thus, the scatter about the line is S/N = 0.5/5 = 0.1. (This is not the least-squares line. See Exercise 3.3.)

The least-squares (LS) fit is preferred for several reasons. It produces a function that generally corresponds to our free-hand curve in simple cases. It is relatively easy to compute. Also, if the data satisfies certain assumptions, the LS fit can be shown to have some nice statistical properties. For instance, suppose the true function is the line $y = mx + b$ and the fitted equation, computed by the LS method, is $y = \hat{m}x + \hat{b}$. The numbers $\hat{m}$ and $\hat{b}$ (read "m hat" and "b hat") will not necessarily equal m and b. But if we were to repeat the experiment many, many times, each time computing $\hat{m}$ and $\hat{b}$, then the averages of the values of $\hat{m}$ and $\hat{b}$ would approach the true values m and b. (Check one of the statistics books in the bibliography if you want to pursue this further.)

3.2 Fitting A Constant Function

The simplest function to fit is the constant function, $f(x) = b$. If the true function is constant, then y does not depend on x at all. The problem then is to estimate the one true value of y based on measured values y_i, y_2, ..., y_N. The deviations are $y_i - b$. The least-squares criterion says: Choose b so as to minimize the sum

$$S = \sum_{i=1}^{N} (y_i - b)^2. \tag{3.2}$$

For example, in Fig. 4(a) we have $y_1 = 1$, $y_2 = 1.5$, $y_3 = 3$, $y_4 = 4.5$, $y_5 = 5$. Thus,

$$S = S(b) = (1-b)^2 + (1.5-b)^2 + (3-b)^2 + (4.5-b)^2 + (5-b)^2.$$

This expression gives S as a quadratic function of b. (See Exercise 3.1.) When we expand the squares and collect terms, we find that

$$\begin{aligned} S &= (1-2b+b^2) + (2.25-3b+b^2) + (9-6b+b^2) \\ &\qquad + (20.25-9b+b^2) + (25-10b+b^2) \\ &= (1+2.25+9+20.25+25) - (2+3+6+9+10)b + 5b^2 \\ &= 57.5 - 30b + 5b^2. \end{aligned}$$

To find the value of b that minimizes this, we use the following fact from algebra, which will be used again in the next section. (See the Appendix for a proof.)

LEMMA. A quadratic function of the form

$$Y = AX^2 + BX + C \tag{3.3}$$

with positive squared term $(A > 0)$ attains its minimum value when $X = -B/(2A)$. The minimum value of the function is $Y = C - B^2/(4A)$.

When we compare $S = 57.5 - 30b + 5b^2$ with Eq. (3.3), we find that S corresponds to Y, b corresponds to X, $A = 5$, $B = -30$, and $C = 57.5$. Therefore, the value of b that minimizes S is

$$\hat{b} = \frac{-B}{2A} = -\frac{(-30)}{2(5)} = 3.$$

The minimum value of S is

$$\hat{S} = C - \frac{B^2}{4A} = 57.5 - \frac{(-30)^2}{4(5)} = 57.5 - 45 = 12.5.$$

Note that $\hat{b}$ is just the mean value $\overline{y}$ of y that we computed in Section 2.

This example illustrates what happens in general. To minimize S in (3.2), expand the squares and collect terms:

$$\begin{aligned} S &= \sum_{i=1}^{N} (y_i-b)^2 \\ &= \sum_{i=1}^{N} (y_i^2-2y_ib+b^2) \\ &= \sum_{i=1}^{N} y_i^2 - 2\sum_{i=1}^{N} y_ib + \sum_{i=1}^{N} b^2. \end{aligned}$$

Now $\Sigma y_i = N\overline{y}$ and $\Sigma b^2 = Nb^2$, so that

$$S = \sum_{i=1}^{N} y_i^2 - 2N\overline{y}\, b + Nb^2.$$

Comparing this with Eq. (3.3), the standard form in the lemma, we find that

$$A = N,\ B = -2N\overline{y},\ C = \Sigma y_i^2.$$

Therefore, the value of S is minimized for

$$\hat{b} = \frac{-B}{2A} = -\frac{(-2N\overline{y})}{2N} = \overline{y},$$

and the minimum value of S is

$$\hat{S} = C - \frac{B^2}{4A} = \Sigma y_i^2 - \frac{4N^2\overline{y}^2}{4N} = \Sigma y_i^2 - N\overline{y}^2.$$

Now we have two expressions for $\hat{S}$, the one in the previous sentence and $\hat{S} = \Sigma(y_i-\overline{y})^2$, obtained by substituting $\hat{b} = \overline{y}$ into Eq. (3.2). Setting these equal and dividing by N, we get

$$(3.4) \qquad \frac{\hat{S}}{N} = \sum_{i=1}^{N} \frac{(y_i-\overline{y})^2}{N} = \sum_{i=1}^{N} \frac{y_i^2}{N} - \overline{y}^2.$$

By definition (2.2), the expression in the middle is var(y). This shows that formula (2.3) for the variance is correct.

In summary:

The least-squares fit of a constant function $f(x) = b$ to data points $y_i, y_2, \ldots, y_N$ is obtained by minimizing

$$S = \sum_{i=1}^{N} (y_i - b)^2 .$$

The estimated value of b is

$$\hat{b} = \bar{y},$$

and the minimum value of S is

$$\hat{S} = \sum_{i=1}^{N} y_i^2 - N\bar{y}^2 = N \, \mathrm{var}(y) .$$

Exercises

3.1 Complete the table below by computing

$$S = (1-b)^2 + (1.5-b)^2 + (3-b)^2 + (4.5-b)^2 + (5-b)^2$$

for the values of b given. Make a plot of S vs. b, and sketch a parabola through the points. Notice that the low point on the parabola is at b = 3.

b	1	2	3	4	5
S			12.5		

3.2 In an introductory surveying class, 20 students in 10 teams measured the length of a property line with a steel tape. The results are shown below. Estimate the true length by the LS method by fitting a constant function f(x) = b to the data. Compute the scatter about the estimated value (i.e. the variance).

Team i	1	2	3	4	5
Measured Length, y_i (feet)	147.9	145.9	145.5	143.2	149.9

i	6	7	8	9	10
y_i	150.7	148.3	152.9	151.5	144.8

3.3 Copy the five-point scatter diagram in Fig. 3 and plot the line $y = 1.1x - 0.3$ on it. Compute the sum of the squares of the deviations from this line. Note that it is less than the value of S for the line $y = x$.

3.4 a) Using the data in Example 2, compute

$$S = \sum_{i=1}^{30} (y_i - mx_i)^2$$

for the value of m you got in Exercise 1.2.

b) Now compute S for the value of m you got in Exercise 2.2. Which value of m gives the better fit?

3.5 Derive (2.3) directly by showing that

$$\sum_{i=1}^{N} (y_i - \overline{y})^2 = \sum_{i=1}^{N} y_i^2 - N\overline{y}^2.$$

Hint: $$(y_i - \overline{y})^2 = y_i^2 - 2y_i\overline{y} + \overline{y}.$$

The middle term on the right side of this last expression sums to $-2N\overline{y}^2$.

---

4. FITTING A LINE

4.1 Introduction

Let's assume now that the regression curve is a straight line, so that the true value of y for each x is given by y = mx + b for some unknown values of m and b. We want to estimate m and b by the LS method, that is, by choosing m and b so as to minimize

$$S = S(m,b) = \sum_{i=1}^{N} (y_i - mx_i - b)^2. \tag{4.1}$$

First we shall consider the special case of lines through the origin. That is, we'll assume b = 0 and try to minimize S letting m vary. The result will be useful in its own right and will also generalize to the case of lines through any given point. Then we shall consider lines of the same slope; we'll hold m constant and try to minimize S by letting b vary. Finally, we'll consider all possible lines, letting both m and b vary.

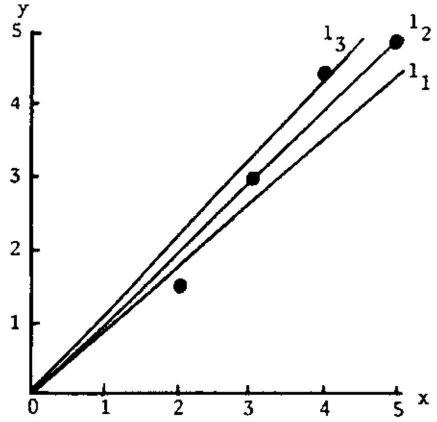
4.2 Fitting a Line Through the Origin

Suppose we know that the regression line goes through the origin, so that b = 0 and y = mx. We are looking for the value of m that minimizes

$$S = \sum_{i=1}^{N} (y_i - mx_i)^2. \tag{4.2}$$

For example, let's look for the line y = mx that best fits the 5 points in Figure 3. Some possible lines and the corresponding values of S are shown in Figure 6. The line y = x gives the smallest S among those shown. Does it have the smallest S among all lines through the origin? Let's see. From (4.2),

$$S = S(m) = (1-m)^2 + (1.5-2m)^2 + (3-3m)^2$$
$$+ (4.5-4m)^2 + (5-5m)^2.$$



line	m	S
l_2	0.9	1.25
l_2	1.0	.50
l_3	1.1	.85

Figure 6. Lines through the origin.

Expanding the squares and collecting terms, we get

$$S = (1-2m+m^2) + (2.25-6m+4m^2) + (9-18m+9m^2)$$
$$+ (20.25-36m+16m^2) + (25-50m+25m^2)$$
$$= 57.5 - 112m + 55m^2.$$

This is a quadratic function of m, so the Lemma applies with A = 55, B = -112, and C = 57.5. The value of S is minimized when m equals

$$\hat{m} = \frac{-B}{2A} = -\frac{(-112)}{2(55)} = 1.02.$$

The minimum value of S is

$$\hat{S} = C - \frac{B^2}{4A} = 57.5 - \frac{(-112)^2}{4(55)} = .48.$$

Thus, the LS line through the origin is

$$y = 1.02x.$$

(Perhaps it is surprising that y = x was not the LS line through the origin. In general, the points farthest from the origin have the greatest effect on the value of $\hat{m}$. Thus the LS line runs between the points (4,4.5) and (5,5) in Fig. 3.)

In general, we can proceed in the same way to minimize (4.2). Expand the squares and collect terms:

$$\sum_{i=1}^{N} (y_i - mx_i)^2 = \sum_{i=1}^{N} (y_i^2 - 2x_iy_im + x_i^2m^2)$$

$$= \sum_{i=1}^{N} y_i^2 - 2 \sum_{i=1}^{N} x_iy_im + \sum_{i=1}^{N} x_i^2m^2.$$

Now apply the Lemma, with

$$A = \Sigma x_i^2, \ B = -2\Sigma x_iy_i, \text{ and } C = \Sigma y_i^2.$$

The value of S is minimized when m equals

$$\hat{m} = \frac{-B}{2A} = - \frac{-2 \sum_{i=1}^{N} x_iy_i}{2 \sum_{i=1}^{N} x_i^2}$$

$$= \frac{\sum_{i=1}^{N} x_iy_i}{\sum_{i=1}^{N} x_i^2}.$$

The minimum value of S is

$$\hat{S} = C - \frac{B^2}{4A} = \sum_{i=1}^{N} y_i^2 - \frac{(-2 \sum_{i=1}^{N} x_iy_i)^2}{4 \sum_{i=1}^{N} x_i^2}$$

$$= \sum_{i=1}^{N} y_i^2 - \frac{(\sum_{i=1}^{N} x_iy_i)^2}{\sum_{i=1}^{N} x_i^2} .$$

The formula for $\hat{m}$ can be written as

$$\Sigma x_iy_i = \hat{m} \ \Sigma x_i^2.$$

Substituting this into the formula for $\hat{S}$, we get

$$\hat{S} = \sum_{i=1}^{N} y_i^2 - \frac{(\hat{m} \sum_{i=1}^{N} x_i^2)^2}{\sum_{i=1}^{N} x_i^2}$$

$$= \sum_{i=1}^{N} y_i^2 - \hat{m}^2 \sum_{i=1}^{N} x_i^2.$$

In summary:

The least squares fit of a line through the origin, $y = mx$, to data points (x_1,y_1), (x_2,y_2), ..., (x_N,y_N) has slope m equal to

(4.3) $$\hat{m} = \frac{\sum_{i=1}^{N} x_i y_i}{\sum_{i=1}^{N} x_i^2}.$$

The minimum value of S is

(4.4) $$\hat{S} = \sum_{i=1}^{N} y_i^2 - \hat{m}^2 \sum_{i=1}^{N} x_i^2.$$

Example 2 continued (Heat of fusion)

Since adding no ice should result in no temperature change, the true line goes through the origin. We compute the LS line, $y = \hat{m}x$.

	x_i	y_i	x_i^2	$x_i y_i$	y_i^2
	13.4	19	179.56	254.6	361
	12.8	19	163.84	243.2	361
	15.1	22	228.01	332.2	484
	19.7	30	388.09	591.0	900
	17.5	27	306.25	472.5	729
	19.3	29	372.49	559.7	841
	12.5	18	156.25	225.0	324
	11.5	18	132.25	207.0	324
	14.2	21	201.64	298.2	441
	11.2	17	125.44	190.4	289
	10.5	16	110.25	168.0	256
	11.8	17	139.24	200.6	289
	10.2	16	104.04	163.2	256
	13.6	20	184.96	272.0	400
	18.1	29	327.61	524.9	841
	17.8	27	316.84	480.6	729
	18.1	28	327.61	506.8	784
	14.2	23	201.64	326.6	529
	15.9	25	252.81	397.5	625
	18.3	27	334.89	494.1	729
	19.3	29	372.49	559.7	841
	13.9	21	193.21	291.9	441
	15.6	24	243.36	374.4	576
	10.7	16	114.49	171.2	256
	18.5	28	342.25	518.0	784
	17.2	25	295.84	430.0	625
	15.0	22	225.00	330.0	484
	14.2	20	201.64	284.0	400
	19.0	28	361.00	532.0	784
	17.2	26	295.84	447.2	676
TOTALS	456.3	687	7198.83	10846.5	16359

$\hat{m} = \frac{10846.5}{7198.83} = 1.5067$ $\quad \hat{S} = 16359 - (1.5067)^2(7198.83) = 16.6$

The LS line through the origin is $y = 1.51x$. The class estimated the value of β as $55\hat{m} = 83$ cal./g. (The accepted value is closer to 79.7 cal./g.)

In Example 2, we knew that the regression line passed through the point (0,0), so we made our estimate pass through that point also. More generally, suppose we want our estimate to pass through the point (x_0,y_0). The equation for a line through (x_0,y_0) is $y = y_0 + m(x-x_0)$, so the expression to be minimized (3.1) is

$$S = \sum_{i=1}^{N} [y_i - y_0 - m(x_i-x_0)]^2. \tag{4.5}$$

We could proceed as we did for (4.2) by expanding the squares and collecting terms, but there is an easier way! Let's make the substitutions $y_i' = y_i - y_0$ and $x_i' = x_i - x_0$. (x_i' and y_i' will be numbers calculated from the data.) Then we can write (4.5) as

$$S = \sum_{i=1}^{N} (y_i' - mx_i')^2$$

This has the same form as (4.2), using the points (x_i',y_i') instead of (x_i,y_i). All we have done is to take (x_0,y_0) as our origin and compute the coordinates of the data points relative to that. Using the results (4.3) and (4.4), we find S is minimized when m equals

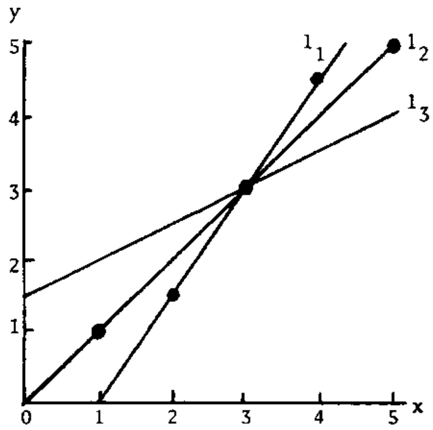
$$\hat{m} = \frac{\sum_{i=1}^{N} x_i'y_i'}{\sum_{i=1}^{N} (x_i')^2} = \frac{\sum_{i=1}^{N} (x_i-x_0)(y_i-y_0)}{\sum_{i=1}^{N} (x_i-x_0)^2}, \tag{4.6}$$

The corresponding value of S is

$$\hat{S} = \sum_{i=1}^{N} (y_i') - \hat{m}^2 \sum_{i=1}^{N} (x_i)^2 \tag{4.7}$$

$$= \sum_{i=1}^{N} (y_i-y_0)^2 - \hat{m}^2 \sum_{i=1}^{N} (x_i-x_0)^2.$$

For example, consider lines through the point (3,3) in Fig. 3. Some of these are shown with the corresponding values of S in Fig. 7. The line $y = x$ has the smallest S among the ones shown, but we suspect that S would be smaller for a line with slope slightly greater than 1. (This would decrease the deviations of the points (2,1.5) and (4,4.5) from the line.) To verify this and find $\hat{m}$, we compute the coordinates relative to (3,3):



line	m	S
l_1	0.5	4.0
l_2	1.0	0.5
l_3	1.5	2.0

Figure 7. Lines through (3,3).

	x_i	y_i	x_i'	y_i'	$(x_i')^2$	$x_i'y_i'$	$(y_i')^2$
	1	1.0	-2.0	-2.0	4	4.0	4.0
	2	1.5	-1.0	-1.5	1	1.5	2.25
	3	3.0	0	0	0	0	0
	4	4.5	1.0	1.5	1	1.5	2.25
	5	5.0	2.0	2.0	4	4.0	4.0
TOTALS			0	0	10	11	12.5

Thus,

$$\hat{m} = 11/10 = 1.1$$

and

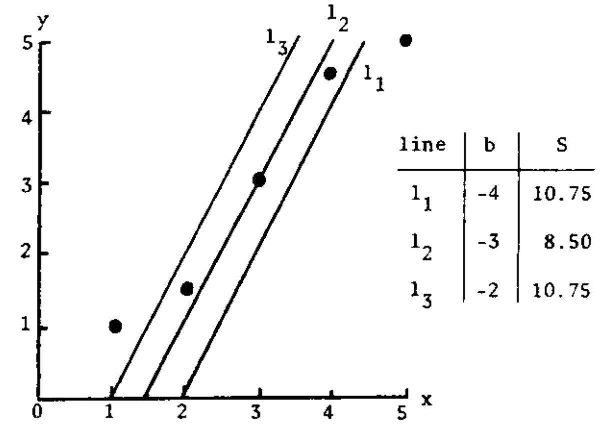
$$\hat{S} = 12.5 - (1.1)^2(10) = 0.4.$$

We conclude that the LS line through the point (3,3) is

$$\begin{aligned} y &= 3 + 1.1(x-3) \\ &= 1.1\,x - 0.3. \end{aligned}$$

4.3 Fitting a Line with Given Slope

Now suppose we want our line to have a given slope m. We hold m fixed and vary b, trying to minimize the value of S in Eq. (4.1). Lines with the same slope are parallel, so we are looking for the line in a complete set of parallel lines that gives the smallest S value.



line	b	S
l_1	-4	10.75
l_2	-3	8.50
l_3	-2	10.75

Figure 8. Lines of slope 2.

For example, consider lines with slope 2 in the five-point example of Fig. 3. Some of these and the corresponding values of S are shown in Fig. 8. The middle line, y = 2x - 3, has the smallest S among the lines shown. Does it have the smallest value of S among all lines of slope 2? Let's see. From Eq. (4.1),

$$\begin{aligned} S &= \sum_{i=1}^{5} (y_i - 2x_i - b)^2 \\ &= (1-2-b)^2 + (1.5-4-b)^2 + (3-6-b)^2 \\ &\qquad + (4.5-8-b)^2 + (5-10-b)^2 \\ &= (-1-b)^2 + (-2.5-b)^2 + (-3-b)^2 + (-3.5-b)^2 \\ &\qquad + (-3-b)^2 + (-5-b)^2. \end{aligned}$$

This equation expresses S as a quadratic function of b. We could expand the squares, collect terms, and apply the Lemma, but again there is an easier way. The last expression for S above has the same form as (3.2) with the data values -1, -2.5, -3, -3.5, and -5. We have already seen that the minimum value of S for that expression occurs when b equals the mean data value. Therefore,

$$\hat{b} = \frac{(-1 - 2.5 - 3 - 3.5 - 5)}{5} = \frac{-15}{5} = -3.$$

We conclude that the LS line with slope 2 is y = 2x - 3.

Notice that the line y = 2x - 3 passes through the point $(\bar{x},\bar{y}) = (3,3)$. In any scatter diagram, the point

$(\bar{x},\bar{y})$ is the center of mass of the data points. Does the LS line of given slope m always pass through the center of mass $(\bar{x},\bar{y})$? Once more we examine (4.1). This time, following the short cut used in the previous paragraph, we view the problem as one to fit the constant b to the data points $y_i - mx_i$. (Remember m is a fixed number for now.) The value of b that gives the minimum S is the mean:

$$\hat{b} = \sum_{i=1}^{N} \frac{(y_i - mx_i)}{N} = \frac{\sum_{i=1}^{N} y_i}{N} - m \frac{\sum_{i=1}^{N} x_i}{N} = \bar{y} - m\bar{x}.$$

This shows that $\bar{y} = m\bar{x} + \hat{b}$. In other words, the point $(\bar{x},\bar{y})$ is indeed on the line.

4.4 Fitting the General Line

We are finally in a position to find the general LS line. We allow m and b both to vary in Eq. (4.1), and we ask for the values $\hat{m}$ and $\hat{b}$ that minimize S. For any particular m, the value of S will be least when $b = \bar{y} - m\bar{x}$, i.e. when the line goes through the center of mass $(\bar{x},\bar{y})$. But among all lines through $(\bar{x},\bar{y})$, the value of m which minimizes S is given by (4.6). We conclude:

The general least-squares line is $y = \hat{m}x + \hat{b}$, where

$$\hat{m} = \frac{\sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{N} (x_i - \bar{x})^2} \tag{4.8}$$

and

$$\hat{b} = \bar{y} - \hat{m}\bar{x}. \tag{4.9}$$

The line goes through the center of mass $(\bar{x},\bar{y})$.

Formula (4.8) is not very convenient for computations, since it involves computing coordinates relative to the center of mass. Rounding off $\bar{x}$ and $\bar{y}$ leads to errors in $x_i - \bar{x}$ and $y_i - \bar{y}$ that accumulate in the sums. Therefore, another formula is desirable. By (2.5), the denominator in (4.8) can be written as

$$x_i^2 - N\bar{x}^2.$$

The numerator can also be simplified:

$$\sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{N} x_i(y_i - \bar{y}) - \sum_{i=1}^{N} \bar{x}(y_i - \bar{y}).$$

The last term on the right is zero, because $\Sigma(y_i - \bar{y}) = 0$.

Since $\Sigma x_i = N\bar{x}$,

$$\sum_{i=1}^{N} x_i(y_i-\bar{y}_i) = \sum_{i=1}^{N} x_iy_i - \sum_{i=1}^{N} x_i\bar{y} = \sum_{i=1}^{N} x_iy_i - N\bar{x}\bar{y}.$$

So the formula for $\hat{m}$ preferred for computation is

$$\hat{m} = \frac{\sum_{i=1}^{N} x_iy_i - N\bar{x}\bar{y}}{\sum_{i=1}^{N} x_i^2 - N\bar{x}^2}. \tag{4.10}$$

4.5 The Uptake of Nitrogen by Grasses

In a greenhouse experiment on the growth of a grass mixture, known amounts of nitrogen were added to the compost in which the grass was grown, and the amount of nitrogen in the mature grass was measured. The results for 11 pots are shown in Table 5.

TABLE 5

UPTAKE OF NITROGEN BY GRASSES

Nitrogen in compost, x (grams per pot)	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
Nitrogen in grass, y (grams per pot)	0.12	0.23	0.32	0.25	0.26	0.52	0.53	0.63	0.63	0.69	0.73

The scatter diagram (Fig. 9) shows the relation of y to x to be approximately linear.

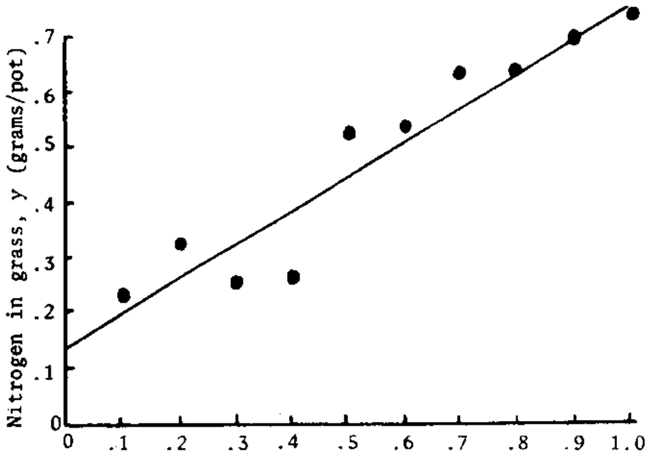


Figure 9. Uptake of nitrogen.

Let's compute the sums and the coefficients for the LS line.

Data Point	x_i	x_i^2	y_i	y_i^2	x_iy_i
1	0.0	0.00	0.12	0.0144	0.000
2	.1	.01	.23	.0529	.023
3	.2	.04	.32	.1024	.064
4	.3	.09	.25	.0625	.075
5	.4	.16	.26	.0676	.104
6	.5	.25	.52	.2704	.260
7	.6	.36	.53	.2809	.318
8	.7	.49	.63	.3969	.431
9	.8	.64	.63	.3969	.504
10	.9	.81	.69	.4761	.621
11	1.0	1.00	.73	.5329	.730
Totals	5.5	3.85	4.91	2.65	3.130

$\bar{x} = \frac{5.5}{11} = 0.5$ $\qquad\qquad$ $\bar{y} = \frac{4.91}{11} = 0.4464$

Using (4.10),

$$\hat{m} = \frac{3.130 - (11)(0.5)(0.4464)}{3.85 - (11)(0.5)(0.5)} = \frac{0.675}{1.10} = 0.614.$$

From (4.9),

$$\hat{b} = 0.4464 - 0.614(0.5) = .139.$$

The LS line is

$$y = 0.61x + 0.14.$$

This is the line drawn in Fig. 9.

Exercises

4.1 An experiment was done in which temperatures were taken at various depths in a gold mine in South Africa. The following table, derived from the data, uses a point 324 feet below the earth's surface as an origin for both temperature and depth. The number of feet below the origin is x, and y is the number of degrees Fahrenheit above the temperature at the origin (which was 70.0°F).

x	300	600	1750	2500	3400	3700
y	1.0	4.0	8.0	14.5	18.0	25.0

Make a scatter diagram for these data. Compute the LS line through the origin using (4.3), and plot the line. (Data from Watson, T.L., "Underground Temperatures," Science 33:828-831 (1911).)

4.2 Compute the LS line for the data in Example 1, using formulas (4.10) and (4.9). Compare your result with the computer result stated in Example 1.

4.3 Compute the LS line for Example 4 (Galton). Note that the slope is less than 1.

4.4 a) Compute the LS line for Exercise 2.5 (self-paced learning). Plot the line in the scatter diagram.

b) Now treat y as the independent variable, x as the dependent variable, and compute the LS line of the form $x = m'y + b'$. Plot this line also. Notice that it is not the line obtained in (a).

4.5 Show that the line $y = mx + b$ goes through the center of mass if and only if the sum of the deviations from the line is zero. That is, show that

$$\bar{y} = m\bar{x} + b \quad \text{if and only if} \quad \sum_{i=1}^{N} (y_i - mx_i - b) = 0.$$

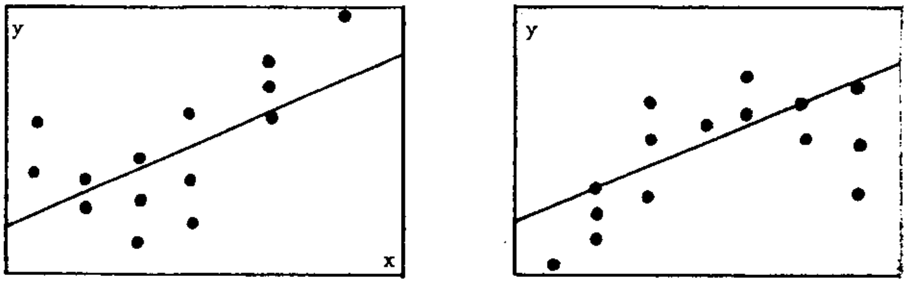
This is another reason why we should expect the LS line to pass through the center of mass.

5. HOW TO TELL IF YOUR LINE FITS WELL

5.1 Introduction

It is usually desirable to get some ideas as to how well your line fits the data. After all, you usually will not know for sure that the true curve is a line. If the least-squares line doesn't fit well, you might try to fit a parabola or some other curve instead. Also, you will probably like to know how close your estimated values of y are to the true values, for each x, especially if the line is to be used for prediction.

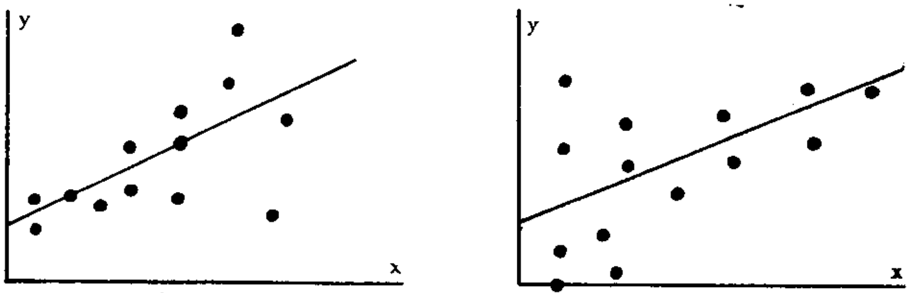
You can tell something about how well your equation fits just by looking at the curve on the scatter diagram. Do the data points tend to lie mostly above (or below) the fitted line at the ends, but below (respectively, above) it in the middle? Then the regression curve is probably not a straight line. (See Fig. 10.) Do the data points show more scatter toward one end of the fitted line than toward the other? Then the estimated values of y will be unreliable at that end. (See Fig. 11.) Does one data point seem to stand out, lying much farther from the line than the others? Check it out! If it is an error, refigure the LS line without it. If not, it might be worth fame and fortune to discover its cause!



upward curvature

downward curvature

Figure 10. Regression curve not a straight line.



variance increases with x

variance decreases with x

Figure 11. Variance in y depends on x.

Various tests of goodness of fit are given in statistics books, and confidence regions can be given around the LS line such that the true value of y lies within the region, say, 90% of the time. These subjects are beyond the scope of this unit. Let's just say that the line fits well if it accounts for most of the scatter in y, that is, if the scatter of the data points about the line is small compared to the variance in y.

5.2 Scatter About the LS Line

The scatter about the LS line is the mean squared deviation:

(5.1)
$$\frac{\hat{S}}{N} = \frac{\sum_{i=1}^{N} (y_i - mx_i - b)^2}{N}.$$

The square root of $\hat{S}/N$ is called the standard error of the estimate. It shows how far a typical observed value of y is from the line. To compute $\hat{S}$ in the case of a line through the origin, you would use Eq. (4.3). In the case of a general line, you have to make the change of variables $x_i' = x_i - \bar{x}$ and $y_i' = y_i - \bar{y}$ in (4.3). The result is

$$\hat{S} = \sum_{i=1}^{N} (y_i - \bar{y})^2 - \hat{m}^2 \sum_{i=1}^{N} (x_i - \bar{x})^2 .$$

Dividing by N and recalling the definition of variance (2.2), we get

$$(5.2) \qquad \frac{\hat{S}}{N} = \mathrm{var}(y) - \hat{m}^2 \, \mathrm{var}(x).$$

5.3 The Correlation Coefficient Squared

The scatter "accounted for" or "explained by" the line is the difference between the total scatter, var(y), and the scatter about the line, $\hat{S}/N$. From (5.2), we see this is just $\hat{m}^2$ var(x). The fraction of the total scatter that is accounted for by the line is called the correlation coefficient squared, denoted by r^2.

$$(5.3) \qquad r^2 = \frac{\text{scatter accounted for by line}}{\text{total scatter}} = \frac{\hat{m}^2 \, \mathrm{var}(x)}{\mathrm{var}(y)} .$$

As a rule of thumb, r^2 should be at least 80% for an acceptable fit.

Substituting the expression for $\hat{m}$ from (4.10) into (5.3) and using (2.3) and (2.5) for the variances, we get a formula for r^2 that is useful for computations:

$$(5.4) \qquad r^2 = \frac{\left(\sum_{i=1}^{N} x_i y_i - N\overline{xy}\right)^2}{\left(\sum_{i=1}^{N} \bar{x}_i^2 - Nx^2\right)\left(\sum_{i=1}^{N} y_i^2 - N\bar{y}^2\right)} .$$

Figure 12 shows scatter diagrams with values of r^2 ranging from 0 to 95%. The value of r^2 will be 100% only if all the data points lie on a straight line.

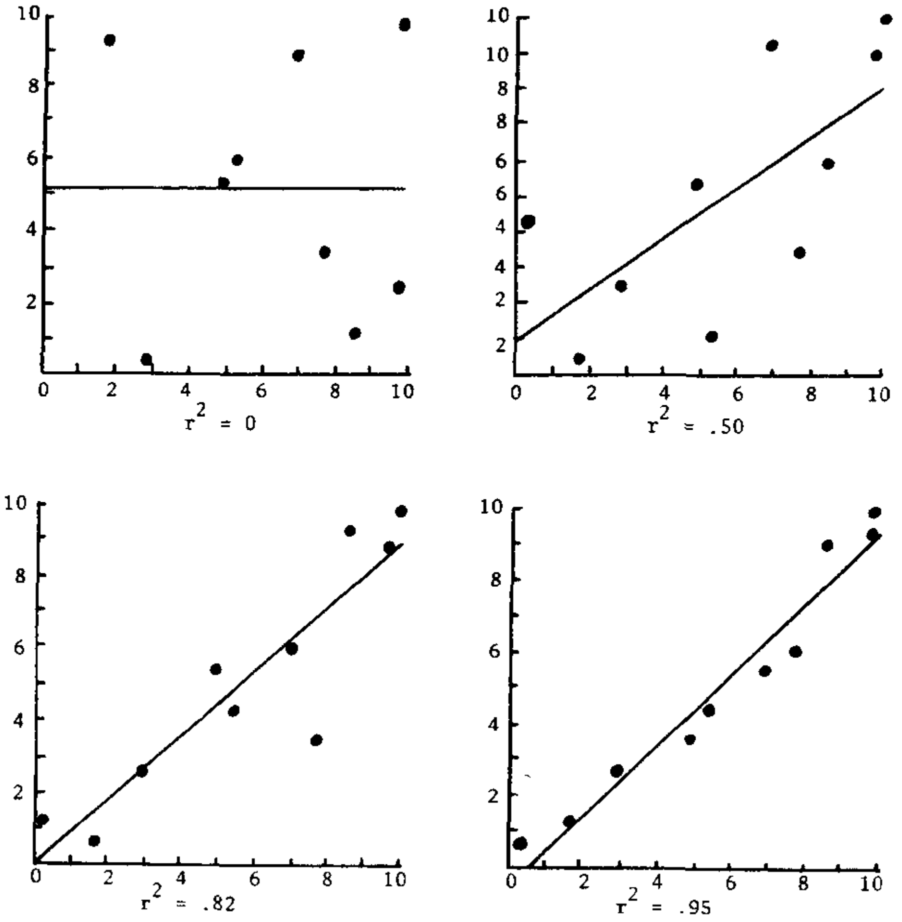
5.4 Example 5 Continued (Uptake of Nitrogen)

Continuing where we left off in Section 4.5, we find

$$\mathrm{var}(x) = \frac{\Sigma x_i^2}{N} - \bar{x}^2 = \frac{3.85}{11} - (0.5)^2 = 0.1,$$

and

$$\mathrm{var}(y) = \frac{\Sigma y_i^2}{N} - \bar{y}^2 = \frac{2.65}{11} - (0.4464)^2 = 0.0416.$$



$r^2 = 0$

$r^2 = .50$

$r^2 = .82$

$r^2 = .95$

Figure 12. Correlation coefficient squared.

The scatter about the LS line is

$$\frac{\hat{S}}{N} = \text{var}(y) - \hat{m}^2 \text{ var}(x) = 0.0416 - (0.614)^2(0.1) = .0039.$$

The standard error of the estimate is $\sqrt{.0039} = 0.06$. Most of the observed values y_i lie within 0.06 units of the fitted values $\hat{m}x_i + \hat{b}$. By (5.3), the correlation coefficient squared is

$$r^2 = \frac{\hat{m}^2 \text{var}(x)}{\text{var}(y)} = \frac{(0.614)^2(0.1)}{(0.0416)} = 0.91,$$

or, by (5.4),

$$r^2 = \frac{[3.13 - (11)(0.5)(0.4464)]^2}{[3.85-(11)(0.5)^2][2.65-(11)(0.4464)^2]}$$

$$= \frac{0.455}{(1.1)(0.458)} = 0.90.$$

This means that about 90% of the total variation in y is accounted for by the line. The least-square line gives a good fit to the data.

Exercises

5.1 Compute r^2 for Example 1 (organic acid content), and compare your result with the values claimed there.

5.2 a) Compute the LS line for Example 3 (response of sea anemones). Plot the line on the scatter diagram. Does it seem to fit well? Why not?

b) Compute r^2 for this line. Does the value of r^2 reflect the lack of fit?

5.3 Compute $\hat{S}/N$ and r^2 for the experiment in Exercise 2.4 (growth of heifers). Does the size of r^2 seem to reflect the magnitude of the scatter?

5.4 a) Compute r^2 for Exercise 2.5 (self-paced learning).

b) In Exercise 4.4, you computed two LS lines for the self-paced learning data: $y = \hat{m}x + \hat{b}$ and $x = \hat{m}'y + \hat{b}'$. Verify that $\hat{m}\hat{m}' = r^2$. (This is true in general.)

---

6. REFERENCES

1. Daniel, C.D., and Wood, F.S., Fitting Equations to Data. Wiley Interscience, 1971.

2. Lindgren, B.W., Statistical Theory, Macmillan, 1962, pp. 443-457.

3. Mosteller, F., Rourke, R., and Thomas, G., Probability with Statistical Applications, Addison-Wesley, 1961, pp. 360-400.

7. APPENDIX

The following Lemma was stated in Section 3 and used in Sections 3 and 4.

LEMMA. A quadratic function of the form $Y = AX^2 + BX + C$, with positive squared term $(A > 0)$, attains its minimum value when $X = -B/(2A)$. The minimum value of the function is $Y = C - B^2/(4A)$.

The proof is by completing the square:

Start with:	$Y = AX^2 + BX + C$
Divide by A:	$Y/A = X^2 + (B/A)X + C/A$
Subtract C/A:	$Y/A - C/A = X^2 + (B/A)X$
Add half the coefficient of X, squared:	$Y/A - C/A + (B/2A)^2 = X^2 + (B/A)X + (B/2A)^2$ $Y/A - C/A + B^2/4A^2 = (X + B/2A)^2$
Isolate Y/A:	$Y/A = C/A - B^2/4A^2 + (X + B/2A)^2$
Multiply by A:	$Y = C - B^2/4A + A(X + B/2A)^2$

Since A is positive, the last term of the right is never negative. Its smallest value is zero, which occurs when $X = -B/2A$. The corresponding value of Y is $C - B^2/4A$. This completes the proof.

8. ANSWERS TO EXERCISES

Section 1

1.1

acid no., x	acid content, y	estimated acid content, y = .3216 x + 35.45
62	55	55.4
100	66	67.6
159	88	86.6

Sources of error include measurement error in x and day-to-day variations in the content of the chemical samples affecting x and y differently (uncontrolled variables).

1.2 The estimated value of m is 1.51. The estimated value of β is 83.

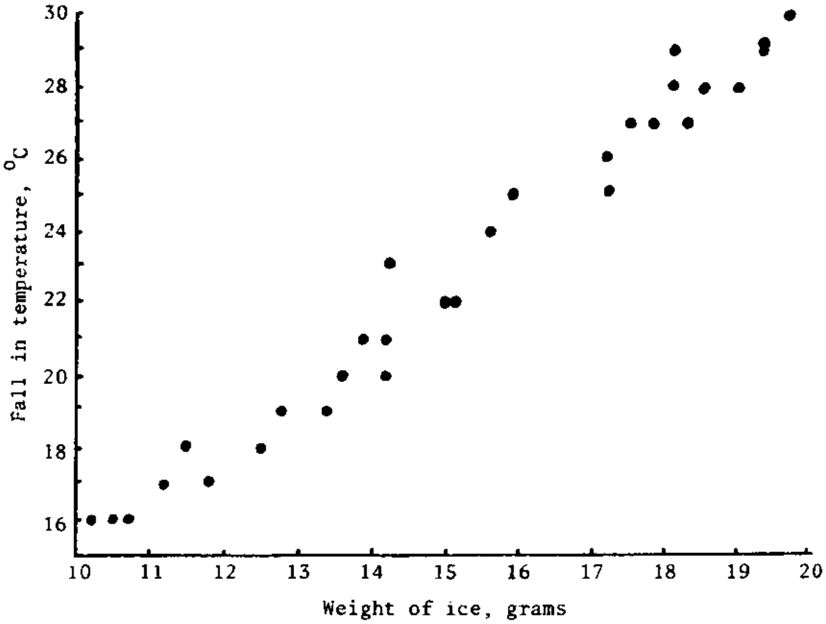
1.3 Measurement error in x; error in y due to uncertainty in knowing when a response terminates; sampling error; individual differences among anemones.

1.4 67.2, 69.2. Yes, because these averages y's are closer to the overall population mean than are the x's.

1.5 See page 5.

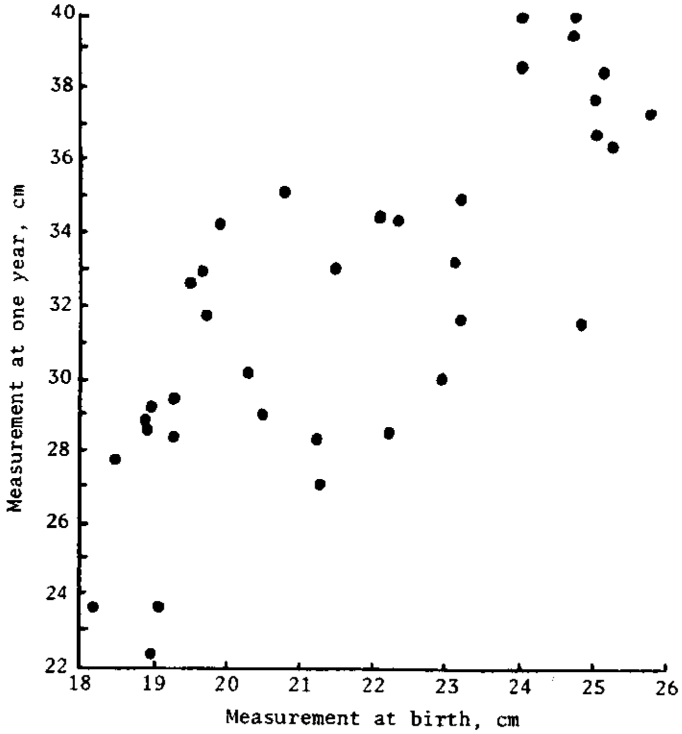
Section 2

2.2

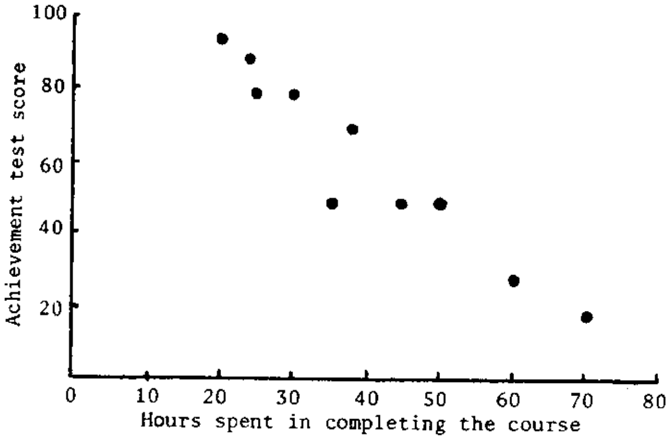


2.3 Log function, power function, and others.

2.4



2.5



2.6 $\overline{y} = 68.6$, $\text{var}(y) = 255$

2.7 $\overline{y} = 32.1$, $\text{var}(y) = 21.7$

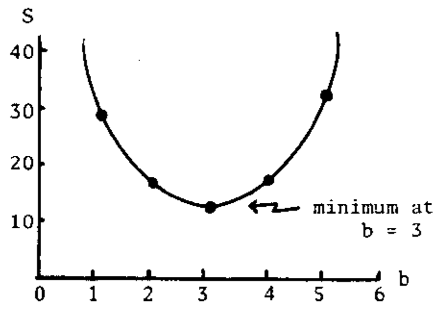
2.8 $\overline{x} = 3$, $\text{var}(x) = 2$

2.9 $\overline{x} = 39.7$, $\text{var}(x) = 243$, $\overline{y} = 60.5$, $\text{var}(y) = 627$

Section 3

3.1

b	S
1	32.5
2	17.5
3	12.5
4	17.5
5	32.5



3.2 $\hat{b} = \overline{y} = 148.1$, $\text{var}(y) = 9.12$

3.3 $S = 0.4$

3.4 (a) $S = 16.6$

Section 4

4.1 $y = 0.0059\ x$

4.2 $y = 0.3216\ x + 35.46$

4.3 $y = 0.548\ x + 30.9$

4.4 (a) $y = -1.54x + 121.6$
(b) $x = -0.508\ y + 75.9$, or $y = -1.67\ x + 126.9$

Section 5

5.1 $r^2 = 0.995$

5.2 (a) $y = 0.56\ x + 4.6$ (b) $r^2 = 0.82$

Although the value of r^2 indicates a possibly acceptable fit, the LS line does not fit well due to the curvature of the data points about it.

5.3 $\hat{S}/N = 8.0$, $r^2 = 0.63$

5.4 (a) $r^2 = 0.92$
(b) $mm' = (-1.54)(-0.598) = 0.92$

9. SAMPLE TEST

1. State four purposes for fitting equations to data, and describe an example illustrating each. (Objective 1)

2. What is experimental error? (Objective 2)

3. What is meant by the least-squares fit of a given type of function $y = f(x)$ to data points (x_i, y_i), $i = 1, 2, 3, \ldots, N$? (Objective 6)

4. ______________is a measure of the scatter of the data points about the fitted curve. (Objective 7)

Consider the following experiment:

Raw material used in the production of a synthetic fiber is stored in a place without humidity control. Measurements of the relative humidity in the storage room and the moisture content by weight of a sample of the material are taken on 10 days selected at random over a six week period. The results are shown in Table A.

TABLE A

Moisture Content of Raw Material

Relative Humidity, x (percent)	46	30	34	52	38	44	40	45	34	60
Moisture Content, y (percent)	10	7	9	13	8	12	11	11	7	14

The managers want to estimate the moisture content of the fiber from the humidity of the room.

5. Make a scatter diagram using this data. (Objective 3)

6. Sketch a free-hand line or curve through the data points. (Objective 4)

7. a. Compute the mean and variance in y for Table A.
   b. What does variance measure? (Objective 5)

8. Compute the least-squares line through the data points in the experiment. (Objective 8)

9. a. Compute r^2, the correlation coefficient squared, for this data.
   b. What does r^2 measure? (Objective 9)

10. Does the LS line fit the data well? Discuss. (Objective 10)

10. ANSWERS TO SAMPLE TEST

1. See page 5.

2. (pp. 6-7) Experimental error is the difference between the observed value of a variable and its hypothetical "true" value.

3. The least-squares fit of a given type of function $y = f(x)$ to data points (x_i, y_i) is the particular function of that type that minimizes the sum of the squares of the derivations. (See p. 15.)

   $$\text{Least-squares fit} \Leftrightarrow \text{minimize } S = \Sigma[y_i - f(x_i)]^2$$

4. The mean squared deviation, $S/N = \Sigma[y_i - f(x_i)]^2/N$. (See pp. 15-30.)

5. See Figure A.

6. See Figure A. Your line should lie between the dashed curves.

7. a. The sums are computed in Table B below.

   The mean y is $\bar{y} = \Sigma y_i/N = 102/10 = 10.2$.

   The variance in y is

   $$\text{var}(y) = \frac{(\Sigma y_i^2 - N\bar{y}^2)}{N} = \frac{53.4}{10} = 5.34.$$

   b. The variance measures the scatter about the mean or the total scatter, assuming no dependence on x. (See pp. 8-11.)

8. Taking the sums from Table B, we get

   $$\hat{m} = \frac{\Sigma x_i y_i - N\overline{xy}}{\Sigma x_i^2 - N\bar{x}^2} = \frac{182.3}{744.1} = 0.245,$$

   $$\hat{b} = \bar{y} - \hat{m}\bar{x} - 10.2 - (0.245)(42.3) = 0.169.$$

   The LS line is $y = 0.25\ x - 0.17$. This is shown in Figure A. (See pp 26-28.)

9. a. From Table B,

   $$r^2 = \frac{(\Sigma x_i y_i - N\overline{xy})^2}{(\Sigma x_i^2 - N\bar{x}^2)(\overline{\Sigma y}_i^2 - n\bar{y}^2)} = \frac{(182.4)^2}{(744.1)(53.6)} = 0.834,$$

   or about 83%.

   b. r^2 measures the fraction of the total variance in y that is accounted for by the curve. (See pp 31-32.)

10. The scatter diagram does not show significant curvature, and the scatter does not seem to depend on x. There are no "outliers",

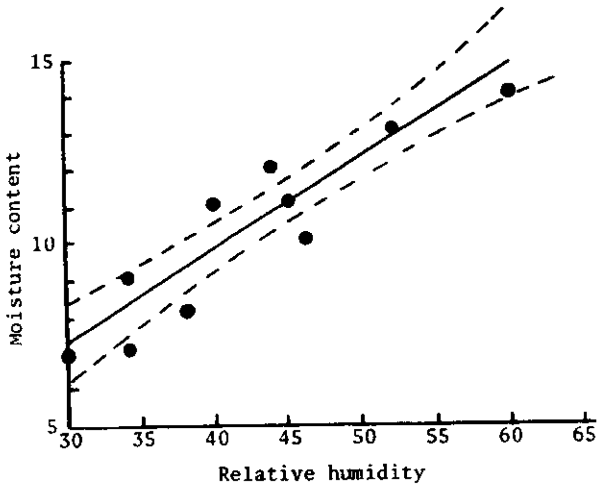


Figure A. Storage of synthetic material.

or points that lie exceptionally far from the line. Also, the value of r^2 is acceptable. We conclude that the regression curve is probably a line.

However, there aren't many data points (N = 10 is small), and the scatter about the LS line is rather large. The standard error of the estimate is 0.95, or about 10% of $\overline{y}$. Thus, the LS line does not give a precise estimate of y. Whether the managers can use it depends on how good an estimate they need.

We suspect that one reason for the large scatter about the LS line is that the moisture content of the material stored depends on other factors besides the current relative humidity. For example, there may be a time lag between changes in the humidity and changes in the moisture content of the material. More careful experimenting would be necessary to determine this. (See p. 29.)

TABLE B

Computations

	x_i	y_i	x_i^2	x_iy_i	y_i^2
	46	10	2116	460	100
	30	7.0	900	210	49
	34	9.0	1156	306	81
	52	13	2704	676	169
	38	8.0	1444	304	64
	44	12	1936	528	144
	40	11	1600	440	121
	45	11	2025	495	121
	34	7.0	1156	238	49
	60	14	3600	840	196
TOTALS	423	102	18637	4497	1094

$\Sigma x_i^2 - N\overline{x}^2 = 18637 - (10)(42.3)^2 = 744.1$

$\Sigma y_i^2 - N\bar{y}^2 = 1094 - (10)(10.2)^2 = 53.6$

$\Sigma x_i y_i - N\bar{x}\bar{y} = 4497 - (10)(42.3)(10.2) = 182.4.$