Plots of relationship 5

We have learned something about scratching down batches--sets of similar numbers. We found more things to do--saw that more things could be learned about data by doing them--than we might have expected. We turn now to the use of simple plots, to the plotting of y against x, and find, again, more things to do and more gains from doing them than we might expect.

The most quoted passage of all the Sherlock Holmes stories concerns the unusual behavior of the dog in the night time. Holmes called attention to its unusualness--Watson couldn't see what was unusual--Holmes pointed out that the dog did nothing.

The moral is clear--we ought to judge each occurrence against the background of--or a background derived from--other "nearby" occurrences. We do not ordinarily think of "zero" as unusual--yet a 24-hour period when no one died anywhere in the world, or a winter with no snow at all in the White Mountains, would be recognized by each of us as strikingly unusual.

As in the large, so in the small. Once we have plotted some data, and found out how it behaves generally, our next step is to look at each of its elements--at each "point"--and ask how it seems to deviate from the general behavior of all the "points". To do this it very often helps to make a new plot, one that focuses on such deviations--in brief, "it very often pays to plot residuals".

As we have begun, so will we continue. Here, as well as later, we will be repeatedly engaged in splitting up data according to one version or another of the key relation

given = fit PLUS residual.

Here the fit is our current description--always incomplete, always approximate--of the overall behavior of the data. Each individual observation is split up into a sum of this fit and what is left over, called

a residual.

Residuals are our main tool in going further. They are to the data analyst what powerful magnifying glasses, sensitive chemical tests for bloodstains, and delicate listening devices are to a story-book detective. They permeate all sorts of data analysis and appear in many guises.

The most conventional plots follow a simple pattern. A quantity, habitually shown on a vertical scale, which we think of as

a response.

Another quantity, habitually shown on a horizontal scale, which we call

a factor

which is usually

a circumstance

and which we think of as possibly explanatory or descriptive. The data--in whole or part--will consist of pairs of numbers of the form

(factor, response)

which will be plotted as points using these scales.

Most plots--at least in the popular press--concentrate on the fit. Too often, however, such plots are only to remind us that the anticipated relationship is there--that, for example, the population of the U.S.A. still increases. As such they are usually matters of elementary exposition rather than analyses. They show us "the big picture" that we already knew about. (In the sense of Linus in the familiar comic strip, they are just "security blankets".) At other times, such plots "in the large" are to show us the unexpected--either an unsuspected relationship or an unsuspected strength or weakness of an anticipated relationship. For these purposes, plots "in the large" are part of data analysis. They tell us of investigation's successes, perhaps very effectively, but they still need to be supplemented by pictures of residuals--pictures that tell us whether there is yet more to investigate.

For us the most useful plot will be one that might reveal the unexpected or the unobvious. Sometimes a plot "in the large" will do this. Usually, however, it is the plot of residuals that has the greatest use and the greatest impact.

review questions

How ought we judge occurrences or numbers? What is a "residual"? What is the key relation involving residuals? What is a fit? Is it final, complete, or exact? What is a "response"? A "factor"? A "circumstance"? How are they usually plotted? How is data involving one factor and one response usually written down? Plotted? When are plots "in the large" useful? When are they part of data analysis? How do we ask whether there is more to investigate? What kind of plot is most likely to help us?

5A. How to plot *y* against *x*

Plots are important. Great differences in ease of construction and great differences in effectiveness of use depend on apparently minor, purely technical procedures. The remarks in this section are intended to finish getting you over most of the elementary hurdles. NOW GO BACK and read the parts of section 2C, dealing with "tracing paper", "scale values," and "plotting without graph paper". All this still applies here.

choice of ruling

If plotting is to be easy, you want graph paper with at least three different thicknesses of ruling:

- ◇ light lines for "units".
- ◇ medium lines for "fives".
- ◇ heavy lines for "tens".

A wide variety of graph papers can be obtained with these characteristics. (Some papers have extra heavy lines for "twenties". You will have to learn for yourself whether these help you more than they get in your way.)

DO NOT use papers ruled in "fours" and "eights", or in "sixes" and "twelves" for plotting data given in ordinary (decimal) numbers. (For monthly data, of course, "twelves" in one direction do help.)

If you want to plot fast, easily, and accurately, avoid "dime-store" sheets with only two thicknesses of ruling and, above all, avoid quadrille sheets with only one. (Quadrille sheets are very useful, for almost everything except plotting graphs.) (If you need to save money, see below.)

choice of scale units

When you come to plot, you must choose units. Don't try to make one step of ruling (light, medium, or heavy) equal to 3 units, 7 units, 0.03 units, 0.007 units, or any such uncomfortable number. Stick to one step being 1, 2, or 5 times a power of ten. (One square = 20,000 or one step = 0.05 are examples that can work out quite well.)

You will find it hard enough to learn to be a fast and accurate plotter with two three-speed gear shifts (1, 2, or 5 across the graph and 1, 2, or 5 up the graph).

If you need to use an abnormal scale--three units to the square, for example--convert your numbers into plotting units--as by dividing by 3--by slide rule, hand arithmetic, or what have you, BEFORE you start to plot. It will take less time, overall, to make a good picture.

consequences of our purpose

Throughout this account we shall be interested in graphs to be looked at, not to be used to find numbers. Our graphs are means for looking at the data, not stores of quantitative information. This means:

- ◇ we will want to keep our eyes on the points.
- ◇ we will usually **not** connect one point to the "next" one. (We are likely to draw in fitted lines or curves.)
- ◇ we will want to suppress the rulings of graph paper, at least from our mind's eye.

◇ we will need only a few numbered ticks along the axes, horizontal or vertical, of the final picture.

◇ we will want to use symbols large enough to stand out (and if we need more than one kind, different kinds should be both clearly different and--usually--almost equally noticeable).

There cannot be too much emphasis on our need to see behavior.

We must play down or eliminate anything that might get in the way of our seeing what appears to go on.

kinds of grids

The use of tracing paper makes it possible for any of us to really arm himself with an easy-graphing capability of broad scope. A tracing pad and one sheet of each of 20 kinds of graph paper puts us in business to do many things. (Actually, it usually seems worthwhile to have a reasonable number of sheets, perhaps in the form of a pad, of at least one or two of the most used kinds of graph paper. Many will wish to first plot on the graph paper and then copy on the tracing paper.)

"Ordinary" graph paper frequently comes 8 small squares to the inch (beware; it may be lined in "4's" or "8's" which you should never, never use--unless what you are plotting comes in eighths of an inch, or, like stock market prices, in eighths of a dollar), 10 to the inch, 12 to the inch (beware, it may be ruled in "6's" or "12's"), 20 to the inch, and 10 to the centimeter. The writer likes the coarser rulings at least as well as the fine; it is up to you to learn your own preference.

Beware of 12-to-the-inch rulings, since too many are *ruled* in "sixes" and "twelves"; but bear in mind that 12-to-the-inch ruled in fives and tens--like Codex Nos. 31,253 and 32,253, for example--is a very good base for graphs where the final product is a typed page. (Typewriters like 12 to the inch.) It has little special advantage for most other purposes, although it is quite satisfactory.

Semilogarithmic (uniform scale one way, log scale the other) and full logarithmic (log scales both ways) graph papers come in a variety of scales and patterns. To have a reasonable selection is to save time and encourage inquiry. One CAN always look up the logs and then plot on uniform-scale graph paper, but WILL one? If chapter 3 is at hand, probably yes. (We rarely get much more than two-decimal accuracy out of a plot on log paper, so exhibit 2 of chapter 3 will usually do as well as special graph paper.) Otherwise? As we saw above, logs are often the way to make data reveal itself.

A variety of other graph papers are useful, as we shall see later. We note here that paper with two square-root scales is available (Codex No. 31,298 or No. 32,298) and that trilinear or isometric paper--with three sets of rulings at 120° to one another--is often quite useful. (It is available from various

manufacturers.) Reciprocal (one-way) paper is available both from Codex and from Keuffel and Esser.

shape of plot

Naturally enough, graph paper usually comes in the same general shape and size (letter size) as ruled writing pads and typing paper. Clearly it can be used in two ways:

- narrow edge at the top of the plot, or
- broad edge at the top.

By analogy with the use of ruled writing pads, it is natural to make plots with the narrow edge up. Full use of the area then usually makes a plot that is taller than it is wide.

There are some purposes for which this is a good shape. The ever-faster-rising curve of early growth is an example.We can quite well use taller-than-wide plots for such simple pictures, most of which tell us that we haven't yet made an analytically useful plot. Sometimes, indeed, such simple plots can be clearer when they are taller than wide, rather than wider than tall.

Most diagnostic plots involve either a more or less definite dependence that bobbles around a lot, or a point spatter. Such plots are rather more often better made *wider* than tall. Wider-than-tall shapes usually make it easier for the eye to follow from left to right.

Perhaps the most general guidance we can offer is that smoothly-changing curves can stand being taller than wide, but a wiggly curve needs to be wider than tall (sometimes after a smooth part has been taken out).

When a plot is made wider than tall, convention says it should be turned in the direction illustrated by many plots in this chapter, even if this makes a down-to-up legend upside down when we first see the plot as we turn the pages.

ticks and numbers along axes

We use marks along axes for two quite different purposes: (1) to plot the points; (2) to look at them. Different purposes call for different techniques, and the graph-paper–tracing-paper combination makes separate techniques easy. If one begins by plotting on the graph paper, it helps to have many numbers and ticks on the scale. (Remember to have the horizontal scale **above** the plot, and the vertical scale to the left*--both out from under the plotting hand.) To look effectively at the traced result, it helps to have ONLY a few numbers and ticks. Exhibit **1** shows five pairs of vertical axes, one in each pair for plotting and one for looking. Notice the technique, in each PLOTTING scale, of:

- putting 4 ticks--or dots--between consecutive numbers (often useful: NOT to be religiously followed; sometimes no tick between, or one tick between, is better).

* To the right, for southpaws.

◇ putting a tick--or dot--for every one or two steps of some digit in what is plotted (sometimes it pays to do this at every five, instead).

(Some even standardize on dots for each one and ticks for each two.) All this is focused on making it as easy as possible to find where to put the point. Doing less than this slows down our plotting and wastes both effort and temper.

Notice also the technique, in each LOOKING scale, of using:

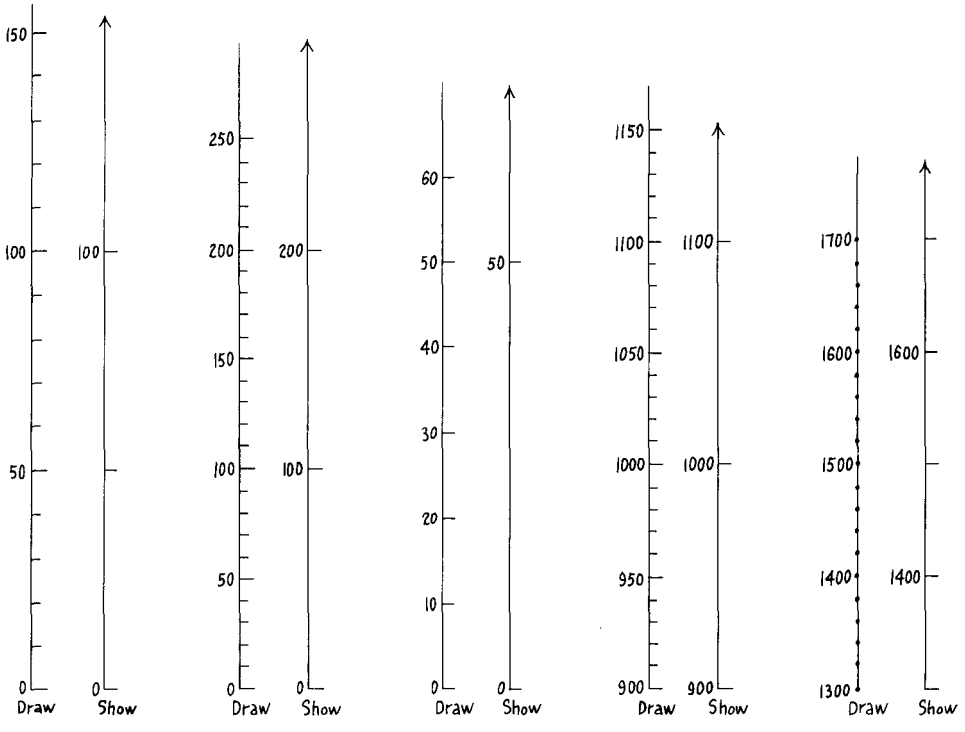
◇ only four or five ticks.

◇ only two or three numbers.

Doing more than this distracts our attention from what we ought to see. (If the scale is irregular, we may need more ticks and values. Scales for dates, where individual values like 1066, 1776, or 1929 are well remembered, often deserve more ticks and values.)

exhibit **1** of chapter 5; illustrative

Pairs of scales, showing the difference between scales for plotting (marked "draw") and for looking (marked "show")



Clearly, separating plotting from looking can be a great help. Three cheers for tracing paper or transparent plastic.

review questions

What is the minimum for well-ruled graph paper? What is optional? What choices of scale units are reasonable? What should we do if we need an unreasonable choice? Should we connect our points with lines or curves? Do graph-paper grids help us to plot points? To see what our points look like? How many ticks and numbers for looking? For plotting? Where should we put the scales for plotting? How do we use tracing paper? What are some conveniently available kinds of graph paper? Who should use graph paper with heavy 4's and 8's? With heavy 6's and 12's? What shapes of plot are desirable? Does this book follow the rule? How can we plot without graph paper?

5B. Looking at subtraction

Undoubtedly, the form of graphical representation we have been most exposed to in school is one involving two variables called x and y, in which y is said to **depend on** x.

In data analysis, a plot of y against x may help us when we know nothing about the logical connection from x to y--even when we do not know whether or not there is one--even when we know that such a connection is impossible.

Before we can make full use of such plots, we need to understand--in terms both of doing and of feeling--certain things about such a plot, including:

- ◇ how to subtract one "curve" from another.
- ◇ how to find a numerical formula for a straight line drawn on graph paper.
- ◇ what effect subtracting different--two or more--straight lines from the same data points has.
- ◇ how to try to re-express either y or x, or both, so as to make the data appear more nearly straight.
- ◇ why graph paper--on which we plot easily--expresses the essence of how points represent numbers much better than the kind of picture we will usually want to look at--which has axes with ticks and numbers.

When we have x alone, subtraction is simply and easily represented by sliding arrows along. Subtracting 3 from 5 is a matter of drawing an arrow starting at +3 and extending to +5, and then sliding this arrow until it starts at 0. Its new endpoint, +2, is the result of subtraction. Exhibit **2** shows this example and three others that involve various combinations of minus signs.

In dealing with data we usually deal with subtraction of y's rather than of x's. For a hypothetical ABC Corporation, in 1960, sales were 44 million and

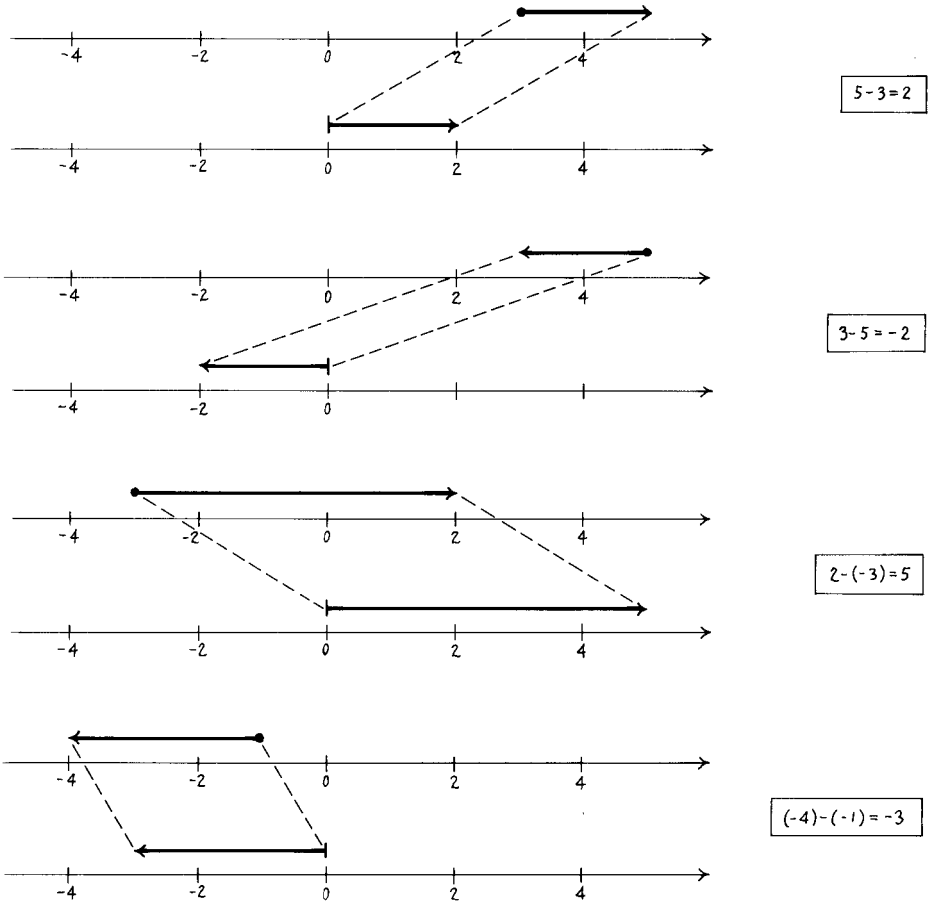
expenses were 32 million. We can easily subtract expenses from sales to find profit before taxes of 12 million, as shown in exhibit **3**. Again, we are sliding the arrow to put its base at zero. This time we have to be careful to keep it at 1960.

In this exhibit and the next--for clarity of what we are doing, rather than for clarity of result--we have used two time scales placed side by side. Ordinarily we would use one time scale and slide each arrow along a single vertical line--as for the broken arrows on the left sides of these exhibits.

Exhibit **4** retains the pattern and shows sales, expenses, and profits for 12 consecutive years. To avoid confusing detail, the arrows and their sliding are shown for only three years, 1951, 1957, and 1960.

exhibit **2** of chapter 5: illustrative

Four examples of subtraction



The ABC Corporation in 1960

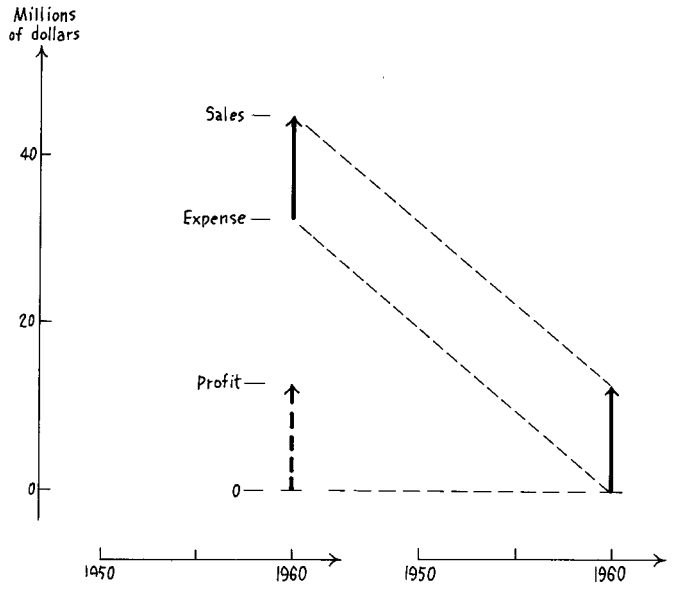
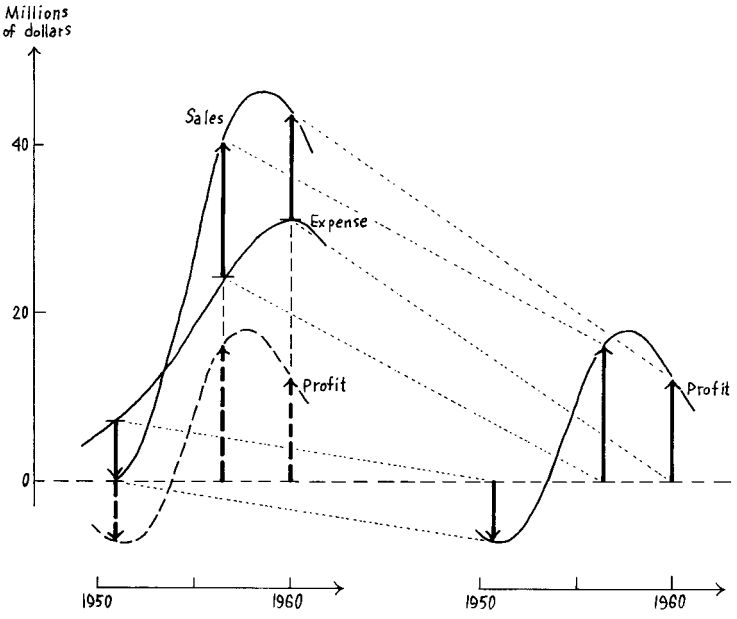


exhibit **4** of chapter 5: ABC Corp.

Twelve years of the ABC Corporation



The case where we will be most concerned with graphical subtraction is the case where we solve the basic relation

data = incomplete description PLUS residual

= fit PLUS residual

for the residual, finding

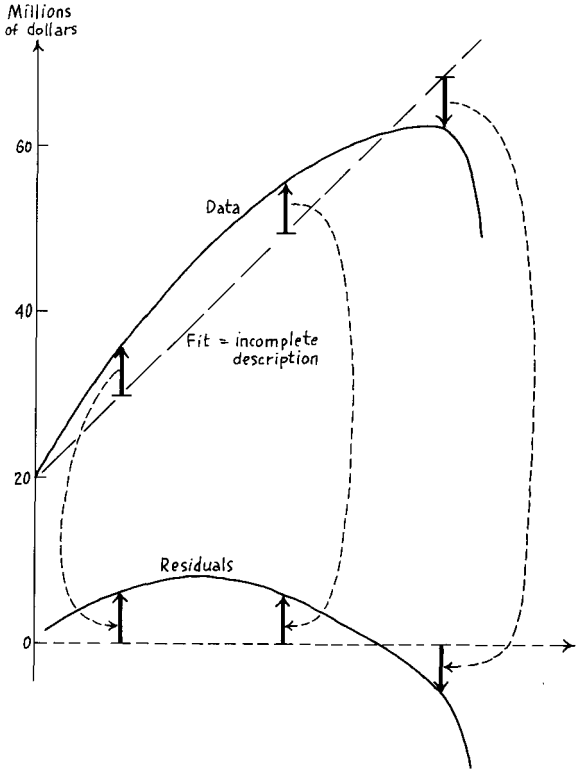
residual = data MINUS incomplete description.

untilting

Exhibit **5** shows an example of this where the incomplete description is a straight line. Here we have slid the vertical arrows down (or up) the vertical on which they lie. Again we have shown only three arrows of the many possible.

exhibit **5** of chapter 5: illustrative

Subtraction of an incomplete description from data to form residuals



It is natural to think--and speak--of such a subtraction of a straight-line partial description as an *untilting*. Natural and useful. Yet such use of words seems likely to distort our thinking a little if we are not wary. Untilting suggests, at least to some of us, a rigid motion in which the final curve comes from the initial curve by a rigid motion--a rotation about some point. This is quite wrong.

There are many ways to see this. An easy one is to look at the length of the curve between the two crossings of the straight line. This distance is obviously greater for "data" than for "residuals". Thus we do not have rigid motion.

A convenient picture involves a deck of cards, on whose edges--the edges on one side--we have drawn both "data" and "partial description". Now let us:

- ◇ clamp the deck together and saw off its bottom edge on a slant--a slant that is parallel to, because it is a constant vertical distance from, the partial description.
- ◇ unclamp the deck and strike it on a table to line up its new bottom edges horizontally.
- ◇ clamp the deck again.

The marks for "partial description" will now lie along a horizontal line, since they are a constant distance from the new bottom edges. If we call this upper line zero, the marks for "data" will now show us the "residuals".

The sliding of the cards with respect to one another exactly corresponds to sliding the various vertical arrows with respect to one another. This is a proper mechanical picture of graphical subtraction of one y from another. It works--rigid motion does not.

Of course, all this works for curved fits as well as for straight ones. (At least if we can make a curved cut in our deck of cards.)

review questions

Why would we want to subtract one curve from another? A curve from some point? How does subtracting a straight line behave? Does it have anything to do with rotation? What is a mechanical model for subtracting a straight line?

5C. Subtracting straight lines

If we have taken our data, plotted it, and drawn a straight line through it, and now we wish to use the straight line as an incomplete description, our next task is to subtract the line from the data. Sometimes, as in exhibit 4, we can do the subtraction graphically. Often, however, this is more work than we like.

finding the line

To do the subtraction by arithmetic, we have to turn the straight line into numbers. The easy way to do this is to choose two points on the line--let us call them (x_1, y_1) and (x_2, y_2)--read off their coordinates, and say that an equation for the line is

$$y = y_1 + b(x - x_1)$$

where the slope, b, of the line is given by

$$b = \frac{y_2 - y_1}{x_2 - x_1}$$

Clearly, where $x = x_1$, we have $x - x_1 = 0$ and the equation gives $y = y_1$. When $x = x_2$, the second term on the righthand side of the equation is

$$\frac{y_2 - y_1}{x_2 - x_1}(x_2 - x_1) = y_2 - y_1$$

and we have

$$y = y_1 + (y_2 - y_1) = y_2$$

as we should.

To make the arithmetic of using such an equation easy, we should choose x_1 so that the values of $x - x_1$ are as simple as possible. To make finding the equation easy, we should choose x_2 so that $x_2 - x_1$ is a simple number. (Of course, we have to balance the advantages of simplicity against the advantages of getting well out toward the end.) There need never be great difficulty in turning a straight line into an equation.

an example

Exhibit **6** shows a plot of the population of England and Wales at every decade from 1801 to 1931. An eye-fitted line has been drawn in and the necessary simple arithmetic--shown at the lower right--performed.

In carrying out the eye-fitting of a line and its conversion into numbers it is important:

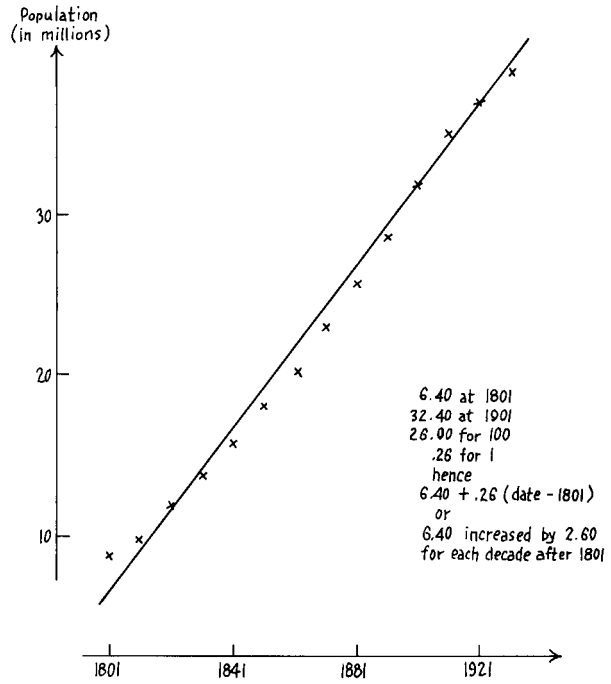
- ◇ to eye-fit the line on a picture without excessive background--use the tracing paper (or transparent plastic) version WITHOUT graph paper underneath.
- ◇ to find two points through which to pass the line, put the transparent version back on the graph paper, and locate the points with its aid.

It takes two different looks at a point-swarm to do all we want do.

Exhibit **7** shows the calculation of the residuals, not only from the line of exhibit 6 (line 1) but also from another line (line 2). (The rest of this exhibit will be discussed in a moment.)

exhibit **6** of chapter 5: England and Wales

Population in millions at successive censuses (1801–1931)



When, as in this example, the x-values step along in steps of constant size, one easy way to find the y-values is to begin at the low end and add the constant difference. For line 1, this means beginning with 6.40 and adding 2.60 repeatedly.

When the steps in x are not all the same (with the possible exception of short gaps), we have to do a little more arithmetic.

subtracting different straight lines

If we look at data, and draw what looks like a reasonable straight line through it, we are not likely to draw exactly the line that will serve us best. What are the consequences of a somewhat unsatisfactory line?

- what will it do to the residuals?
- how hard will it be to move to a better choice?

An easy way to approach both these questions is to ask what happens if we subtract first one straight line from y and then a second straight line from the

residuals thus formed. Algebra is easier than geometry here. If we begin by subtracting $a + bx$ we will form

$$y - (a + bx)$$

as our residual, where (x, y) is the data point and $a + bx$ is some fit or other. We have just seen examples of this (in exhibit 7). If, from the residuals thus formed, we subtract $A + Bx$, we will form

$$[y - (a + bx)] - (A + Bx) = y - [(a + A) + (b + B)x].$$

exhibit **7** of chapter 5: England and Wales

The population of England and Wales, with residuals from various lines (populations in millions)

A) DATA and CALCULATIONS

		line 1*		line 2*			supplementary line*	
year	pop'n	fit	resid	fit	resid		fit†	resid‡
1801	8.89	6.40	2.49	6.	2.89	s	1.73	1.16
11	10.16	9.00	1.16	8.5	1.66	t	1.64	.02
21	12.00	11.60	.40	11.	1.00	o	1.56	−.56
31	13.90	14.20	−.30	13.5	.40	p	1.48	−1.08
41	15.91	16.80	−.89	16.	−.09	h	1.39	−1.48
1851	17.93	19.40	−1.47	18.5	−.57	e	1.30	−1.87
61	20.07	22.00	−1.93	21.	−.93	r	1.22	−2.15
71	22.71	24.60	−1.89	23.5	−.79	e	1.14	−1.93
81	25.97	27.20	−1.23	26.	−.03	u	1.05	−1.08
91	29.00	29.80	−.80	28.5	.50	n	.96	−.46
1901	32.53	32.40	.13	31.	1.53	t	.88	.65
11	36.07	35.00	1.07	33.5	2.57	i	.80	1.77
21	37.89	37.60	.29	36.	1.89	l	.71	1.18
31	39.95	40.20	−.25	38.5	1.45	a	.62	.83
						f		
						t		
						e		
						r		
						e		
						x		
						h		
						8		

*** Line 1: 6.40 + .26 (date − 1801); Line 2: 6 + .25 (date − 1801).**
Supplementary line: 1.73 − .0085 (date − 1801).
† One more decimal kept in additions.
‡ These residuals are from the fit of the supplementary line to the residuals from line 2.

The result of two subtractions--of $a + bx$ and $A + Bx$--is always the same as the result of a single subtraction, of

$$(a + A) + (b + B)x$$

--of a subtraction of the

sum of the two lines.

Those who wish can show the same fact geometrically.

This fact about subtracting lines goes far toward answering our two questions:

◇ if we subtract one line and, on looking at the residuals, find them still tilted, we are free to draw a line among these residuals and subtract it further, thus finding new residuals. The new residuals correspond to subtracting a single line--the sum of the two actually subtracted. We need not go back and start again. This is particularly handy when the residuals are much smaller numbers than are the data.

◇ if we subtract an unsatisfactory line, and discover that we have done this by looking at the residuals, we could always correct this by a further subtraction. Accordingly, our first residuals will differ from the better residuals by being somewhat tilted. Since slight tilts do little to hide what we are looking at the residuals for--evidence of further structure or of unusual values--it will rarely be necessary to do the second subtraction (unless we want to publish the residuals). We can see what we need to see in the slightly tilted residuals. And, if we want to find the equation of the better line, we can draw a correction line and sum the expressions of the original and correction lines.

Subtracting lines is simple and convenient in many ways.

back to the example

Exhibit **8** shows the residuals from line 2 of exhibit 7, plotting them against date. One eye-fitted line is shown. (Clearly, there could be considerable debate about which line to fit to this sequence of points.) This is naturally called a

supplementary line

since it is fitted to the residuals from a first fit. The result is

$$1.73 - .0085(\text{year} - 1801).$$

Since the residuals fitted came from the line

$$6 + .25(\text{year} - 1801)$$

the 2nd residuals--visible in exhibit 8 as deviations from the line, given

numerically in the righthand column of exhibit 7--are residuals from the sum of these two lines, namely

$$(1.73 + 6) + (-.0085 + .25)(\text{year} - 1801)$$

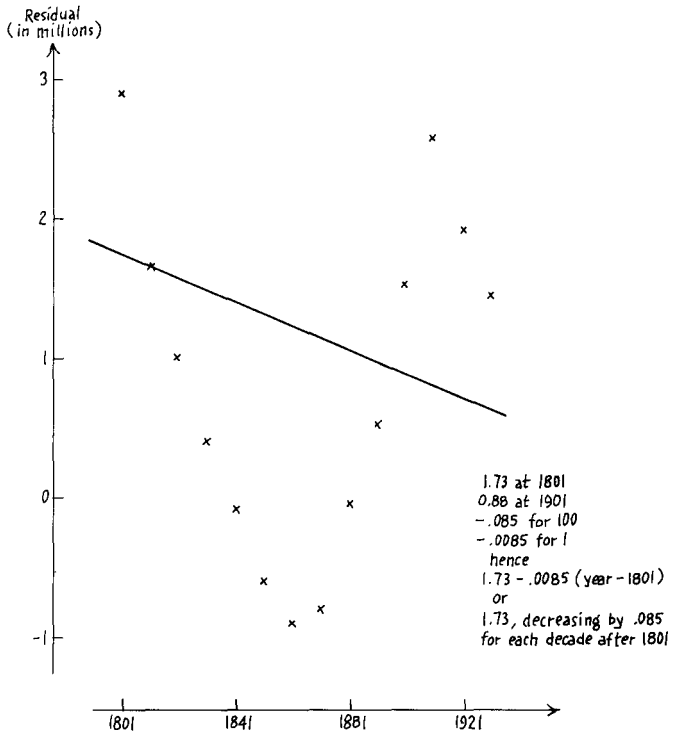
or, simplified

$$7.73 + .2415(\text{year} - 1801).$$

Taking out first a preliminary line and then a supplementary one often, as in this example, helps us keep our (hand-done) arithmetic simpler. If we use "easy numbers" in the preliminary fit, we can often do the two fittings for less than the price of one. In addition--in fact a more important consideration--we get a look at a picture of some residuals, something that can have a variety of advantages, and often does.

exhibit **8** of chapter 5: England and Wales

The residuals--from line 2 of exhibit 7--plotted against date and eye-fitted with a line.



review questions

How does one fit a straight line to two points? How does this help us in eye-fitting a line? What did we choose as an example? What happens if we subtract two lines, one after the other, from either some points or a curve? Can this make our arithmetic simpler? Why?

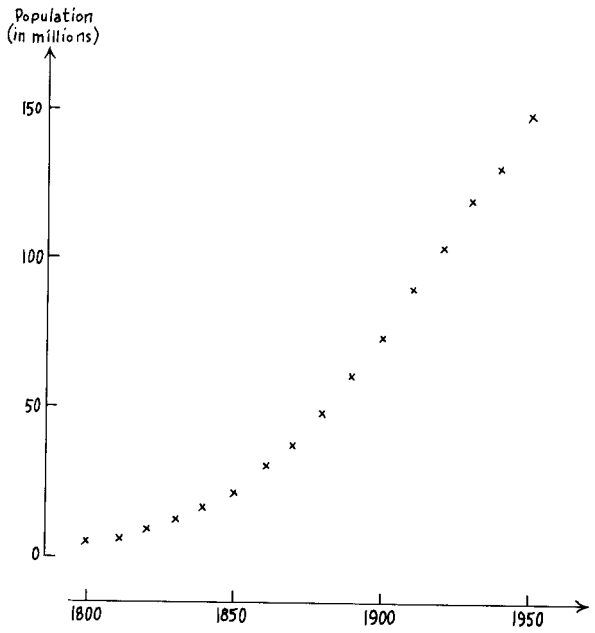
5D. Plotting the population of the U.S.A.

Many people would think that plotting y against x for simple data is something requiring little thought. If one only wishes to learn but little, it is true that a very little thought is enough, but if we want to learn more, we must think more.

A convenient example of how thinking more lets us learn more is given by the population of the U.S.A. as counted by the census every ten years from 1800 to 1950. Exhibit **9** shows the "little thought" version, in which millions of people are plotted against date. What can we see in this graph?

exhibit **9** of chapter 5: U.S.A. population

Population of the U.S.A. (in millions; linear scale)



For something like the first half of the period the curve is hollow upward, so that over this period the population was growing at a steadily increasing rate. In fact, it might have been growing at something like a constant percentage each decade. For something like the second half of the period, the growth seems to approximate a straight line. Beyond this, the value for 1940 seems to be somewhat low.

All this is helpful. If we had never before looked at the population of the United States as a function of time, we would have rightly felt that we had learned quite a lot from exhibit 9. But once we have come this far, must we stop? Let us use what we have so far learned to help us look more deeply into the growth of the U.S.A.'s population.

In the present instance this is easy to do. What do we have as bases for further steps? Two things come in sight in exhibit 9:

- ◇ the early years were years of accelerated growth, possibly at a constant percent per year.
- ◇ in the later years, the population grew by about the same number of people each decade.

We can check up on these appearances and, more importantly, try to use them to go further.

To check up on constant-percent growth, the easy thing is to look up the logs of the population sizes, and then plot them against date. (If we didn't expect to go further, we could just plot the raw values on semilogarithmic graph paper instead.) Exhibit **10** shows the result of making such a plot. The earlier part of the plot now looks quite straight, even if we put our eye close to the paper so that we can look right along this hypothetical line. This looks very much as if constant percentage growth per decade is a good description of U.S.A. population growth in the early 1800's. Let us keep this appearance in mind, and plan to come back to it.

the later decades

Before going further with the early decades, however, we shall turn to the apparent linearity of population growth seen in the original graph, exhibit 9, for the later decades. Exhibit **11** shows the result of drawing in a comparison line. To find an equation for this line, note that in 1870 the line has a height of about 35--35 million people--and in 1950 it has a height of somewhat less than 150, say 147. the line through (1870, 35) and (1950, 147) has slope

$$\frac{147 - 35}{1950 - 1870} = \frac{112}{80} = 1.4$$

and thus the equation is

$$y = 35 + 1.4(x - 1870)$$

This calculation is made by simple steps in the form of writing down words and

numbers in the lower right of the exhibit. We recommend such a form (in general with "at" a value of x rather than "in" a date) whenever we eye-fit a line.

This line follows the data quite well, leaving us with a confirmed feeling about both approximate straightness and the dip in 1940. Need we stop here? Surely not. The straight line is an incomplete description of how the data behaves in later years. One of the great arts of data analysis consists of subtracting out incomplete descriptions and examining the residuals that are left. Let us do just this.

coming to details

Exhibit **12** shows the residuals from the line of exhibit 11, for the period from 1800 to 1950. (For example, at 1880 the census population is 50.2 millions and the fit is $35 + 1.4(10) = 49$, so that the residual is +1.2 millions.) The detailed behavior of the population in the later half of this period is now fairly well revealed. The earlier half of the period is, however, telling us little. (Especially since the comparison line

$$35 + 1.4(\text{date} - 1870)$$

gives rather large residuals before 1840, making the comparison in earlier years unlikely to be helpful.) If we gave up looking at the early half, and

exhibit **10** of chapter 5: U.S.A. population

Population of the U.S.A. (in millions; logarithmic scale)

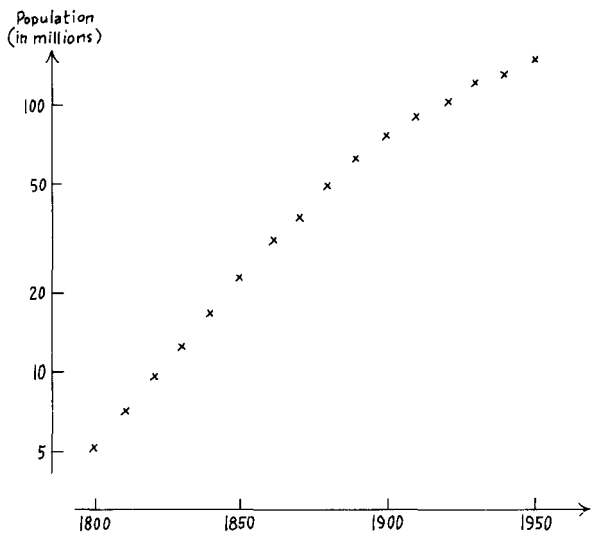


exhibit **11** of chapter 5: U.S.A. population

Population of the U.S.A. (linear scale with comparison line)

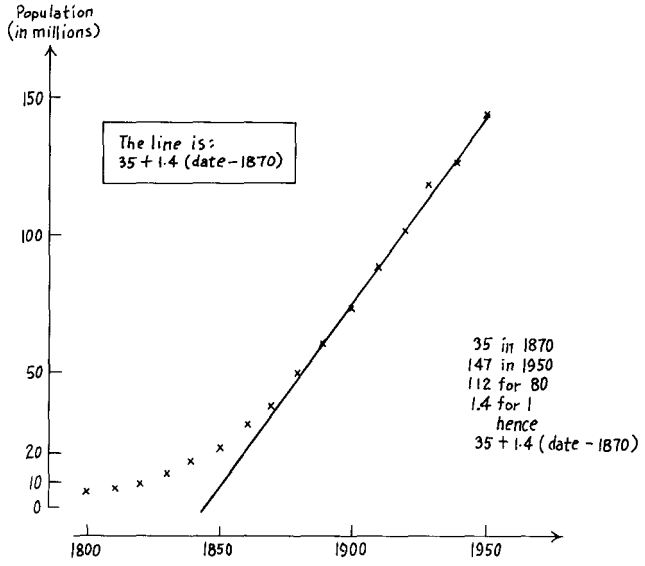
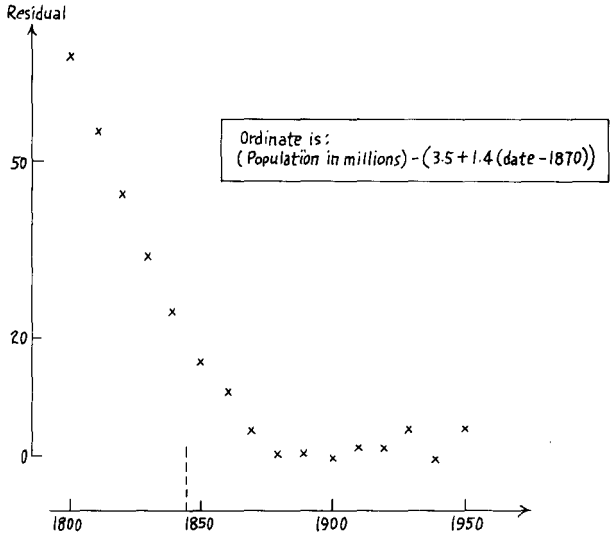


exhibit **12** of chapter 5: U.S.A. population

Population of the U.S.A. (residuals = deviations from the specified straight line)



focused on the later half, we could put the data under a very much more powerful microscope. Why not do just this?

Exhibit **13** shows the same values as the righthand side of exhibit 12 but at 15 times the vertical scale. We can now see that the 1940 population was about five million less than would fit smoothly into the adjacent values. (Why do you think this happened?) And more, for we can now see that 1920 was a couple of millions low also (unless 1930 is thought to be unduly high).

Our magnifying glass is now working at full capacity, at least until we identify some further partial description and arrange to subtract it out also. To learn more about the later years of U.S.A. population growth, we would need either to get year-by-year estimates or to study the mechanisms that are involved.

the earlier details

So much for the later years--what of the earlier ones? We left unfinished business when we said that the lefthand side of exhibit 10 seemed quite straight. We can attack this logarithmic straightness in the early years just as we attacked the linear straightness in the later ones.

exhibit **13** of chapter 5: U.S.A. population

Later population of the U.S.A. (expanded deviations from the specified straight line)

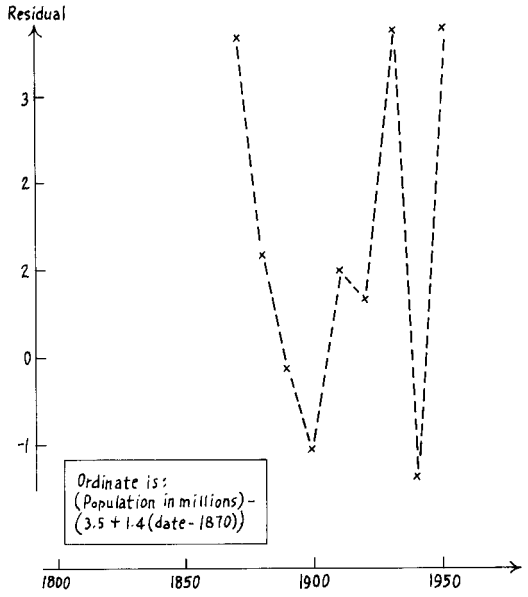


Exhibit **14** shows the result of drawing a straight line on exhibit 10. It is encouraging enough for us to hasten to the residuals, shown in exhibit **15**. (In 1880, a population of 50.2 millions gives a log of 7.70, while the fit is 6.75 + .012(80) = 7.71, so that the residual is −0.01. More decimals were used in calculating the points for exhibit 15.) As we ought to have expected, the residuals prove useful in the early years but of very dubious value in the later ones. We can again afford to use the magnifying glass on the relatively flat section, as is done in exhibit **16**.

Exhibit 16, once we realize that ±0.01 in log is about ±2.3% in size, gives us a quite delicate view of U.S.A. population growth during the nineteenth century. After 1860, population growth was not as fast as before. Moreover, 1800 appears to have been additionally depressed by 3 or 4 per cent. Why? Again, we have gone as far with our microscope as seems reasonable without further inputs.

Experts believe many of the detailed fluctuations now so clear to us are due to variations in completeness of the census, rather than to changes in population growth. Clearly, the data cannot contradict the experts. The fluctuations are in the numbers, whatever their source. It is worthwhile to find them, whether they tell us about the growth of U.S. population or the deficiencies of U.S. censuses.

exhibit **14** of chapter 5: U.S.A. population

Population of the U.S.A. (logarithmic scale with comparison line)

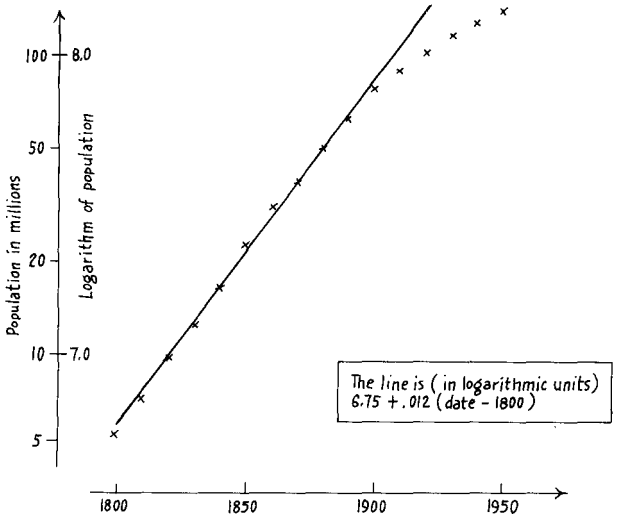


exhibit **15** of chapter 5: U.S.A. population

Population of the U.S.A. (deviations from logarithmic straight line)

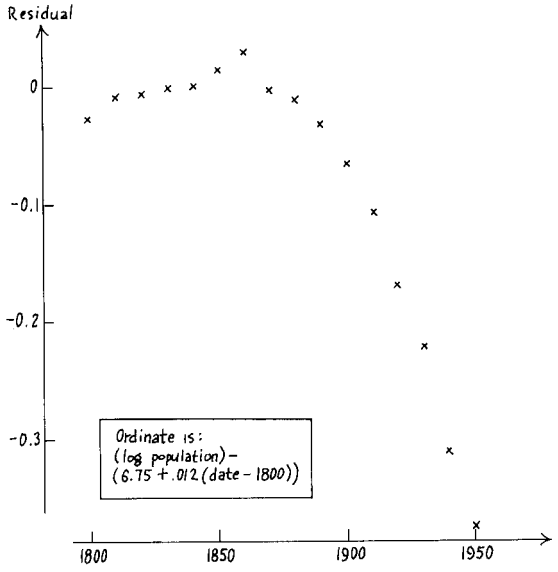
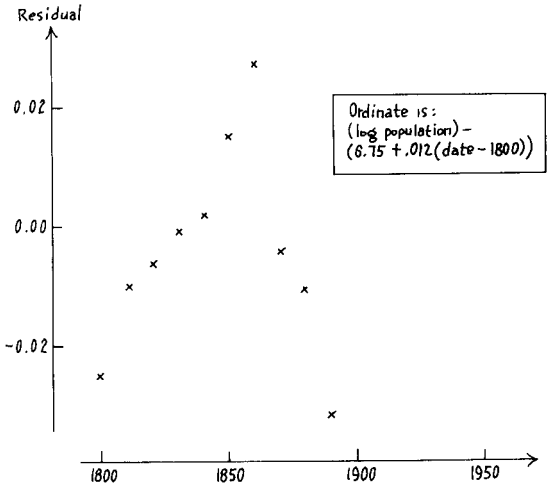


exhibit **16** of chapter 5: U.S.A. population

Earlier population of the U.S.A. (expanded deviations from given logarithmic line)



If we had to choose a set of pictures to summarize U.S.A. population growth as completely as we now can, we would probably choose to show all four of the following:

- ◇ exhibits 14 and 11, to show the general patterns of growth.
- ◇ exhibits 16 and 13, to show local behavior.

Together, these four would be responsive to the request: Make useful plots of U.S. population against date. (If we were population specialists, we would know about logistic functions and be able to fit a single incomplete description all the way from 1800 to 1950. This would simplify the plotting of a single graph of residuals, and would probably allow us to summarize the situation in two plots, one showing the fit and the other the residuals.)

What are the lessons to be learned from this example? Not merely that thought can help us see deeper. We have seen specific examples of very general principles, including these:

- ◇ **choosing scales to make behavior roughly linear always allows us to see local or idiosyncratic behavior much more clearly.**
- ◇ **subtracting incomplete descriptions to make behavior roughly flat always allows us to expand the vertical scale and look harder at almost any kind of remaining behavior.**

Whatever the data, we can try to gain by straightening or by flattening. When we succeed in doing one or both, we almost always see more clearly what is going on.

review questions

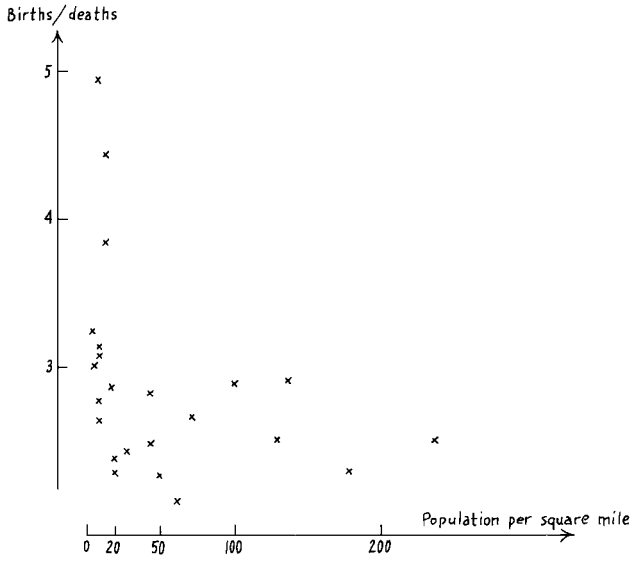
Can we plot without thinking? How much will we learn? What can we see from exhibit 9? How can we use what we see? What does exhibit 11 tell us to do next (go forward as far as you can)? Did we do it? Where? What does exhibit 14 tell us to do next (go forward as far as you can)? Did we do it? Where? Which pictures would we choose to "tell all" about the U.S. population? Why? What are two important lessons to be learned from this example?

5E. Plotting the ratio of births to deaths

The *County and City Data Book* of the U.S. Bureau of the Census contains much varied information. In particular, the 1962 edition gives for each state the number of live births for 1959, the number of deaths for 1959, and the density of population for 1960. One who believed in the "wide open spaces" might feel that the ratio of births to deaths would appear to be influenced by the density of population, at least if the South and the Atlantic coast states were set aside. A plot of the ratio of births to deaths against

exhibit **17** of chapter 5: births and deaths

Births/deaths and population density



population density for the remaining states appears as in exhibit **17**. About all we can say from this plot is that the point spatter seems crudely L-shaped. The use of a linear scale for population per square mile has squeezed so many of the states up against the vertical axis that we can't be sure what is going on. If we are to see what, if anything, is going on, we must adjust the left-to-right scale so that the states are less jammed together.

Exhibit **18** shows the result of using a log scale for population density. We now see that the three states with unusually high ratios of births to deaths have low, but not very low densities, and do not appear to be typical of very low-density states. Looking only at the other states, there may be a faint tendency for a higher ratio of births to deaths to go with lower population density. (By setting aside the three states, we could double the vertical scale for the rest, putting this tendency under a slightly stronger microscope. Doing this teaches us almost nothing new, as the reader may verify.)

another try

If we are to give up on population density, what next? The three unusual states (and their birth/death ratios) were: New Mexico (4.95), Utah (4.46), and Arizona (3.83). Looking again in the *County and City Data Book*, we discover

exhibit **18** of chapter 5: births and deaths

Births/deaths and population density by states (density on logarithmic scale)

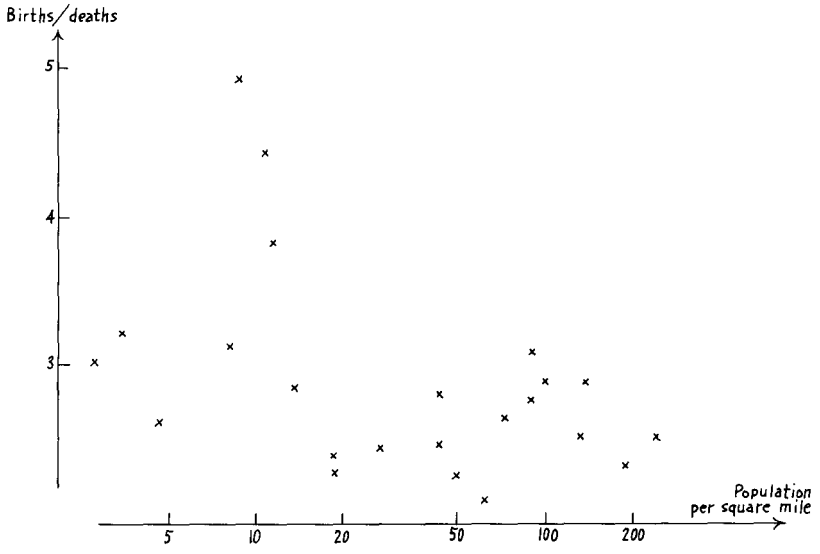
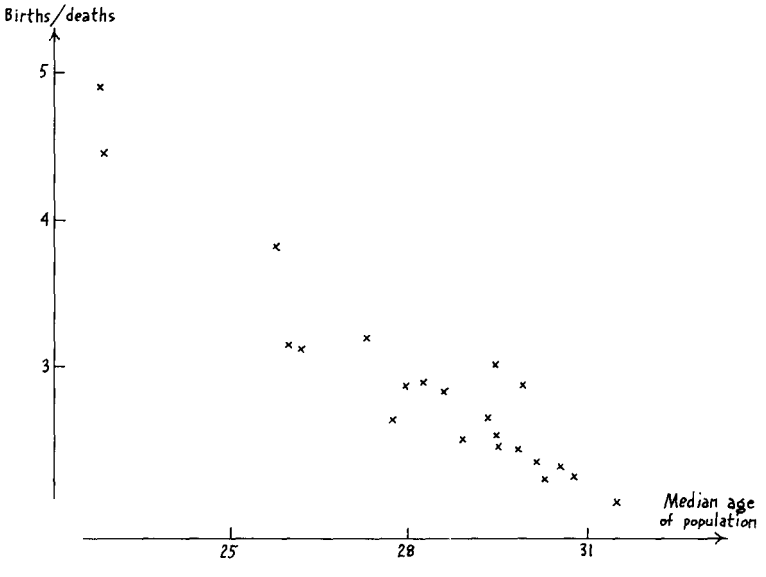


exhibit **19** of chapter 5: births and death

Births/deaths and median age by state



that all three have young populations, for instance, as measured by median age. Thus it is natural to plot births/deaths against median age.

Exhibit **19** shows the result. Clearly, median age does a much better job of appearing to explain the ratio of births to deaths than did population density, though the apparent dependence is, of course, not perfect. If we are to go further, we will need to find and subtract out some partial description of this apparent dependence.

When we "put eye to paper" and look along the point spatter in exhibit 19, we seem to see a definite tendency to curvature (hollow above). If we could eliminate this, we could reasonably compare the individual points with a straight line. How might we approach such a simpler partial description?

One thing to try is changing the down-to-up scale. It is easy to see that using squares to fix this scale would make curvature worse, so we may as well try going in the oppøsite direction by using logs, a choice which is attractive because of the more symmetric way it treats births and deaths. Notice how the identities

$$\log\left(\frac{\text{births}}{\text{deaths}}\right) \equiv \log \text{births} - \log \text{deaths}$$

$$\equiv -(\log \text{deaths} - \log \text{births}) \equiv -\log\left(\frac{\text{deaths}}{\text{births}}\right)$$

exhibit this symmetry.

Exhibit **20** shows the result, complete with a convenient comparison line. The point spatter is now much more nearly straight. When we plot the corresponding residuals against median age and identify the more extreme states, we find the results shown in exhibit **21**.

going to the map

The two states with notably high residuals are adjacent to one another on the map. The four states with middle median age and notably low residuals also touch one another. Clearly, we need to see the residuals on a map.

Exhibit **22** shows the residuals spread across a map of the United States. The roughly regular structure of these residuals is rather easily seen--adjacent states are clearly more often similar than are distant ones. (Might the somewhat surprisingly positive residual for Illinois be due to the unusual size of Chicago?) To go further here requires either:

- ◇ more careful allowance for the age of the population, or
- ◇ more knowledge about the mechanisms affecting birth and death rates in general. (Would state-by-state information on economic conditions help?)

Indeed, both of these are likely to be needed!

exhibit **20** of chapter 5: births and deaths

Births/deaths (logarithmic scale) and median age by state

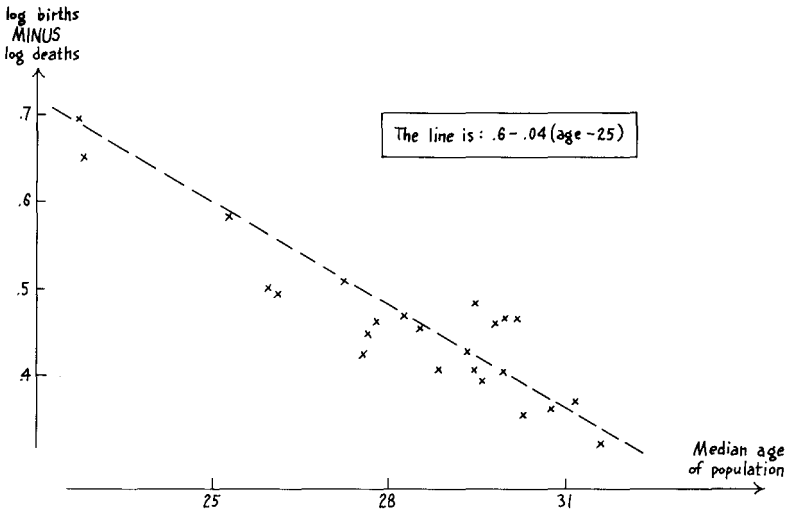
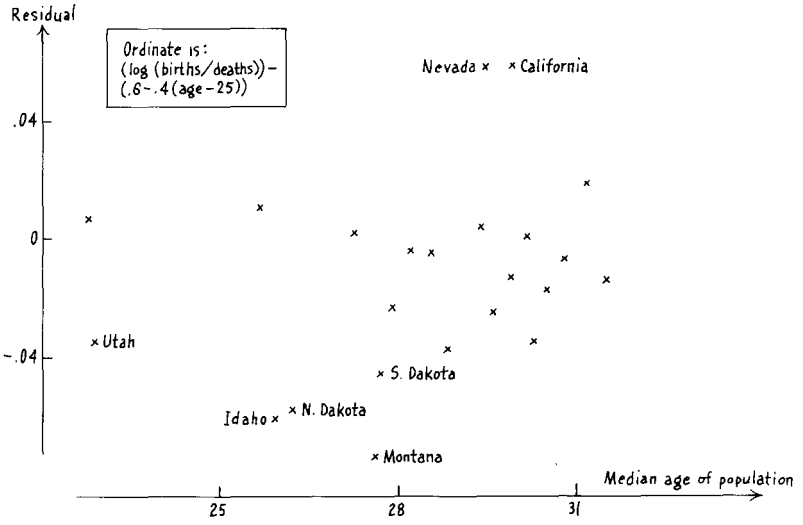


exhibit **21** of chapter 5: births and deaths

Residuals of log $\frac{\text{births}}{\text{deaths}}$ against median age by state



In this example, we have again seen the same main points as in the previous one:

- ◇ **changing scales to make dependences roughly linear usually helps.**
- ◇ **flattening by subtraction makes it much easier to see what is going on at more subtle levels.**

The fact that our dependences were approximate rather than exact did not alter these main points at all.

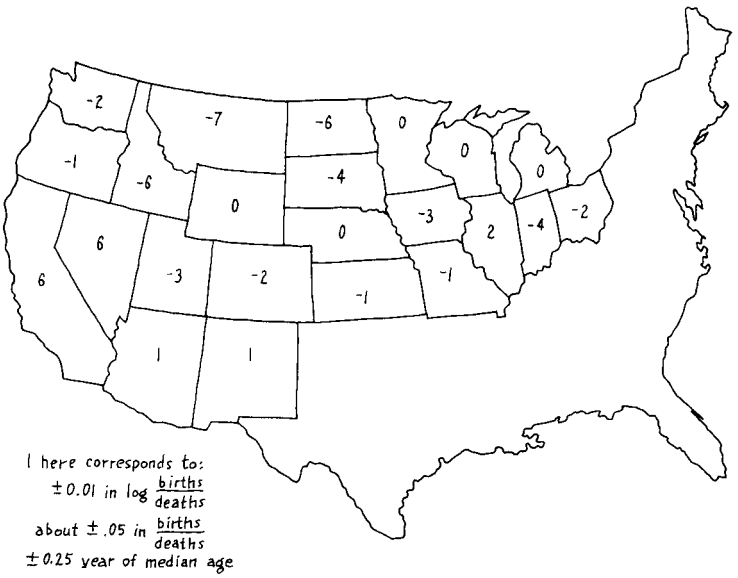
Approximate dependence did, however, bring in one new aspect, as we saw in exhibits 17 and 18:

- ◇ **the usefulness of changing scales to reduce confusion caused by crowding.**

In both examples, it is clear that we never expect the data to be ON the line--only that it might, if we are lucky, be NEAR the line. Once we put the first example under the microscope, what we saw is not intrinsically different from the other example. If we had contracted out the 1850 census to two

exhibit **22** of chapter 5: births and deaths

The residuals of exhibit 17 mapped



different contractors, we would have gotten two different numbers for the population of the U.S.A. The potential values of the U.S.A. population--those we might reasonably have found--do not lie ON a curve; they merely lie QUITE NEAR one. We have happened not to buy more than one census value at a time. This makes the first example LOOK a little different, but will not keep us, in the next chapter, from interchanging the axes used for date and population. Once we face the uncertainties of "what the numbers might have been," almost all data is at best "just NEAR a line or curve".

review questions

Where did we get the data for the example of this section? What did our first try teach us? Why did it pay to use a log scale for population density? What did we try next? Why? How well did it work? Why did we try a log scale for births/deaths? What two exhibits combine to tell the story (no map yet)? Why did we go to a map? How well did it work?

5F. Untilting defines "tilt"

We are now well aware of how much we can gain by flattening our picture, because this lets us expand its vertical scale. It is well to have things straight before flattening them; we can then "blow up" the picture even more, but flattening of even unstraightened pictures can help too.

We want a procedure that flattens the data out, whether or not it is straight. Here "flattening" must refer to the general run of the data, and not to its detailed behavior. Something that looks like a plausible set of residuals from a straight line is "flattened", though it may not be flat.

Exhibit **23** shows a rather extreme example of curvature, and exhibit **24** the results of "flattening" exhibit 23. Clearly, exhibit 24 is far from being flat, so "flattening" is a poor term. Equally, it is poor use of words to talk about "the slope" of exhibit 23, since there is a very small slope toward the left of the plot and a very large slope to the right.

Tilt is a short word, and conveys an appropriate feeling. We shall say that exhibit 24 is

untilted,

and that the slope of the line whose subtraction converts exhibit 23 into exhibit 24--thus clearing away an important part of what is going on so that we can see better what remains--is the

tilt

of exhibit 23.

exhibit **23** of chapter 5: illustrative

A tilted set of data

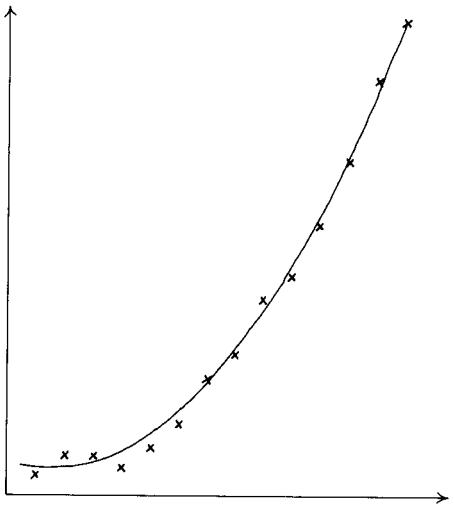
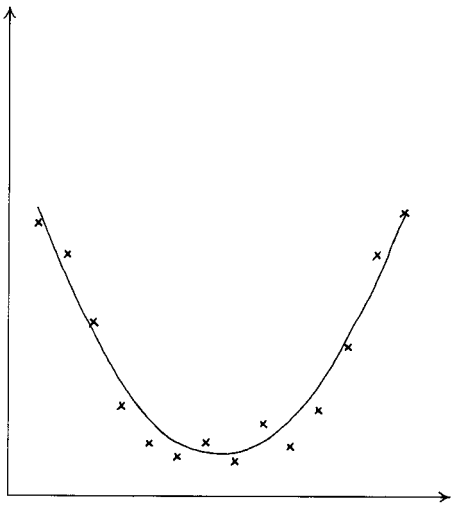


exhibit **24** of chapter 5: illustrative

The data of exhibit 23 untilted



Thus, we take the question "How much is y tilted against x?" to mean: "What value of b leaves $y - bx$ apparently untilted?" Equally, we shall take the b thus found as our assessment of tilt, and the values of $y - bx$ as what we have been able to do about freeing y of tilt against x.

There can be a variety of different--often only slightly different, usually not very much different--definitions of "apparently untilted". For each such definition there is a--comparably slightly different--value of b and thus both an assessed value for the tilt and a set of apparently untilted values. These differ--from definition to definition--by only comparably slight amounts.

The existence of such alternatives--and the need to choose between them--usually bothers us not at all.

In the great majority of cases, our concern with tilt is like the woodsman's concern with bushes and fallen trees on a trail he plans to use:

- we may be concerned to know that there appears to be some tilt (though we may have known that there would be a tilt in this particular direction long before we collected the data).
- we may even want to know how much tilt there seems to be.
- **we are almost certain,** whatever else, **to want to clear the tilt out of our way.**

We have already, in specific instances, cleared our way by using an eye-judged tilt to flatten our graphs. We frequently need to do something similar in fixing numbers for future analysis. Especially because we are going to use any tilt we assess in such a way, we will meet our needs quite well enough if we do a reasonably good job of assessing our tilt. (We need not worry as to in what way, if any, our assessment might be **best possible**--or even whether anyone can define "best possible" sensibly.)

review questions

What is a tilt? Why use a special word? What is it to be untilted? How do we define tilt? Are there many definitions, or few, or only one definition of "untilted"? Do we need a "best possible" tilt? Can you define a "best possible" tilt?

5H. How far have we come?

In this chapter we have met plots of y against x. Perhaps we have even come to know them a little.

The state of our progress is not measured by the specific techniques we have seen or understood--though that kind of progress is essential. Rather, our progress is measured by our acceptance of such propositions as these:

1. **Graphs are friendly.**

2. **Arithmetic often exists to make graphs possible.**

3. **Graphs force us to note the unexpected; nothing could be more important.**

4. **Different graphs show us quite different aspects of the same data.**

5. **There is no more reason to expect one graph to "tell all" than to expect one number to do the same.**

6. **"Plotting y against x" involves significant choices--how we express one or both variables can be crucial.**

7. **The first step in penetrating plotting is to straighten out the dependence or point scatter as much as reasonable.**

8. **Plotting y^2, $\sqrt{y}$, log y, $-1/y$ or the like instead of y is one plausible step to take in search of straightness.**

9. **Plotting x^2, $\sqrt{x}$, log x, $-1/x$ or the like instead of x is another.**

10. **Once the plot is straightened, we can usually gain much by flattening it, usually by plotting residuals (with regard to the partial description implied by the straight line we may not have quite drawn in yet).**

11. **When plotting point scatters, we may need to be careful about how we express x and y in order to avoid concealment by crowding.**

In particular, we have learned--or been reminded--how to:

- ◇ understand about subtracting one curve from another.
- ◇ use two points to find an equation for a straight line.

Our two examples differed in one way. The years of census are fixed by law, and at each a population is measured. It is rather easy to try to think of the points as having been selected from a curve with one population at each possible date. Births/deaths, on the other hand, whether compared with population density or with median age of population, provides a much more symmetric situation. The boundaries of states are fixed by law, and for each state two things happen, an x and a y. There is no possibility of assuming that all the data are ON a curve; the most we can hope is that the data are NEAR a line or curve.

5P. Additional problems

See exhibits **25**, **26**, **27**, **28**, **29**, **30**, **31**, **32**, **33**.

exhibit **25** of chapter 5: data and problems

Data for the 15 smallest counties of 3 states (from 1962 *County and City Data Book*)

A) SOUTH CAROLINA

		Population (1960)				Families (1959)	Local Gov't (1957)	
(*)	Name	Total	% less than 5 yrs school†	% at least 65	per sq. mile	% less than $3000 income	% budget for education	Total budget ($1000's)
2424	McCormick	8,629	28.2	8.8	23	59.8	62.0	566
2130	Allendale	11,362	34.1	7.9	27	60.1	63.1	1082
2051	Jasper	12,237	37.1	7.5	19	60.2	79.7	1082
2050	Calhoun	12,256	26.1	7.9	33	68.2	76.6	659
1852	Saluda	14,554	17.0	9.7	33	50.6	68.9	656
1753	Edgefield	15,735	23.1	8.0	33	55.3	70.1	785
1720	Bamberg	16,274	26.6	8.3	41	58.5	63.1	1340
1622	Hampton	17,425	31.8	7.2	31	58.0	57.8	1312
1608	Barnwell	17,659	23.2	7.4	32	47.5	82.9	1565
1405	Fairfield	20,713	30.8	7.8	30	54.2	56.8	1596
1356	Abbeville	21,417	23.0	8.7	42	42.1	56.4	1717
1339	Lee	21,832	31.7	6.4	53	68.6	75.8	1671
1211	Dorchester	24,383	23.1	6.6	43	49.6	67.9	1684
1092	Colleton	27,816	29.3	7.5	27	57.9	46.7	3090
1066	Marlboro	28,529	29.6	6.8	59	58.3	63.4	1948
(Whole state)		((34,262))	(20.3)	(6.3)	(79)	(39.5)	(63.4)	—

State has 46 counties.

B) GEORGIA

3077	Echols	1,876	28.0	8.5	4	55.6	59.5	122
3051	Quitman	2,432	34.8	9.3	14	70.0	79.5	205
3028	Glascoch	2,672	35.6	56.4	19	61.1	67.9	252
2979	Webster	3,247	34.5	8.7	17	71.2	73.3	225
2978	Schley	3,256	23.0	10.3	20	67.4	69.7	221
2970	Taliaferro	3,370	27.9	13.4	17	68.5	48.8	391
2950	Dawson	3,590	22.4	8.6	17	64.7	71.3	293
2934	Long	3,874	22.7	7.5	10	60.1	52.0	408
2875	Towns	4,538	11.3	10.2	27	63.7	33.4	724
2674	Baker	4,543	33.2	8.9	15	74.1	56.4	622
2873	Clay	4,551	26.8	10.5	20	66.6	59.2	417
2832	Lanier	5,097	30.1	7.5	31	57.4	64.5	686
2810	Charlton	5,313	26.5	6.2	7	44.0	49.1	703
2808	Heard	5,333	22.7	10.7	18	56.4	57.1	652
2805	Wheeler	5,342	29.4	9.5	18	63.7	54.4	418
(Whole state)		((12,038))	(17.6)	(7.4)	(68)	(35.6)	(45.0)	—

State has 159 counties. →

exhibit **25** of chapter 5 (continued)

C) ALABAMA

2201	Coosa	10,726	18.4	10.7	18	51.5	39.5	1174
2179	Cleburne	10,911	21.5	9.4	19	52.4	48.2	1088
2034	Clay	12,400	14.4	13.0	21	54.0	48.3	1347
1945	Bullock	13,462	32.7	11.7	22	69.4	42.8	1727
1934	Greene	13,600	38.0	9.8	21	74.0	57.9	1514
1876	Lamar	14,271	14.7	11.0	24	51.4	37.5	2507
1870	Bibb	14,357	22.7	9.6	23	54.4	54.7	1405
1828	Winston	14,858	17.3	9.8	24	53.8	55.7	1072
1824	Crenshaw	14,909	23.9	11.2	24	69.5	51.4	1332
1796	Henry	15,286	27.8	9.0	27	63.8	56.0	1313
1794	Washington	15,372	23.3	7.8	14	51.7	54.5	1353
1790	Lowndes	15,417	37.2	9.2	22	72.1	55.4	1243
1729	Fayette	16,148	16.9	10.8	26	54.7	43.0	1425
1715	Cherokee	16,303	16.6	8.9	27	49.1	53.7	1498
1628	Perry	17,358	28.7	10.5	24	69.2	49.4	1689
(Whole state)		((25,738))	(16.3)	(8.0)	(64)	(39.1)	(45.8)	--

State has 67 counties.

*** National rank by population.**
† Of those 25 or over.
( ) Median.

P) PROBLEMS

25a) Panels A to C contain selected information about the 15 smallest counties in South Carolina, Alabama, and Georgia. Plot

y = % with less than five years schooling

against

x = % with less than $3000 income

for at least two states. Continue the analysis. Comment.

S) SOURCE

1962 *County and City Data Book.*

exhibit **26** of chapter 5: data and problems

Some problems

26a) The following data was obtained in preparing a standard curve for the determination of formaldehyde by the addition of chromatropic acid and concentrated sulphuric acid and the reading of the consequent purple color on a Beckman Model DU Spectrophotometer at 570 mμ.

Amount of CH_2O Used	Optical Density
0.1	0.086
0.3	0.269
0.5	0.446
0.6	0.538
0.7	0.626
0.9	0.782

Analyze graphically, using at least two graphs. Comment. (Bennett & Franklin, p. 216; from Roberts.)

26b) The relation between the amount of β-erythrodine dissolved in water and the turbidity of the solution--as read on a colorimeter--is not quite as simple. Some data gives:

Concentration (in mg/ml)	Colorimeter reading
40	69
50	175
60	272
70	335
80	390
90	415

Analyze graphically. Comment. (Bennett & Franklin, p. 217, from Wois-lawski.)

26c) Find two different collections of (x, y) points that interest you, and make useful plots.

S) SOURCE

See exhibit 27.

exhibit **27** of chapter 5: data and problem

Carbon content of 36 clays measured directly and estimated indirectly

A) DATA

Clay #	Direct measurement	Indirect estimate
1	1.53	2.46
2	0.87	1.54
3	0.28	0.70
4	0.27	−0.40
5	3.07	4.82
6	0.25	0.30
7	0.25	0.64
8	0.29	0.78
9	0.12	0.12
10	1.50	2.36
11	1.31	2.14
12	0.31	0.08
13	0.14	−0.01
14	2.98	4.53
15	6.84	9.94
16	2.15	3.68
17	1.35	1.84
18	0.40	0.97
19	4.18	6.14
20	0.22	0.52
21	0.38	0.40
22	0.24	0.46
23	1.79	2.80
24	0.58	2.09
25	6.55	9.68
26	2.54	4.08
27	1.43	2.80
28	2.74	3.93
29	6.08	8.22
30	0.75	0.28
31	0.16	0.35
32	5.06	7.49
33	0.86	1.41
34	0.16	−0.50
35	11.43	15.80
36	0.19	0.18

→

exhibit **27** of chapter 5 (continued)

P) PROBLEM

27a) The amount of carbon in a clay can be measured directly by heating the clay until all the carbon compounds are burned, collecting the carbon dioxide thus formed, and measuring its amount. The amount of carbon can be estimated by combining the amounts of its constituents in a suitable standard way. The results of such measurements on 36 clays from South Devonshire, England, are given in exhibit 27. Analyze graphically. Comment.

S) SOURCE

C. A. Bennett and N. L. Franklin 1954, Statistical Analysis in Chemistry and the Chemical Industry. John Wiley, New York. Table 6.3 on page 218.

exhibit **28** of chapter 5: data and problems

Percentage Democratic in 12 presidential elections for 24 Northeastern and Central States (percentage Democratic of major party vote)

A) DATA

	1920	1924	1928	1932	1936	1940	1944	1948	1952	1956	1960	1964
Colorado	37.7	27.8	34.4	57.0	61.9	48.7	46.6	52.7	39.3	39.5	45.1	61.6
Connecticut	34.5	30.9	45.9	49.4	57.8	53.6	52.7	49.2	44.1	36.3	53.7	67.9
Delaware	43.0	38.9	33.9	48.8	54.9	54.8	54.6	49.4	48.1	44.7	50.8	61.1
Illinois	27.3	28.4	42.6	56.8	59.2	51.2	51.7	50.4	45.0	40.4	50.1	59.5
Indiana	42.3	41.2	39.9	56.0	57.5	49.3	47.1	49.6	41.4	39.9	44.8	56.2
Iowa	26.4	23.0	37.8	59.1	56.0	47.8	47.7	51.4	35.8	40.8	43.3	62.0
Kansas	33.4	27.7	27.3	54.8	53.9	42.7	39.4	45.4	30.7	34.3	39.3	54.6
Maine	30.2	23.3	31.1	43.6	42.8	48.8	47.5	42.7	33.8	29.1	43.0	68.8
Maryland	43.3	47.7	42.6	63.1	62.7	58.8	51.9	49.3	44.2	40.0	53.6	65.5
Massachusetts	28.9	28.5	50.5	52.1	55.1	53.4	52.9	55.9	45.6	40.5	60.4	76.5
Michigan	23.4	14.8	29.1	54.1	59.2	49.8	50.5	49.1	44.2	44.2	51.0	66.8
Minnesota	21.6	11.7	41.4	62.3	66.6	51.9	52.8	58.9	44.4	46.2	50.7	63.9
Nebraska	32.6	38.5	36.4	64.1	58.4	42.8	41.4	45.8	30.8	34.5	37.9	52.6
New Hampshire	39.7	36.7	41.2	49.3	50.9	53.2	52.1	47.1	39.1	33.9	46.6	63.9
New Jersey	29.6	30.6	40.0	51.0	60.1	51.8	50.7	47.7	42.5	34.6	50.4	66.0

→

exhibit **28** of chapter 5 (continued)

New York	29.5	34.3	48.8	56.7	60.2	51.8	52.5	49.5	44.0	38.7	52.6	68.7
North Dakota	18.9	12.7	44.8	71.3	69.2	44.7	45.8	45.4	28.6	38.2	44.5	58.1
Ohio	39.8	28.9	34.7	51.5	60.8	52.2	49.8	50.1	43.2	38.9	46.7	62.9
Pennsylvania	29.2	22.6	34.2	47.1	58.2	53.5	51.4	48.0	47.0	43.4	51.2	65.2
Rhode Island	33.9	37.9	50.3	56.0	56.8	56.8	58.7	58.2	49.1	41.7	63.6	80.9
South Dakota	24.5	21.2	39.4	64.9	56.0	42.6	41.7	47.6	30.7	41.6	41.8	55.6
Vermont	23.5	16.7	33.0	41.6	43.4	45.1	42.9	37.5	28.3	27.8	41.4	66.3
West Virginia	43.9	47.1	41.3	55.1	60.7	57.1	54.9	57.6	51.9	45.9	52.7	67.9
Wisconsin	18.5	17.9	45.3	67.0	67.8	50.9	49.1	52.3	38.8	38.1	48.1	62.2

P) PROBLEMS

Panel A gives the % Democratic vote in each of 24 northeastern and central states for 12 presidential elections, 1920 to 1964. Plot the following:

28a) 1964 against 1956

28b) 1960 against 1920

28c) 1952 against 1932

28d) Any one against any other that you think will have a close relationship.

exhibit **29** of chapter 5: data and problems

More problems

29a) Determination of ethylene chlorohydrin. (264, 270) translates as: "With 26.4 milligrams of ethylene chlorohydrin present, 27.0 milligrams were found". Data-sets (6): (264, 270), (595, 594), (1173, 1183), (1777, 1780), (2355, 2370), (3578, 3576). SOURCE: K. Uhrig 1946. Determination of ethylene chlorohydrin. *Industrial and Engineering Chemistry, Analytical Edition* **18**: 369 only. Table 1 on page 369. PROBLEM: Choose a plot that is likely to be revealing by thinking hard. Explain the reasons for your choice. Make the plot.

29b) Polarographic behavior of ions containing vanadium. (94, 35) translates as: "For a concentration of vanadite ion of 0.094 millimoles per liter, the anodic diffusion constant in microamperes was 0.35 microamperes". Data-sets (8): (94, 35), (278, 98), (508, 178), (880, 309), (1548, 563), (1840, 696), (352, 1285), (505, 1813). (Last point: 5.05 millimoles/liter, 18.13 microamperes.) SOURCE: J. J. Lingane 1945, "Polarographic characteristics of vanadium in its various oxidation states," *J. Amer. Chem. Soc.* **67**:182–188. Table I on page 186. PROBLEM: Make useful plots of diffusion current against vanadite concentration. ➔

exhibit **29** of chapter 5 (continued)

29c) Amount of desired product in a chemical reaction after different reaction times and under different circumstances. (1; 32, 54; 87, 159, 226) translates as: "In run 1, the amount of desired product in moles per liter was 0.032 after 80 minutes, 0.054 after 160 minutes, 0.087 after 320, 0.159 after 640, and 0.226 after 1280 minutes." Data-sets (16 runs under 16 different sets of conditions): (1; 32, 54; 87, 159, 226), (2; 147, 234; 343, 342, 203), (3; 48, 108; 225, 346, 420), (4; 232, 390; 556, 634, 416), (5; 37, 38; 172, 200, 239), (6; 179, 283; 405, 342, 216), (7; 86, 133; 259, 398, 508), (8; 309, 514; 722, 764, 389), (9; 74, 99; 200, 309, 249), (10; 253, 343; 391, 284, 75), (11; 133, 271; 430, 580, 494), (12; 508, 756; 842, 570, 115), (13; 96, 158; 276, 339, 230), (14; 308, 444; 467, 249, 29), (15; 228, 372; 579, 691, 539), (16; 626, 880; 895, 434, 58). SOURCE: G. E. P. Box and W. G. Hunter 1962, "A useful method for model-building," *Technometrics* **4**: 301–318. Table 1 on page 304. PROBLEM: Make useful plots of concentration at 640 minutes against concentration at 160 minutes.

29d) Make useful plots for one or more other pairs of reaction times. (Data-sets in problem (29c) above)

exhibit **30** of chapter 5: data and problems

Yet more problems

30a) Analysis of samples for chrysanthenic acid. (0, 23) translates as: "When 0 micrograms of synthetic racemic chrysanthenic acid were added, the colorimeter scale reading was 23". Data-sets (13): (0, 23), (5, 32), (10, 40), (20, 54), (40, 86), (60, 118), (80, 146), (100, 179), (120, 212), (140, 240), (160, 272), (180, 300), (200, 330). SOURCE: A. A. Schreiber and D. B. McClellan 1954. Estimation of microquantities of pyrethroids. *Analytical Chemistry* **26**: 604–607. Table I on page 605. PROBLEM: Make useful plots of colorimeter reading against amount of chrysanthenic acid.

30b) Residual strength of 8-oz. cotton duck attacked by 4 different kinds of fungus. (3; 97, 105; 103, 101) translates as: "After 3 hours of incubation, the strengths--referred to initial strength = 100--of the sample exposed to Thielaria was 97, that exposed to Humicola was 105, for Chaetomium was 103, for Myrothecium was 101." Data-sets (24): (3; 97, 105; 103, 101), (6; 98, 106; 101, 105), (9; 95, 107; 99, 95), (12; 96, 105; 95, 95), (15; 97, 106; 90, 100), (18; 98, 102; 91, 97), (21; 97, 101; 78, 98), (24; 97, 90; 74, 93), (27; 90, 81; 71, 82), (30; 96, 78; 71, 76), (33; 89, 73; 65, 67), (36; 88, 69; 58, 64), (39; 89, 63; 53, 59), (42; 86, 59; 47, 54), (45; 82, 55; 44, 50), (48; 79, 53; 44, 42), (51; 73, 52; 42, 41), (54; 73, 41; 40, 40), (57; 73, 42; 40, 39), (60; 68, 41; 39, 35), (63; 59, 36; 38, 37), (66; 57, 37; 37, 33), (69; 57, 31; 35, 34), (72; 55, 34; 36, 31). SOURCE: E. Abrams 1950, "Microbiological deterioration of cellulose during the first 72 hours of attack," *Textile Research J.* **20**: 71–86. Table 2 on page 75. PROBLEM: Plot helpful curves for loss of strength from at least two kinds of fungus. ➔

exhibit **30** of chapter 5 (continued)

30c) Rapid analysis for caffeine. (257, 131) translates as: "For a caffeine concentration of 0.257 milligrams in 100 milliliters, the average optical density was 0.131." Data-sets (20): (257, 131), (498, 262), (506, 265), (514, 263), (747, 384), (760, 393), (770, 396), (996, 512), (1013, 518), (1027, 523), (1245, 633), (1266, 643), (1284, 650), (1494, 760), (1519, 768), (1541, 775), (1798, 903), (2054, 1040), (2311, 1160), (2568, 1290). SOURCE: N. H. Ishler, T. P. Finucaine, and E. Borker 1948, "Rapid spectrophotographic determination of caffeine," *Analytical Chemistry* **20**:1162–1166. Table 1 on page 1162. PROBLEM: Make useful plots of optical density against caffeine concentration.

exhibit **31** of chapter 5: data and problems

Still more problems

31a) Survival of automobiles and trucks in use by a public utility. (Oh, 990) translates as: "After 1/2 year, 0.990 of all vehicles were still in service." Data-sets (8): (0h, 990), (1h, 972), (2h, 944), (3h, 895), (4h, 784), (5h, 679), (6h, 593), (7h, 497). SOURCE: S. A. Krane 1963, "Analysis of survival data by regression techniques," *Technometrics* **5**: 161–174. Table on page 168. His source: H. A. Cowles, Jr., 1957. Prediction of mortality characteristics of industrial property groups. Ph.D. Thesis, Iowa State University. PROBLEM: Make helpful plots of fraction surviving against age.

31b) Heat and entropy contents of a sodium silicate. (400, 3080, 885) translates as: "The increases from 'room temperature' (298.16°K) to an absolute temperature of 400°K were 3,080 calories per mole for the heat content of Na_2SiO_3 and 8.85 calories/degree/mole for its entropy content." Data-sets (17): (400, 3080, 885), (500, 6300, 1604), (600, 9650, 2214), (700, 13190, 2760), (800, 16910, 3256), (900, 20730, 3708), (1000, 24700, 4124), (1100, 28770, 4511), (1200, 32940, 4874), (1300, 37210, 5216), (1361, 39870, 5416), (1361, 52340, 6332), (1400, 54010, 6453), (1500, 58390, 6748), (1600, 62570, 7024), (1700, 66850, 7284), (1800, 71130, 7528). (The last point, at 1800°K, gives a heat content change of 71,130 calories/mole and an entropy content change of 75.28 calories/degree/mole.) SOURCE: B. F. Naylor 1945, "High-temperature heat contents of sodium metasilicate and sodium disilicate," *J. Amer. Chem. Soc.* **67**: 466–467. Table II on page 467. PROBLEM: Make helpful plots of the increase in heat content against temperature.

31c) For Naylor's data (immediately above) make helpful plots of increase in entropy content against temperature.

exhibit **32** of chapter 5: data and problems

And yet more

32a) Equilibrium splitting of plutonium tribromide by water (gases at high temperature). (911, 153) translates as' "For an absolute temperature of 911°K, the observed equilibrium constant was 0.0153/atmosphere." Data-sets (11): (911, 153), (914, 156), (919, 149), (920, 163), (882, 246), (876, 282), (875, 247), (883, 243), (815, 704), (817, 502), (816, 692). SOURCE: I. Shift and N. R. Davidson 1949. Equilibrium in the vapor-phase hydrolysis of plutonium tribromide. Paper 6.24, at pages 831–840 of *The Transuranium Elements*, edited by Seaborg, Katz, and Manning. National Nuclear Energy Series IV-14B. McGraw Hill. Table 2 on page 835. PROBLEM: Make useful plots of equilibrium constant against temperature. Which three of the 11 data-sets do you think the authors rejected?

32b) Sales of Swiss bond issues since World War II. (46, 527) translates as: "In 1946, total sales of Swiss bonds--governmental and private--were 527 million francs". Data-sets (23): (46, 527), (47, 276), (48, 472), (49, 342), (50, 174), (51, 434), (52, 333), (53, 249), (54, 242), (55, 492), (56, 613), (57, 1148), (58, 827), (59, 686), (60, 890), (61, 1023), (62, 1124), (63, 2091), (64, 2503), (65, 2523), (66, 2292), (67, 2446), (68, 2648). SOURCE: *Swiss Statistical Abstract*, issued by the Swiss Credit Bank, November 1969. (Title also in French and German.) Table on page 46. PROBLEM: Make useful plots based on the data from 1950 to 1968.

32c) Comparison of two ways of measuring the water content of samples of the sea bed. (0 to 3; 76, 76) translates as: "For a sample from 0 to 3 inches below the surface of the sea level, measurement of % water by drying in an oven gave 76%, measurement by analyzing for chloride--and using the known concentration of chloride in deep sea water--gave 76%". Data-sets (14): (0 to 3; 76, 76), (3 to 6; 68, 72), (6 to 9; 69, 69), (9 to 12; 67, 67), (12 to 15; 60, 64), (15 to 18; 62, 62), (18 to 21; 60, 60), (21 to 24; 58, 59), (24 to 27; 57, 57), (27 to 30; 55, 56), (30 to 33; 55, 55), (33 to 36; 55, 55), (36 to 39; 53, 54), (39 to 42; 54, 54). SOURCE: L. J. Anderson 1948, "Conductometric titration of chloride in sea water and marine sediments," *Analytical Chemistry* **20**: 618–619. Table II on page 619. PROBLEM: Find differences between water by chloride and water by oven. Make a stem-and-leaf display of these differences. Comment on the appearance of this display. Find and plot residuals from a straight-line fit of water content against depth, separately for each method of finding water content. What do you conclude about the two methods of measuring water?

exhibit **33** of chapter 5: data and problems

Median ages of urban and rural populations and estimated colonial population

A) MEDIAN AGE, URBAN and RURAL POPULATIONS--at U.S. Censuses

Year	Median age	Urban population	Rural population
1950	30.4	88,927,464	61,769,897
40	29.5	74,923,702	57,245,573
30	27.1	68,954,823	53,830,223
20	26.1	54,157,973	51,552,647
10	24.9	41,998,932	49,973,334
1900	23.8	30,159,921	45,834,654
1890	22.9	22,106,265	40,841,449
80	21.6	14,129,735	36,026,048
70	20.6	9,902,361	28,656,010
60	20.2	6,216,518	25,226,803
50	19.5	3,543,716	19,648,160
40	17.9	1,845,055	15,224,398
30	17.2	1,127,247	11,738,773
20	16.5	693,255	8,945,198
10	15.9	525,459	6,714,422
1800	15.7	322,371	4,986,112
1790	15.9	201,655	3,727,559

Median age = A90 Median age of white males
Urban population = A195 Population of "urban territory"
Rural population = A206 Population of "rural territory"

B) ESTIMATED POPULATION OF AMERICAN COLONIES

Year	Estimated population
1780	2,780,369
70	2,148,076
60	1,593,625
50	1,170,760
40	905,563
30	629,445
20	466,185
10	331,711
1700	250,888
1690	216,372
80	151,507
70	111,935
60	75,058
50	50,368
40	26,634
1630	4,646

→

exhibit **33** of chapter 5 (continued)

P) PROBLEMS

33a) Analyze the median ages of panel A carefully.

33b/c) Analyze the urban/rural populations of panel A carefully.

33d) Take logs of the populations in panel B, and compare them with the extension of the fit given in text for 1790 to 1860.

33e) Fit the data of panel B, re-expressing it if necessary.

S) SOURCES

Historical Statistics of the U.S. Colonial times to 1957. Washington 1960.
(Panel A entries are from series A90, A195, A206 as indicated.

Straightening out plots (using three points) 6

chapter index on next page

We are now sure that we want to first straighten out and then flatten out plots. Straightening out is important--in the language of the opening of Chapter 4, a "big deal"--we want to learn to do it as easily as we reasonably can. This chapter is devoted to techniques and examples.

The sort of re-expression that concerns us is--as we have said--almost entirely re-expression of amounts, including large counts. (Here, counts that are never smaller than 3 are surely "large", and others may be.) These are the kinds of numbers where it is natural to take **powers, roots,** and **logs.**

In dealing with them, we will want to be sure that our origins are reasonably chosen. It is only for amounts measured from reasonable origins that we are likely to get full value from changing to a power, a root, or a logarithm.

Powers and roots of amounts are again amounts. Logs of amounts are **balances**. For these purposes--as for so many others--large counts are merely a special kind of amount. Accordingly, (nonzero) powers and roots of large counts are amounts, while logs of large counts are balances.

In thinking about our problems of re-expression--which need not be the same sort of thing as analyzing the data involved--we need to think about whether x--or y--varies much or little. So long as we deal with amounts, the natural way to make comparisons is by ratios--including percents. Thus we are interested in such facts as:

$$\frac{\text{largest } x}{\text{smallest } x} = 3,$$

All x are within $\pm 50\%$ of a middle value,

or

$$\frac{\text{largest } x}{\text{smallest } x} = 1.1,$$

largest x = smallest x plus 10%.

When we work in logs, we are concerned with balances, not amounts, and the natural way of expressing spread is either in log units or--occasionally--by the ratios into which **differences** in log units can be re-expressed.

review questions

What will we try to do in this chapter? What sort of re-expression concerns us? Need we bother about choice of origin? What is important about how much x--or y--varies? When it is an amount or count? When it is a balance?

6A. Looking at three points

We have seen various illustrations of straightening out data. So far, either these have come from simple rational considerations, or else they have come--"out of the air"--with little apparent reason for the choice. How are we to behave when we have some other kind of data? Do we have to try all possible combinations of an expression of y with an expression of x on the whole data? Or can we save most of the effort this would involve?

The purpose of these changes of expression is to straighten out the data. If the data looks curved, in some overall way, we can make this obvious by picking out three representative points. For the early growth of the U.S.A. population, for example, we could choose the points corresponding to 1800, 1850, and 1890. These three points are:

$$(1800,\ 5.3)$$
$$(1850,\ 23.2)$$
$$(1890,\ 62.9)$$

A simple way to see that three points do not lie on a single straight line is to find the slopes of the straight lines through the first pair and the second pair, respectively; this gives:

$$\frac{23.2 - 5.3}{1850 - 1800} = \frac{17.9}{50} = 0.36$$

and

$$\frac{62.9 - 23.2}{1890 - 1850} = \frac{39.7}{40} = 0.99$$

which are quite different. The second slope is greater, so the curve is hollow upward. (Draw yourself a sketch.)

If any pair of choices of expression is going to straighten out the early portion of the U.S.A. population curve, these same choices will have to do a reasonably good job in straightening out these three points. We can save a lot of effort by screening our pairs of expressions on these three points. We will then need to try only the one best--or perhaps the few best--on the whole data.

We can often ease the task a little by choosing the spacing of the three points to simplify matters further. Had we chosen

(1810, 7.2)
(1850, 23.2)
(1890, 62.9)

we could have compared

$$16.0 = 23.2 - 7.2$$

with

$$39.7 = 62.9 - 23.2$$

since both changes in x are the same. (1850 − 1810 = 40 = 1890 − 1850).

Sometimes we can go almost this far by picking points so that changes in x have a simple ratio--1 to 2, 1 to 3, 3 to 2, etc.--instead of being equal.

Even trying all reasonably possible pairs of expressions on three points is an effort. Can we save some of this by looking into the direction in which changing an expression shifts curvature?

review questions

What relation ought three selected points have to all the available points? How do we ask three points about curvature? How can we make asking easier?

6B. Re-expressing *y* alone

Exhibit 19 of chapter 3 shows various expressions of y plotted against y. We see at once that the higher curves are hollow upward, while the lower curves are hollow downward.

To say that a curve is **hollow upward** means that if we take three points on the curve, the middle point is **below** the line joining the other two. Similarly, **hollow downward** means that the middle point is **above** the line joining the other two. Since we are trying to get a middle point **onto** the line joining the outer two, these are just the sort of facts that matter to us.

We can say more about the simple ladder of ways of expression, which includes

$$y^3$$
$$y^2$$
$$y$$
$$\sqrt{y}$$
$$\log y$$
$$-\frac{1}{\sqrt{y}}$$

$$-\frac{1}{y}$$

$$-\frac{1}{y^2}$$

$$-\frac{1}{y^3}$$

than we just have. We stated one special case of the following:

◇ if one expression is straight, those above it are hollow upward, those below it are hollow downward.

Exhibit 19 of chapter 3 has already shown us that this is true when y is straight. Exhibit 20 of chapter 3 shows us that this is true when $\log y$ is straight. The appearance of both those exhibits makes it plausible that the statement is true if any expression of the simple ladder is straight. (The doubting reader should try to use the following two statements to construct a general proof.)

◇ if our three points are hollow UPward, we look further DOWN the ladder for straightness.

◇ if our three points are hollow DOWNward, we look further UP the ladder for straightness.

These must be true. Consider the first: If a new expression is to be straight, and our present expression is hollow upward, the present expression has to be higher up the ladder than the new expression. To find the new expression, we must move down the ladder from the present expression.

So far as re-expressing y goes, the rule is simple:

◇ move on the ladder as the bulging side of the curve points.

U.S. population again

Applying this to the early U.S. population--since we have seen that the curve is hollow upward--moves us down the ladder. Let us try this, trying $-1/y$ first.

Turning to exhibit 6 of chapter 3 we can find

$$-1/5.3 = -0.188$$
$$-1/23.2 = -0.043$$
$$-1/62.9 = -0.016$$

The three points and the two slopes are now

$$(1800, -0.188)$$
$$(1850, -0.043)$$
$$(1890, -0.016)$$

$$\frac{-0.043 - (-0.188)}{1850 - 1800} = \frac{0.145}{50} = 0.0029$$
$$\frac{-0.016 - (-0.043)}{1890 - 1850} = \frac{0.027}{40} = 0.0007$$

The slope for the first pair of points is now four times that for the second pair. When we used y, that for the second was three times that for the first. We would like equal slopes, so it is natural to try next about halfway from y to $-1/y$. Thus we are led to try $\log y$. Turning to exhibit 3 of chapter 3, we see that

$$\log 5.3 = 0.72$$
$$\log 23.2 = 1.37$$
$$\log 62.9 = 1.80$$

so that the three points and two slopes are

$$(1800, 0.72)$$
$$(1850, 1.37)$$
$$(1890, 1.80)$$

$$\frac{1.37 - 0.72}{1850 - 1800} = \frac{0.65}{50} = 0.013$$
$$\frac{1.80 - 1.37}{1890 - 1850} = \frac{0.43}{40} = 0.011$$

Now the slopes agree with each other rather well. We should now go ahead--calculating $\log y$ and plotting x and $\log y$ for either many more or all of the points. Exhibits 7 and 10 of chapter 5 have already shown us how well this choice works.

fitting lines to three points

Once we have our three points fairly well on a line, we may as well fit a line to them. The three points just fixed offer a reasonable example.

To fit a line to three reasonably-spaced points, we usually do well to fit a slope to the two endpoints, and then take the mean of the three adjusted

values to find the constant. For (1800, 0.72), (1850, 1.37), (1890, 1.80), this leads to

$$\frac{1.80 - 0.72}{1890 - 1800} = \frac{1.08}{90} = 0.012$$

and since it seems easy to work with

$$x = \text{date} - 1800$$

we form the three values of

$$y - .012(\text{date} - 1800)$$

namely

$$\begin{aligned} 0.72 - .012(0) &= 0.72 \\ 1.37 - .012(50) &= 0.77 \\ 1.80 - .012(90) &= 0.72 \end{aligned}$$

for which the mean is 0.74 to two decimals. (The agreement of the two 0.72's is an important check on our arithmetic.) Thus our fit is

$$\text{Population(log of millions)} = 0.74 + 0.012(\text{date} - 1800)$$

Three points can take us a long way. If they are well chosen, they can do very well for us.

review questions

How do three points indicate hollow upward? Hollow downward? In which direction are expressions hollow that fall above (on the ladder of ways of expression) a straight expression? Those below? How do these rules help us in straightening the growth of the U.S. population? How do we fit a line to three (reasonably spaced) points? Why did we not use the median of the three adjusted values? Should we be surprised that 0.012 (in the 3-point fit) falls between 0.011 and 0.013 (for the two pairs of points)?

6C. Re-expressing *x* alone

We now know how to be guided in choosing a better expression for *y*. What if we wish to leave *y* alone and re-express *x*?

If we flip our picture over, interchanging the *y*- and *x*-axes, we convert one problem into the other. (We have to look through the back side of the paper after the flip, but a straight line stays a straight line.) Since it is now the same problem, all the same arguments apply.

This means that if the curve bulges toward large x and we are to re-express x, we ought to move x up the ladder, while if the curve bulges toward small x, we should move x down the ladder.

back to the population of the U.S.

Exhibit **1** shows three points from the population curve after the interchange of x and y. We see that the bulge is toward larger x (upward on this interchanged plot), so if we are to re-express x we would have to move upward on the x ladder.

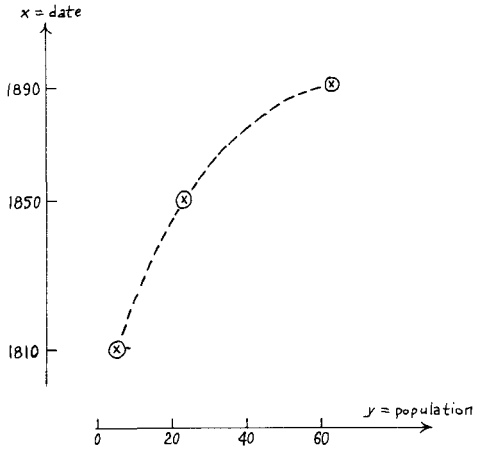
Should we try re-expressing x in this situation? No, because we can see that re-expression is unlikely to help us. As amounts, 1810, 1850, and 1890 are very similar, the outer values differing only a few percent from the middle one.

Re-expressing a variable that changes by only a few percent rarely gets rid of more than a barely detectable curvature.

One way to make the x's less alike as amounts would be to figure dates from some origin later than the birth of Christ. To figure from 1776 would be quite unrealistic, since there had been much immigration before that date. To figure from 1600 would be more realistic, but even this would be neglecting a substantial Amerind population.

exhibit **1** of chapter 6: U.S. population

The early population of the U.S., *x* and *y* interchanged



Let us try "date − 1600" anyway. An example should be useful, so long as it is not quite foolish. To keep our numbers of convenient size, we may as well measure "date − 1600" in centuries.

Our new x values for the last 3 points are 2.1, 2.5, and 2.9. They now vary by about ±15%, which is much more than before. We have some hope, but we cannot expect to find it easy to take care of substantial curvature, such as we still face. Since the change from date in years to [(1/100)(date) − 16] is a trivial re-expression--one that involves only multiplication by, and addition of, constants, our need to move up the x ladder is unchanged.

Let us begin by going to the cube of the new x. We find

$$(2.1)^3 = 9.261$$
$$(2.5)^3 = 15.625$$
$$(2.9)^3 = 24.389$$

The three points are:

$$(9.3, 7.2)$$
$$(15.6, 23.2)$$
$$(24.4, 62.9)$$

and the two slopes are:

$$3.2 = \frac{23.2 - 7.2}{15.6 - 9.3}$$
$$4.5 = \frac{62.9 - 23.2}{24.4 - 15.6}$$

The slope is still larger for the second interval, though only by about 2 to 1 rather than by 3 to 1. We have made progress, but clearly need to go further still.

Trying x^6, which is easy to find by squaring x^3, gives

$$(85, 7.2)$$
$$(244, 23.2)$$
$$(501, 62.9)$$

and

$$.102 = \frac{23.2 - 7.2}{244 - 85}$$
$$.114 = \frac{62.9 - 23.2}{591 - 244}$$

This is a lot closer, but we are not there yet.

Trying x^8 gives

$$\begin{aligned}&(378,\ \ 7.2)\\&(1526, 23.2)\\&(5002, 62.9)\end{aligned}$$

and

$$.0139 = \frac{23.2 - 7.2}{1526 - 378}$$

$$.0114 = \frac{62.9 - 23.2}{5002 - 1526}$$

We seem at last to have gone too far.

Trying x^7 should come quite close. We find

$$\begin{aligned}&(180,\ \ 7.2)\\&(610, 23.2)\\&(1725, 62.9)\end{aligned}$$

$$.0372 = \frac{23.2 - 7.2}{610 - 180}$$

$$.0357 = \frac{62.9 - 23.2}{1725 - 610}$$

Agreement is now fairly good.

If we are to re-express

$$(\text{year} - 1600),$$

we find that our three points suggest the use of either

$$(\text{year} - 1600)^7$$

or, equivalently,

$$\left(\frac{\text{year} - 1600}{100}\right)^7.$$

We ought at least look at the results of doing this. Exhibit **2** shows the numbers, exhibit **3** the gross picture, exhibit **4** some differences. As either exhibit 2 or exhibit 4 shows, the deviations of the U.S. population, expressed in millions, from

$$0.72 + 0.036\left(\frac{\text{date} - 1600}{100}\right)^7$$

are less than 150,000 from 1800 to 1830 and from 1870 to 1890, with an "extra" million or two counted in 1840, 1850, and 1860.

exhibit **2** of chapter 6: U.S. population

The results of re-expressing x in dealing with U.S. population in the nineteenth century

Date	z $\frac{(\text{date} - 1600)^7}{100}$	y Population (in millions)	$.036z$	Diff.
1800	128	5.31	4.61	0.70
1810	180	7.24	6.48	0.76
1820	249	9.64	8.96	0.68
1830	340	12.87	12.24	0.63
1840	459	17.07	16.52	1.55
1850	610	23.19	21.96	1.23
1860	803	31.44	28.91	2.53
1870	1046	38.58	37.66	0.92
1880	1349	50.16	48.56	1.60
1890	1725	62.95	62.10	0.85
1900	2187	76.0	78.7	−2.7
1910	2751	92.0	99.0	−7.0
1920	3436	105.7	123.7	−18.0

exhibit **3** of chapter 6: U.S. population

The early U.S. population plotted against $z = \left(\frac{\text{date} - 1600}{100}\right)^7$

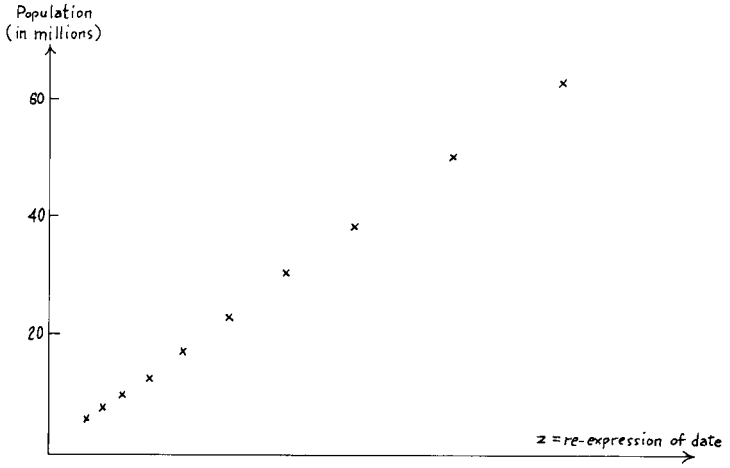
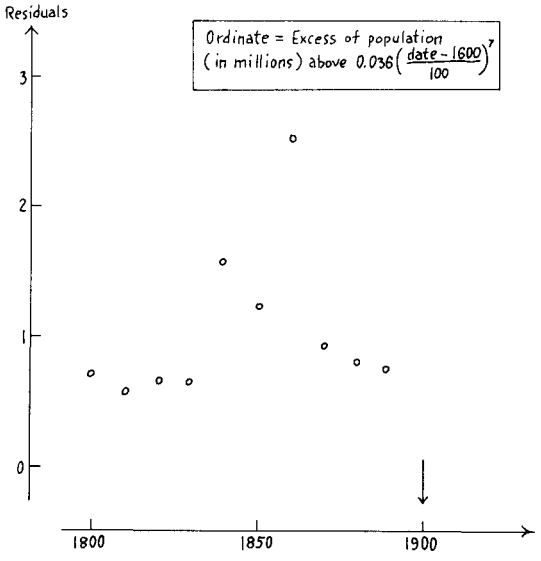


exhibit **4** of chapter 6: U.S. population

The result of flattening exhibit 3



a caveat

So far, either as quality of fit or ability to set residuals in clear view goes, plots such as exhibits 3 and 4 are as good as the plots of the last chapter, which related log population to date in years. If our purpose is to examine residuals--as it so often is--either plot is effective and useful.

When then should we make a distinction between the two plots? Surely we do need to make a distinction when we want to find an easily communicated description. That the U.S. population grew about 2.8% per year from 1800 to 1890 is relatively easy both to communicate and to understand. This is a major advantage, since we cannot say the same of the y vs. x^7 relationship. (Both, of course, apply over a limited span of years. Both fail--faster and faster--as we move on beyond 1900.) For communication, there is no doubt that log y vs. x is a more useful description.

Like any good fit, either the log y vs. x or the y vs. x^7 plot is subject to dangers of overvaluation. We see that each fits closely, though they cannot be exactly alike. Particularly if we have found only one of the two, there is a very natural tendency to convert "a good fit" into "this is how it had to be" or "a basic law of population growth". **One example of a close fit, by itself, is far**

from representing evidence for such strong statements. The fact that we have found two close fits of quite different form emphasizes our need to learn to avoid this sort of jumping at conclusions. **Conversely, we can make many good uses of a close fit, whether or not it is "a basic law".**

review questions

If we are to re-express x, which way ought we move on the ladder? Why must this be so? What happens if we exchange x and y axes in the U.S. population example? If x varies by only a few percent, what then? What trials did we make in re-expressing dates? Do we expect only one choice of re-expression to straighten a given set of points out thoroughly? What if we find several? Can we infer a "basic law" from one close fit? Why is one straightening of U.S. population growth easier to communicate than the other?

6D. A braking example

Let us next look at an example where re-expressing x seems to be the natural way to make the data more orderly and more describable.

Exhibit **5** shows the speed and distance to stop for 50 cars. Exhibit **6** plots the data. We could fit most of the data points with a straight line. However, a fitted line would give zero stopping distance at a speed between 5 and 10 miles per hour. The one thing we are sure of in this example is that zero stopping distance goes with zero speed, and vice versa. We didn't have to test cars to know that--or to put a point at (0, 0). We must face curvature, and try to eliminate it.

Three reasonable points to take are, then,

$$(0, 0)$$

and the two marked by large x's in exhibit 6,

$$(15, 35) \quad \text{and} \quad (25, 90).$$

The two slopes are

$$\frac{35 - 0}{15 - 0} = 2.3 \quad \text{and} \quad \frac{90 - 35}{25 - 15} = 5.5.$$

To re-express x, since the curve bulges toward large x, we ought to move toward x^2, x^3, etc. Trying x^2 gives

$$(0, 0) \quad (225, 35) \quad (625, 90)$$

with slopes of

$$.15 = \frac{35 - 0}{225 - 0} \quad \text{and} \quad .14 = \frac{90 - 35}{625 - 225}$$

Clearly, using x^2 is a reasonable try.

Exhibit **7** shows the plot of y against x^2, which now looks quite straight. If we take

$$(0, 0) \quad \text{and} \quad (600, 80)$$

as representative points, we are led to try

$$y - .133x^2$$

as a reasonable flattened quantity.

Exhibit **8** shows the result. It is far from wonderful, but seems to be reasonably flat, although the behavior for x^2 near 0 does not fit too well with the known point at (0, 0).

using our knowledge

We need to try something further. We believe in $y = 0$ when $x = 0$. Perhaps we should use this belief in choosing what to plot. How can we do this?

exhibit **5** of chapter 6: braking distances

Speed and distance to stop

Speed, *x* (mph)	Distance to stop, *y* (in feet)
4	2, 10,
7	4, 22,
8	16,
9	10,
10	18, 26, 34,
11	17, 28,
12	14, 20, 24, 28,
13	26, 34, 34, 46,
14	26, 36, 60, 80,
15	20, 26, 54,
16	32, 40,
17	32, 40, 50
18	42, 56, 76, 84,
19	36, 46, 68,
20	32, 48, 52, 56, 64,
(21)	
22	66,
23	54,
24	70, 92, 93, 120,
25	85,

S) SOURCE

Mordecai Ezekiel, 1930. Methods of Correlation Analysis. New York, John Wiley. Table 11, page 41.

Also Table 10, page 43, of 2nd edition 1943. Table 4.1, page 45 of 3rd edition 1959, by Mordecai Ezekiel and Karl A. Fox, has similar but different data.

exhibit **6** of chapter 6: braking distances

Plot of exhibit 5s data

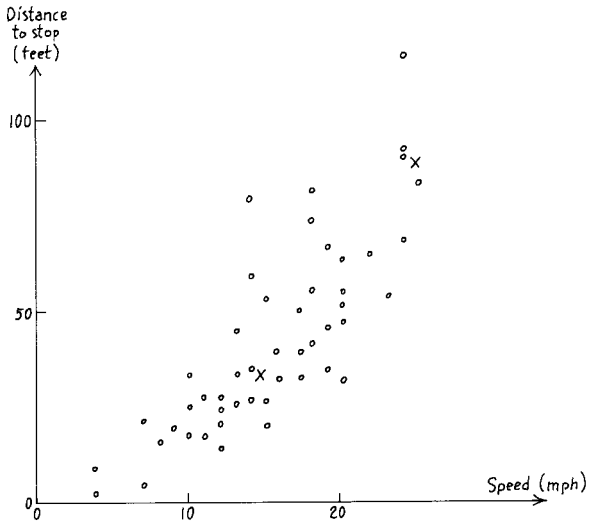
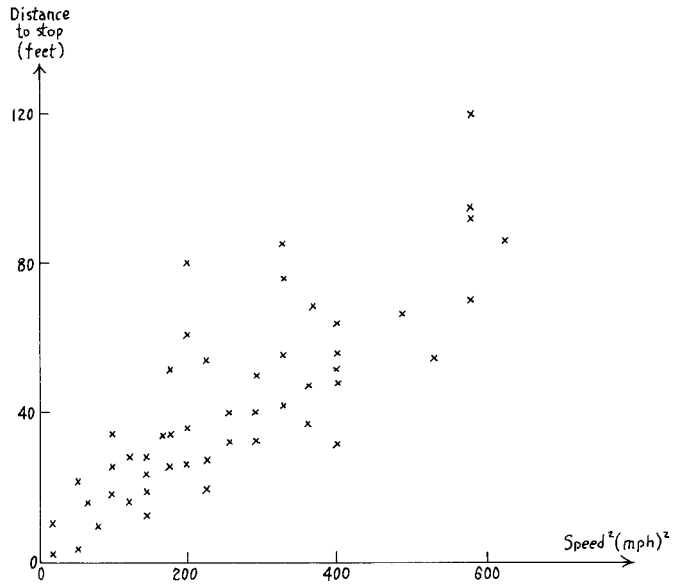


exhibit **7** of chapter 6: braking distances

Speed^2 and distance to stop for 50 motor cars



One way is to plot y/x--rather than y--against x. Any finite value for y/x will make y go to zero as x goes to zero. Thus any reasonable behavior of our y/x vs. x plot will lead to a fit that makes $x = 0$ go with $y = 0$.

Exhibit **9** gives the values of y/x. Exhibit **10** plots y/x against x. The behavior of the point cloud is quite reasonable, so we select two representative points as shown. They are

$$(5, 1.4) \quad \text{and} \quad (25, 3.7).$$

The corresponding line is

$$\frac{y}{x} = .115x + .8.$$

The residuals are written down in exhibit **11** and plotted in exhibit **12**.

We now see:

- ◇ 10 wandering points on the high side--9 between 1.1 and 1.9, one very high--these presumably represent bad brakes or slow-responding drivers.
- ◇ a mass of other points which, if taken by themselves, seem quite level but placed about 0.5 too low.

exhibit **8** of chapter 6: braking distances

Flattening by use of $y - 1.33x^2$

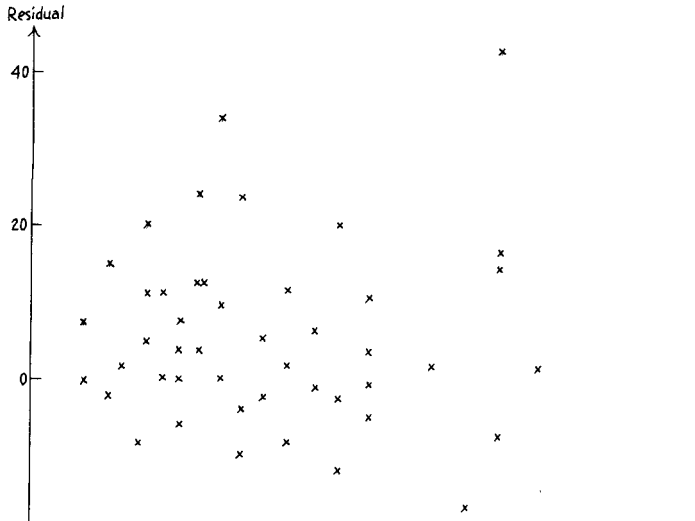


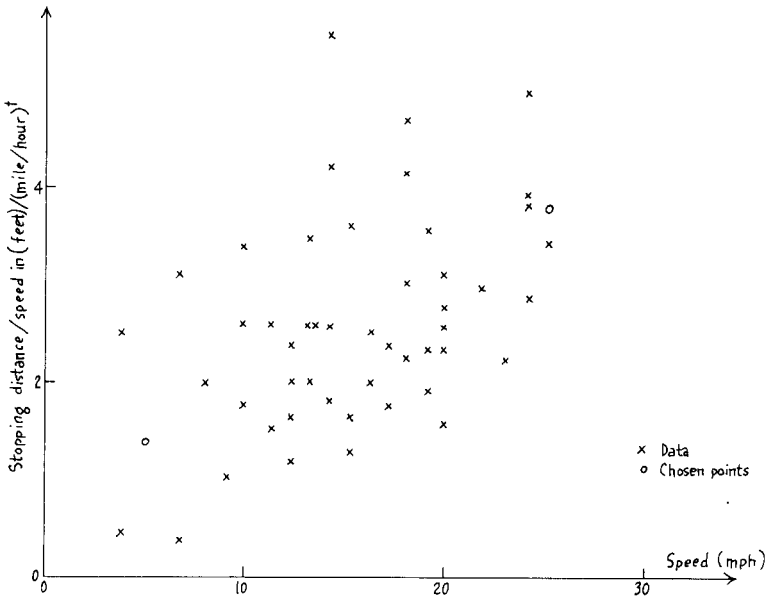
exhibit **9** of chapter 6: braking distances

The values of y/x

x	y/x
4	0.5, 2.5
7	0.6, 3.1
8	2.0
9	1.1
10	1.8, 2.6, 3.4
11	1.6, 2.6
12	1.2, 1.7, 2.0, 2.3
13	2.0, 2.6, 2.6, 3.5
14	1.9, 2.6, 4.3, 5.7
15	1.3, 1.7, 3.6
16	2.0, 2.5
17	1.9, 2.4, 3.0
18	2.3, 3.1, 4.2, 4.7
19	1.9, 2.4, 3.6
20	1.6, 2.4, 2.6, 2.8, 3.2
(21)	
22	3.0
23	2.3
24	2.9, 3.8, 3.9, 5.0
25	3.4

exhibit **10** of chapter 6: braking distances

y/x against x



† 1 foot/(mile/hour) = $\frac{30}{44}$ seconds = $\frac{15}{22}$ second

Thus, our final description can run as follows:

40 points fairly well described by:

$$\frac{y}{x} = .115x + .8 - .5$$

or

$$y = .115x^2 + .3x,$$

accompanied by 9 points with y greater than this expression by about $2x$ and one point greater by about $4x$.

Careful and repeated analysis can lead to effective description.

review questions

What example did we use in this section? What fact were we entitled to be sure of? How does exhibit 6 behave? What three points is it natural to choose? What re-expression do they lead to? How effective does this re-expression seem to be? What did we then think of trying? Why was it natural to think of it? How did the resulting plots look? Do there seem to be stray values? Are you surprised?

exhibit **11** of chapter 6: braking distances

The values of $(y/x) - 0.115x - 0.8$

x	$0.115x + 0.8$	$y/x - 0.115x - 0.8$
4	1.3	−0.8, 1.2,
7	1.6	−1.2, 1.5,
8	1.7	0.3,
9	1.8	−0.7,
10	2.0	−0.2, 0.6, 1.4,
11	2.1	−0.5, 0.5,
12	2.2	−1.0, −0.5, −0.2, 0.1,
13	2.3	−0.3, 0.3, 0.3, 1.2,
14	2.4	−0.5, 0.2, 1.9, 3.3,
15	2.5	−1.2, −0.8, 1.1,
16	2.6	−0.6, −0.1,
17	2.8	−0.9, −0.4, 0.2,
18	2.9	−0.6, 0.2, 1.3, 1.8,
19	3.0	−1.1, −0.6, 0.6,
20	3.1	−1.5, −0.7, −0.5, −0.3, 0.1,
(21)		
22	3.3	−0.3
23	3.4	−1.1
24	3.6	−0.7, 0.2, 0.3, 1.4
25	3.7	−0.3

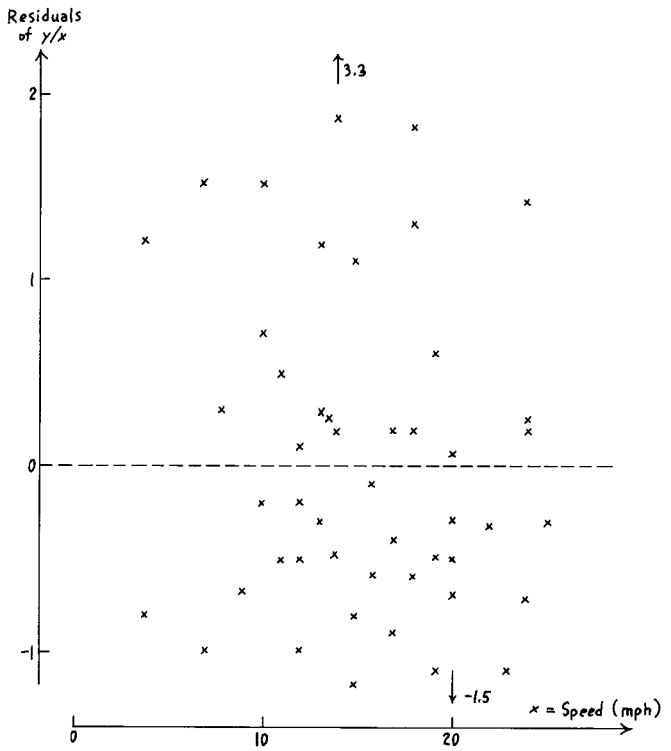
6E. The vapor pressure of H_2O

The vapor pressure of water or ice is the highest pressure of water vapor (steam, if you like) that can exist in equilibrium with the water or ice at a given temperature. Its values are well known, and of considerable practical importance. In terms of temperatures Celsius (once called centigrade, giving 0°C = 32°F and 100°C = 212°F) and pressures in mm of mercury (760 mm is one standard atmosphere), these vapor pressures at different temperatures are given in exhibit **13.**

Just a look at the first two columns of this table shows that a direct plot of p against t will show us only a very rapidly rising pressure. (If 139893.20 can be plotted, even 760.00 can hardly be seen.) Accordingly, we plan to do something to at least partly straighten out this plot. Taking the logarithm of p looks as if it might help.

exhibit **12** of chapter 6: braking distances

Plot of $(y/x) - 0.115x' - 0.8$ against x



Column (3) of exhibit 13 contains the values of log p. This looks conceivably plottable, so we proceed to exhibit **14,** which can be looked over, but is far from straight.

Another step needs to be taken. We could try to find some empirical approach, but we can go more directly to our goal if we recall that plotting log pressure against the reciprocal of the absolute temperature is rather common in physical chemistry. (Or if we use the closing example of the next section.) Here, one kind of absolute temperature is the Celsius temperature increased by about 273.1°C.

Column (4) of exhibit 13 contains values of $-1000(1/T)$, where $T = t + 273.1$°C. (The factor 1000 has been used to avoid unsightly and confusing zeros.) The result of plotting log p against column (4) of exhibit 13 is shown in exhibit **15.** At last we have a reasonably good straight line, one worth calculating residuals about.

Column (5) of exhibit 13 contains values of $\log p - 2.25(-1000/T)$ which, as we see, is reasonably constant. Subtracting 8.8 clearly gives residuals well

exhibit **13** of chapter 6: vapor pressure

The vapor pressure of H_2O and some associated quantities

t(°C)	p(mm Hg)	log p	(4)	(5)
−40	0.105	−.9788	−4.2900	8.6737
−20	0.787	−.1040	−3.9510	8.7858
0	4.5687	.6598	−3.6617	8.8986
20	17.363	1.2396	−3.4118	8.9162
40	54.865	1.7393	−3.1939	8.9256
60	148.88	2.1728	−3.0021	8.9275
80	354.87	2.5501	−2.8321	8.9223
100	760.00	2.8808	−2.6802	8.9113
120	1489.14	3.1729	−2.5439	8.8966
140	2710.92	3.4331	−2.4207	8.8797
160	4636.00	3.6661	−2.3089	8.8611
180	7520.20	3.8762	−2.2070	8.8420
200	11659.16	4.0667	−2.1137	8.8225
220	17395.64	4.2404	−2.0280	8.8034
240	25100.52	4.3397	−1.9489	8.7847
260	35188.00	4.5464	−1.8758	8.7670
280	48104.20	4.6822	−1.8080	8.7502
300	64432.80	4.8091	−1.7449	8.7351
320	84686.80	4.9278	−1.6861	8.7215
340	109592.00	5.0398	−1.6311	8.7097
360	139893.20	5.1458	−1.5795	8.6997

(4) = $-1000(1/T)$, where $T = t + 273.1$°C
(5) = $\log p - 2.25(-1000/T)$

exhibit **14** of chapter 6: vapor pressure

Vapor pressure of H_2O and temperature (logarithmic scale for pressure, which is in mm Hg)

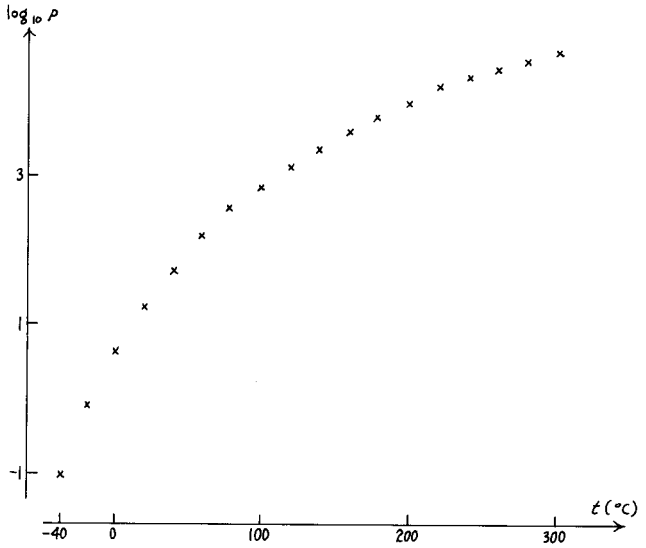
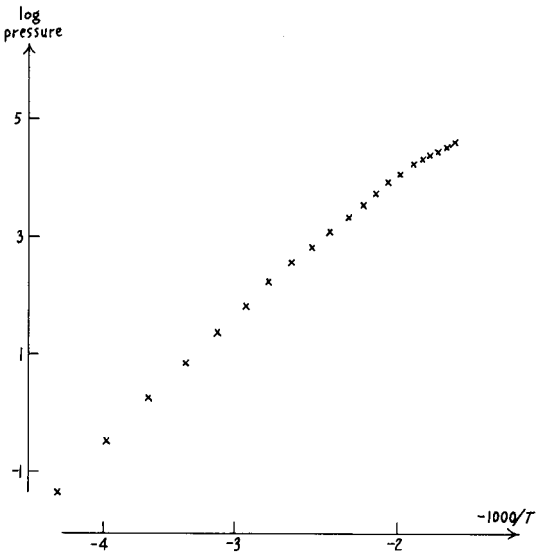


exhibit **15** of chapter 6: vapor pressure

Vapor pressure of H_2O and temperature (logarithm of pressure, reciprocal of absolute temperature)



worth plotting. Exhibit **16** shows the results. The most striking result is the apparent split of our curve into two parts with a corner at the third point from the left. What might this mean?

When we recall that this point comes at 0°C (= 32°F), which is the freezing point of water, we see that a break at this point is quite reasonable. Below 0°C we are dealing with the vapor pressure of solid water (ice), while above this point we are dealing with the vapor pressure of liquid water.

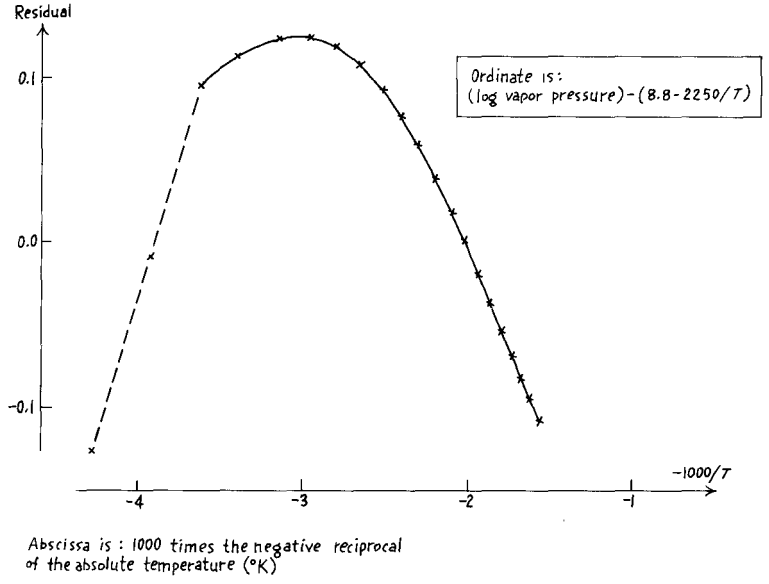
Note that this fact was not forced upon us by our data until we both straightened and flattened the plot.

review questions

What data did we look at in this section? Could we plot it at all in its original form? Did we re-express y? Why? Was this enough? Could we go further? How? How well did we then do? Could we see anything new when we looked at the residuals? Did such a look start off any ideas?

exhibit **16** of chapter 6: vapor pressure

Residuals of log vapor pressure from a straight line in $-1/T$ (water)



6F. Re-expressing the second variable

As we have just seen, it will not always be wise--or satisfactory--to change only one of the two expressions. Sometimes it pays to re-express **both** x and y. In such a situation we may need to do "cut and try". However, once we have agreed to try a specific expression for one of the variables, be it x or y, we can use the same rules to narrow down our choice of the second expression.

As an example, let us take three points from the example just treated in the last section, and see how we might have been guided. From exhibit 13 we have

$$(0, \quad 4.5687)$$
$$(100, \quad 760.00)$$
$$(200, \ 11659.16).$$

Clearly, the middle point is below the line that joins the other two--and also to the right of that line. If we are to re-express y we should look toward log y and $-1/y$. Trying logs gives

$$(0, 0.6598) \qquad (100, 2.8808) \qquad (200, 4.0667)$$

with slopes of

$$\frac{2.8808 - 0.6598}{100 - 0} = \frac{2.2210}{100} = .02221$$

and

$$\frac{4.0667 - 2.8808}{200 - 100} = \frac{1.1859}{100} = .01186.$$

Thus the middle point is now above and to the left of the line. We could try going only as far as $\sqrt{y}$, which is about halfway to log y. If we do this, say by using exhibit 5 of chapter 3, we find

$$(0, \quad 2.12)$$
$$(100, \ 27.6)$$
$$(200, \ 108)$$

and we see that $\sqrt{y}$ goes nowhere nearly far enough. To go on to

$$\sqrt{\sqrt{y}} = y^{1/4} = \sqrt[4]{y}$$

is easy. If we do this, we find

$$(0, \quad 1.44)$$
$$(100, \ 5.3)$$
$$(200, \ 10.4)$$

for which the slopes are

$$\frac{5.3 - 1.44}{100 - 0} = \frac{3.86}{100} = .0386$$

$$\frac{10.4 - 5.3}{200 - 100} = \frac{5.1}{100} = .0510$$

so that we see that we are not yet far enough. Those who love to repeat square roots will now try

$$\sqrt{\sqrt{y}} = y^{1/8} = \sqrt[8]{y},$$

finding

$$(0, 1.20)$$
$$(100, 2.32)$$
$$(200, 3.20)$$

for which the slopes are

$$\frac{2.32 - 1.20}{100 - 0} = \frac{1.12}{100} = .0112$$

$$\frac{3.20 - 2.32}{200 - 100} = \frac{.88}{100} = .0088.$$

This time we have gone too far. As a result, if we are to change only y, we should try something between $y^{1/4}$ and $y^{1/8}$, perhaps $y^{1/6}$ or $y^{1/7}$.

another try

This is, however, not our only reasonable choice. Going to log y did much to straighten out our three points. True, it went a little too far, but perhaps we could do something by keeping log y and re-expressing x. Before we do this, however, we should stop and think for a moment. As we have written it, x is temperature in °C (freezing water = 0°C, boiling water = 100°C). If we are to re-express simply, we ought not to tie our zero to a property of so special a substance as water. (We are looking at water's vapor pressure, but water has a vapor pressure below 0°C, too.) Rather we should figure our temperatures so that we run from the so-called absolute zero, which is just below −273°C.

Our intermediate starting point then should be

$$(273.1, 0.6598)$$
$$(373.1, 2.8808)$$
$$(473.1, 4.0667)$$

with slopes of .02221 and .01186, and a middle point above and to the left of the line. If we are to re-express x, we should move toward $\log x$ or $-1/x$. Let us try $\log x$. We find:

$$(2.44, 0.6598)$$
$$(2.57, 2.8808)$$
$$(2.67, 4.0667)$$

$$\frac{2.8808 - 0.6598}{2.57 - 2.44} = \frac{2.2210}{.13} = 17.1$$
$$\frac{4.0667 - 2.8808}{2.67 - 2.57} = \frac{1.1859}{.10} = 11.8$$

Using logs has helped, but not enough. Let us try $-1/x$, for which we have

$$(-.00366, 0.6598)$$
$$(-.00268, 2.8808)$$
$$(-.00211, 4.0667)$$

$$\frac{2.8808 - 0.6598}{-.00268 - (-.00366)} = \frac{2.221}{.00098} = 2266$$
$$\frac{4.0667 - 2.8808}{-.00211 - (-.00268)} = \frac{1.186}{.00057} = 2080$$

for which the slopes agree better--have a ratio nearer 1--than with any other combination so far tried. Accordingly

$$\log y \quad \text{and} \quad -1/x$$

where x is absolute temperature, seems to be a good choice. As we know from the last section, it proves to be one.

review questions

Does it ever pay to re-express both x and y? Do we have to do this blind? How is guidance for re-expressing the second coordinate to be found? How much is it like guidance when only one coordinate is re-expressed? If we tackle the vapor pressure of water, re-expressing only y, what are we led to? If we pick $\log y$, what are we led to for x?

6G. Wise change of origin as a preliminary

We have already seen two examples where a change of origin was part of a sensible approach to straightening: (1) the early population of the U.S. as a function of time (where a time origin at the birth of Christ--1600 years

before the beginning of European immigration to North America--was far from sensible--once we thought about it! Why 1600 and not 600 or 2600?) and (2) the vapor pressure of water as a function of centigrade temperature. (Why should we take the origin at freezing?)

The same sort of need to think about a sensible origin arises in other problems, sometimes with regard to x and sometimes with regard to y. In one broad class of cases, the question arises because the amount we measure is the sum of contributions, some of which change slowly if at all, while the remaining contribution changes in such a way that re-expression could flatten the plot if this contribution could be measured separately.

The natural approach to this sort of data is to form

observation MINUS background

where "background" is a constant chosen to allow for the slowly changing components. We are not likely to be able to pick the value for "background" either on the basis of general insight or on the basis of doubtfully-related historical data. We expect to learn about it from the same set of data that we are trying to flatten. We are likely to approach choosing a plausible value for "background" by trying various values for it and seeing which one leads to data that can be more thoroughly flattened by further re-expression.

Observation of radioactive decay yields many problems of this sort. Any single kind of radioactive atom decays on a steady percentage basis--so many percent is gone every so many days, years, or millennia. Many processes of isolating--or making--some one kind of radioactive atom also isolate--or make--one or more other kinds. If these other kinds decay--but decay more slowly--their presence can often be adequately allowed for by a constant background.

a radioactive decay example

In 1905, the study of radioactive substances was in its infancy. Meyer and von Schweidler reported the relative activities for an experimental object set out in exhibit **17.** If we plot activity against time, the result is exhibit **18,** which is far from being straight. It is again natural to take logarithms, especially since, in simple radioactive decay, the logarithm of activity should decrease linearly with time.

Exhibit **19** shows the behavior of log activity against time. A noticeable curvature remains. In this situation the most plausible source of curvature is contamination by some other radioactive substance that decays much more slowly than the one of central interest. If such there be, it will not have contributed more than two units of activity (since at 45 days the observed activity is down to 2.1 units).

It is reasonable to explore the consequences of assuming the presence of 1.0 to 1.5 units of activity from such a contaminant by, successively:

◇plotting against time either

$$\log(\text{activity} - 1.0)$$

or

$$\log(\text{activity} - 1.5)$$

which could represent the log of the rapidly-decaying activity (and which are given in columns (4) and 5) of exhibit 17).

◇fitting a straight line, perhaps roughly.

◇plotting the residuals.

exhibit **17** of chapter 6: radioactivity

The decay of radioactivity and associated quantities

time (days)	activity (relative)	(3)	(4)	(5)	(6)	(7)
0.2	36.0	1.556	1.544	1.538	1.551	1.545
2.2	26.0	1.415	1.398	1.389	1.471	1.466
4.0	23.1	1.364	1.344	1.334	1.476	1.474
5	18.9	1.276	1.253	1.241	1.418	1.416
6	17.8	1.250	1.225	1.212	1.423	1.422
8	14.7	1.167	1.137	1.121	1.401	1.401
11	13.4	1.127	1.093	1.076	1.456	1.461
12	11.3	1.053	1.013	.991	1.409	1.411
15	8.5	.929	.875	.845	1.370	1.370
18	5.9	.771	.690	.643	1.284	1.273
26	5.0	.699	.602	.544	1.460	1.454
33	3.4	.531	.380	.279	1.469	1.434
39	2.4	.380	.146	−.046	1.433	1.319
45	2.1	.322	.041	−.222	1.526	1.353

(3) = log(activity)
(4) = log(activity − 1.0)
(5) = log(activity − 1.5)
(6) = .033t + (column 4)
(7) = .035t + (column 5)

S) SOURCE

S. Meyer and E. von Schweidler, 1905. Sitzungsberichte der Akademie der Wissenschaften zu Wien, Mathematisch–Naturwissenschaftliche Classe, p. 1202 (Table 5).

exhibit **18** of chapter 6: radioactivity

The data of exhibit 17

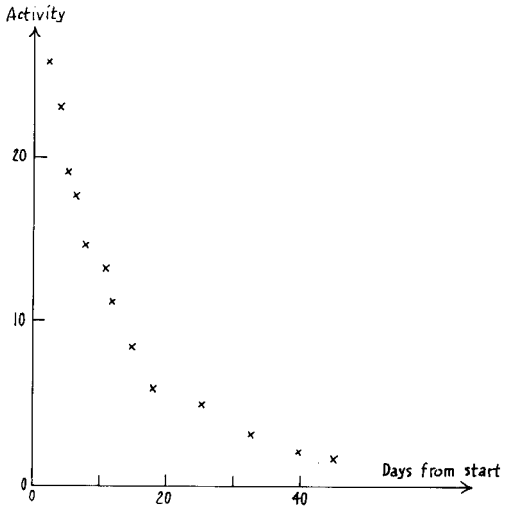
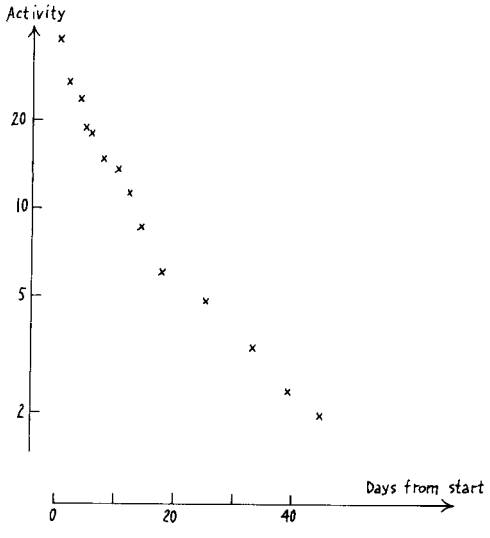


exhibit **19** of chapter 6: radioactivity

The same data on a log scale



The results are shown in exhibit **20.** (Columns (6) and (7) of exhibit 17 give these residuals increased by 1.45, calculated as ".033t + (column 4)" and ".035t + (column 5)", respectively.) Of the two choices, allowance for 1.0 unit of contamination seems to give a more nearly horizontal set of residuals--to lead to a closer fit.

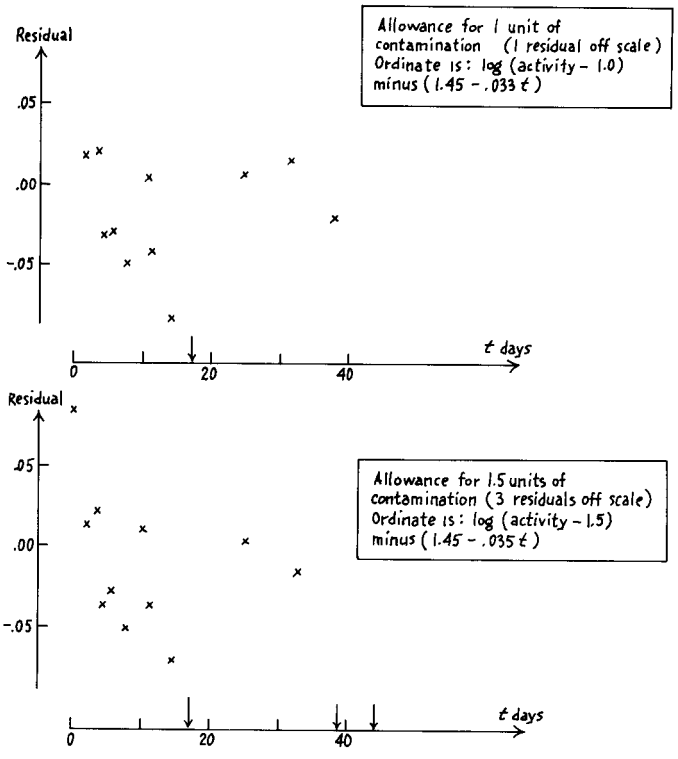
(Perhaps the residuals for 1.0 do trend upward a little. The reader may wish to try 1.1 or 1.2.)

review questions

Ought we expect to make changes of origin? What are three examples? What is a background? When we measure radioactive substances, whether made or isolated, are backgrounds common? What did a plot of raw activity against raw time show? What two things did we do then? Did it all work out well?

exhibit **20** of chapter 6: radioactivity

Residuals from decay line allowing for long-lived contamination (two versions)



6H. How far have we come?

This chapter has been devoted to guidance about how to re-express x and y with a view to straightening a plot. This problem arises most simply when one or both of x and y is an amount--or, really just a special case of an amount, a large count. The natural re-expressions are by powers, roots, and logs.

To keep computation down, we routinely begin with three well-chosen points. The basic rule of thumb is then simple:

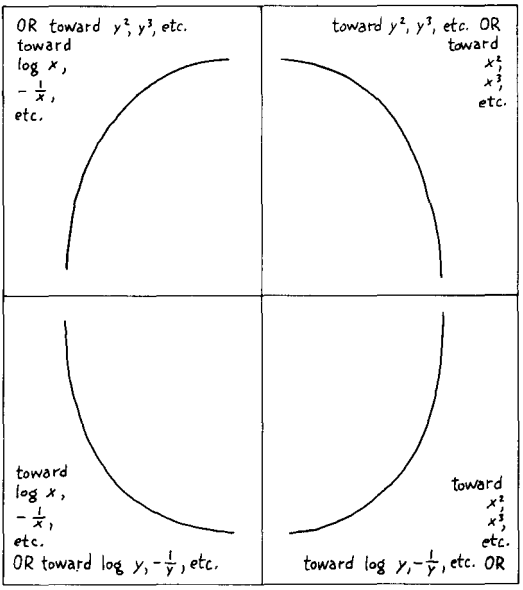
◇ **move on the ladder of expressions in the direction in which the curve bulges.**

Exhibit **21** shows the four possible cases, and the natural steps along both x- and y-ladders. (We can move along either or both.) We try new expressions on more of the data only when the three points have already responded well to them.

In certain cases, a fresh choice of origin--before going to powers, roots, or logs--is valuable. Sometimes the new choice is a matter of common sense, sometimes of how flat the final result proves to be.

exhibit **21** of chapter 6: indicated behavior

How to move for each variable separately; the four cases of curve shape



As we use this technique, we will need to remember that:

◇ straightening by re-expressing x is not the same as re-expressing y.

◇ just because re-expression makes things quite straight, there is no guarantee that we have found a new natural law.

We are now ready:

◇ to approach the analysis of (x, y) data in terms of re-expression followed by straight-line fitting (perhaps in two or more steps) and the examination of residuals.

◇ to turn to another class of important problems.

6P. Additional problems

See exhibits **22 through 26**.

exhibit **22** of chapter 6: data and problems

Three examples of radioactive decay

A) DATA

Days*	Activity†	Days*	Activity†	Days*	Activity†
0.8	6.70	0.8	2.82	0.8	2.05
2.8	6.40	2.8	2.34	1.0	2.03
6.9	5.70	4.8	1.90	1.8	1.79
8.9	5.10	6.8	1.80	2.1	1.77
13.1	4.30	11.8	1.34	3.9	1.54
15.2	4.00	13.9	1.24	5.9	1.35
16.8	3.95	16.8	1.03	7.1	1.29
20.1	3.40	19.8	1.00	9.1	1.23
20.8	3.40	23.8	0.80	12.1	1.01
20.9	3.20	31.8	0.55	12.9	0.96
21.9	3.20	32.5	0.11	16.9	0.82
31.1	2.42			19.9	0.68
36.8	2.30			22.8	0.59
43.8	2.11			25.9	0.51
49.8	2.00			33.8	0.40
53.8	1.99			44.1	0.32
58.8	1.98				
65.9	1.90				
73.8	1.80				
87.1	1.65				

* Since start.
† In volts/minute.

→

exhibit **22** of chapter 6 (continued)

P) PROBLEM

22a) In a later paper, Meyer and Von Schweidler reported the decay of radioactivity for three samples as in panel A. Analyze at least two of these. Comment.

S) SOURCE

S. Meyer and E. von Schweidler 1907. "Untersuchungen über radioaktive Substanzen. VIII Mitteilung: Über ein radioaktives Produkt aus dem Aktinium." Sber. Ak. Wiss. Wien. Math–Nat. Classe 116 IIA1, pp. 315–322 (especially 316–317).

exhibit **23** of chapter 6: data and problems

Vapor pressure of mercury

A) DATA

Temperature (°C)	Pressure (mm Hg)
0	0.0004
20	0.0013
40	0.006
60	0.03
80	0.09
100	0.28
120	0.8
140	1.85
160	4.4
180	9.2
200	18.3
220	33.7
240	59.
260	98.
280	156.
300	246.
320	371.
340	548.
360	790.

P) PROBLEMS

23a) The vapor pressure of mercury is stated to be as in panel A. Analyze graphically. Comment.

→

exhibit **23** of chapter 6 (continued)

23b) Given the three points found partway through section 6F,

(0, 0.6598)
(100, 2.8808)
(200, 4.0667),

what change in expression of this (once new) y will come close to straightening out these points?

23c) Apply the result of the last problem to the data of exhibit 13, fit a straight line, and plot the residuals. How do the results compare with exhibit 16?

23d) Apply the re-expression

y becomes $y^{1/6}$

(suggested by the analysis of section 6E) to the data of exhibit 13, fit a straight line, and find residuals. How do the results compare with exhibit 16? Complete the graphical analysis. Comment.

exhibit **24** of chapter 6: data and problems

More problems

24a) Rates of mortality from breast cancer in different latitudes. (50; 1025, 513) translates as: "In latitude 50°N the mortality index for breast cancer is 102.5 and the mean annual temperature is 51.3." Data-sets (16): (50; 1025, 513), (51; 1045, 499), 52; 1004, 500), (53; 959, 492), (54; 870, 485), (55; 950, 478), (56; 886, 473), (57; 892, 451), (58; 789, 463), (59; 846, 421), (60; 817, 442), (61; 722, 435), (62; 651, 423), (63; 681, 402), (69; 673, 318), (70; 525, 340). SOURCE: A. J. Lea, 1965, "New observations on distribution of neoplasms of female breast in certain European countries," *British Medical J.* **1** (for 1955): 486–490. Table II on page 489. PROBLEM: Can we straighten this plot? (Mortality index against mean annual temperature.)

24b) Plasticity of wool: slow stretching of a single fiber. (1, 321) translates as: "After 1 minute under load, a fiber 53.3 microns in average diameter (coefficient of variation of diameter = 5.4%) had stretched 32.1% beyond its unloaded length." Data-sets (34): (1, 321), (3, 330), (5, 334), (8, 337), (16, 342), (32, 348), (50, 352), (110, 361), (240, 377), (440, 394), (740, 413), (1310, 442), (1460, 449), (1630, 458), (1900, 469), (2090, 478), (2760, 499), (2950, 505), (3080, 509), (3460, 519), (4280, 540), (4970, 556), (5720, 572), (6000, 579), (6320, 586), (7120, 600), (7360, 604), (7540, 607), (8520, 623), (9020, 629), (9230, 633), (9950, 643), (10260, 647), (10680, 654). (The last data set is for 10,680 minutes under load and 65.4% stretch.) SOURCE: O. Ripa and J. B. Speakman, 1951. "The plasticity of wool." *Textile Research J.* **21**: 215–221. Table I on page 217. PROBLEM: How can we straighten the plot?

→

exhibit **24** of chapter 6 (continued)

24c) Amount of interstitial space in young chickens. ("Interstitial space" is a volume defined by the amount of thiocyanate ion taken up outside the blood within 10 minutes after injection and measured by its ratio to blood volume.) (1, 52) translates as: "For chickens 1 week old, the interstitial space accounted for 52% of body weight." Data-sets (8): (1, 52), (2, 42), (3, 39), (4, 38), (6, 37), (8, 36), (16, 25), (32, 22). SOURCE: W. Medway and M. R. Kare, 1959, "Thiocyanate space in growing fowl," *Amer. J. of Physiology* **196**: 873–875. Table 1 on page 874. PROBLEM: How can we straighten the plot?

24d) A more detailed look at interstitial volume (1, 55, 52) translates into: "At age 1 week, the average body weight of 6 chickens was 55 grams, of which the interstitial space was 52%." Data-sets (8): (1, 55, 52), (2, 108, 42), (3, 175, 39), (4, 242, 38), (6, 372, 37), (8, 527, 36), (16, 1137, 25), (32, 1760, 22). SOURCE: As for problem (24c). PROBLEM: How can we straighten the plot? How does the straightness here compare with that for problem (24c)? What expression of amount (not percent) of interstitial space in terms of weight corresponds to our fit?

exhibit **25** of chapter 6: data and problems

Some more problems

25a) Vapor pressure of a boron analog of mesitylene. (130, 29) translates as: "At a temperature of 13.0°C, the vapor pressure of B-trimethylborazole was 2.9 millimeters of mercury." Data/(sets) (13): (130, 29), (195, 51), (225, 85), (272, 103), (318, 146), (384, 213), (457, 305), (561, 514), (644, 745), (714, 1002), (805, 1437), (857, 1769), (915, 2169). SOURCE: E. Wiberg, K. Hertwig, and A. Bolz, 1948, "Zur kenntnis der beiden symmetrischen Trimethyl-borazole ("anorganisches Mesitylen")," *Zeitschrift für Anorganische Chemie* **256**: 177–216. Table on page 191. PROBLEM: How can we straighten this plot? Check your answer!

25b) Effects of small amounts of biotin on the mobility of a microorganism. (5E–7, 1354) translates as: "At a biotin concentration of 5×10^{-7}, the mobility of *Lactobacillus casei* was 1.354 units". Data-sets (7): (0, 1415), (5E-7, 1354), (1E-6, 1311), (5E-6, 1230), (1E-5, 1234), (5E-5, 1181), (1E-4, 1188). SOURCE: V. R. Williams and H. B. Williams, 1949, "Surface activity of biotin," *Journal of Biological Chemistry* **177**: 745–750. PROBLEM: How can we straighten the plot?

25c) Electric current produced by heating aluminum phosphate. (880, 1) translates as: "At a temperature of 880°C––an absolute temperature of 880 + 273 = 1153°K––positive electrification produced a current of 1 unit, where 1 unit = 2×10^{-9} amperes". Data-sets for run A (8): (880, 1), (950, 4), (970, 7), (995, 15), (1030, 35), (1030, 35), (1055, 49), (1110, 126). Data-sets for run B (9): (1036, 1), (1088, 8), (1135, 5), (1160, 8), (1195, 15), (1230, 34), (1245, 35), (1295, 74), (1330, 168). SOURCE: A. E. Garrett, 1910, "Positive electrification due to heating aluminium phosphate," (London, Edinburgh, and Dublin) *Philosophical Magazine* **20**: 571–591. Table on page 581. PROBLEM: How can we straighten the plots?

→

exhibit **25** of chapter 6 (continued)

25d) Growth-promoting effect of a purine for a deficient strain of red bread-mold. (0, 112) translates as: "When 0 moles of guanine per mole of adenine were included with 0.1 milligrams of adenine in 25 milliliters of basal medium, the dray weight of mycelium produced by a purine-deficinet strain of *Neurospora* was 11.2 milligrams". Data-sets (9): (0, 112), (0.25, 135), (0.59, 152), (0.75, 185), (1, 196), (1.5, 203), (2, 203), (2.5, 243), (3, 224). SOURCE: J. L. Fairley, Jr., and H. S. Loring, 1949, "Growth-promoting activities of guanine, guanosine, guanylic acid, and xanthine for a purine-deficient strain of *Neurospora*," *Journal of Biological Chemistry* **177**: 451–453. Table I on page 453. PROBLEM: How can we straighten the plot?

exhibit **26** of chapter 6: data and problems

Still more problems

26a) Measurement of a certain impurity in DDT; change of scale factor with temperature. (21, 248) translates as "At 21°C, the rate of crystallization (in microns per 5 minutes) is 24.8 times the log of the percent of this particular impurity." Data-sets (14): (21, 248), (22, 308), (23, 388), (24, 465), (25, 569), (26, 678), (27, 806), (28, 959), (29, 114), (30, 139), (31, 168), (32, 202), (33, 236), (34, 270). SOURCE: W. McCrone, A. Smedal, V. Gilpin, 1946, "Determination of 2,2, bis-p-chlorophenyl-1,1,1-trichloroethane in technical DDT: A microscopical method," *Industrial and Engineering Chemistry* **18**: 578–582. Table IV on page 582. PROBLEM: How can this plot be straightened?

26b) Demand deposits in post-office savings accounts in Switzerland. (37, 458) translates as: "In 1937, there were 458 million francs in post-office savings accounts". Data-sets (29): (37, 458), (38, 498), (39, 523), (40, 643), (41, 701), (42, 787), (43, 839), (44, 927), (45, 1001), (46, 1079), (47, 1007), (48, 1033), (49, 1090), (50, 1125), (51, 1212), (52, 1248), (53, 1334), (54, 1393), (55, 1443), (56, 1720), (57, 1720), (58, 1896), (59, 2050), (60, 2268), (61, 2643), (62, 3140), (63, 3353), (64, 3513), (65, 3810). SOURCE: *Swiss Statistical Abstract,* issued November 1969 by the Swiss Credit Bank. (Title also in French and German.) Tables at pages 24 and 25. PROBLEM: How can we best straighten a plot for this data? Make the fit and plot the residuals. Summarize all results.

26c) Revenue passenger miles on U.S. passenger airlines. (37, 412) translates as: "In 1937, there were 412,000 revenue passenger miles on U.S. domestic scheduled airlines". Data-sets (24): (37, 412), (38, 480), (39, 683), (40, 1052), (41, 1385), (42, 1418), (43, 1634), (44, 2178), (45, 3362), (46, 5948), (47, 6109), (48, 5981), (49, 6753), (50, 8003), (51, 10566), (52, 12528), (53, 14760), (54, 16769), (55, 19819), (56, 22362), (57, 25340), (58, 25343), (59, 28269), (60, 30514). (The last translates as: "In 1960, 30,514,000 revenue passenger miles".) SOURCE: Robert G. Brown, 1963. *Smoothing, Forecasting and Prediction of Discrete Time Series,* Prentice-Hall. Table C.7 on page 427. His source: *F.A.A. Statistical Handbook of Aviation.* PROBLEM: How can this plot be straightened? Plot residuals from a well-chosen straight-line fit.

26d) Find, from other sources, 2 batches of (x, y) data-sets that interest you and deserve to have their plots straightened by the methods of this chapter. Straighten the plots. Plot the residuals.