ANALYZING STRAIGHT LINE DATA[1]

FORMAN S. ACTON
National Bureau of Standards, Los Angeles, California

WHEN an engineer cautiously approaches a statistician, graph in hand, and asks how he should fit a straight line to these points, the situation is not unlike that moment when one's daughter inquires where babies come from. There is need for tact, there is need for delicacy, but here is opportunity for enlightenment and it must not be discarded casually—or destroyed with the glib answer.

Here is a man with a problem. He has asked a simple question to which there is no answer, and the statistician must take it upon himself to help find out the real questions to which answers are important, pertinent, and possible.

Why do people keep fitting straight lines to data? What is the fascination they hold for the experimental man? It certainly is not the aesthetic division of a plane into two halves—the delight of the geometer and analyst. The biometrician may wish to separate sheep from goats, but the chemist is happiest if his points fall on this boundary and separation is his farthest thought. Perhaps the line is valued because it is a simple two parameter curve—but then the sine wave is equally valuable—and any horizontal line becomes more important than either because it is simpler and possesses only one parameter. I think that for most physicists, chemists, and engineers, the straight line is valued chiefly because it can be extrapolated. Certainly no other curve can claim this property with the same degree of psychological confidence. A man feels much surer when he extrapolates a line than when he extrapolates any other curve—and in this day of bigger operations and larger productions, the engineer must extrapolate. The atomic bomb could not have been built without the straight line and econometricians in the Bureau of the Budget doubtless find it equally indispensable.

Extrapolation is merely one aspect of *prediction*, and here is one of the two important reasons that we fit curves to data. We desire to summarize our experimental experience in the form which may be used to predict the outcome of future similar experiments. This purpose the physical scientist shares with the biological and social scientists. But he differs from them in that he has another purpose more important than prediction: *the estimation of structure*. Our physical sciences have carefully chosen their definitions so as to invoke linear, or nearly linear relations between the various observable quantities, or simple transformations of these quantities. (If exponential behavior is expected, the logarithms are plotted, if pressure and temperature are examined, the reciprocal of the temperature is used; if Boyle's law is violated at heavy pressures, correction coefficients are introduced. By these artifices we retain the intuitional and extrapolatory advantages of the straight line for our machinations, and the statistician reaps the benefit in the simplicities of linear models.) These linearized physical hypotheses or laws contain structural parameters which the experimenter is trying to estimate. Perhaps it is the resistance of a piece of wire, to be found by applying measured voltages and recording the corresponding current; again it may be the angle of optical rotation of a chemical in solution as a function of its concentration, the temperature of the system, and wave length of the light. The system may be simple or complicated, but a theory specifies a functional form containing parameters and the experimental data are used to estimate these parameters. Ultimately we may wish to use this law for prediction, but frequently we merely wish to test our hypothesis that the exponent in the resistance law is *one* (*i. e.*, the current is a linear function of voltage), or some such simple value which the theory hopefully predicts. The experimenter's data contain errors which he probably regards as nuisances to be averaged out or eliminated by fitting a line—a line which is best in some sense. The only quantities of real interest to him are the slope and the intercept of this best line. This is the philosophy which bred the cult of Least Squares, and it is here that a guiding hand can effect the greatest good by pointing out other approaches to the same data which are apt to be more profitable.

Let us take an example. Suppose a physicist is trying to measure the effect of temperature on the fluorescence of a phosphor under a standard intensity of ultraviolet light. He has measured the intensity of the fluorescence while holding the phosphor temperature constant at each of several levels. The data, temperature versus intensity, are not really linear, and they scatter quite a bit. Perhaps a plot of intensity (or square root of intensity) versus the reciprocal of absolute temperature will linearize the data and perhaps it will not, according to the physical law involved, but the scatter has to be dealt with if any precision is to be obtained. The scatter arises in part from errors in measurement, but also from the fact that the light source is not constant and from the aging of the phosphor with oxidation by air or with exposure to ultra-

[1] The preparation of this paper was sponsored (in part) by the Office of Naval Research. It was presented at the annual meeting of The Institute for Mathematical Statistics, December 28, 1951.

violet light. The various errors introduced by the measuring instruments and the operator are lumped together.

I shall not attempt to be completely realistic here, but, if the intensity of the incident ultraviolet radiation is recorded, the uncontrollable fluctuations in this source may well cause enough variation in the final fluorescence to permit at least a crude estimation of the effect of intensity, and hence to permit its removal from the structural picture. If our physical friend's purpose is to estimate the parameters in the equation connecting intensity of fluorescence with temperature, then allowance for this residual variaton in the light source should greatly increase the precision of his estimates. If, on the other hand, his purpose is to calibrate this particular piece of phosphor so that he can use it as a thermometer in some inaccessible place, then the fluctuations of the light source are even more important, because they will also occur in the later experiments. Here he is interested in the *prediction* problem, and the effect of variations in the light source must be estimated in order to be allowed for realistically in the final apparatus. Undoubtedly a different source will be used later, and since its variability will be different from the one used in calibration, an estimation of the magnitude of this effect is essential. (Whether or not this estimation of intensity effects can be made here is dependent on the time constants of the phosphor as compared with the "period" of the light source fluctuations. If the phosphor responds much faster than the light varies, then the estimation of this effect is easy. If the phosphor is sluggish, then it will give an average effect over time, making its estimation difficult, and requiring subsidiary direct experiments on intensity if allowance is to be made for these fluctuations.)

The effect of aging in the phosphor requires different treatment. This is really another structural estimation problem, and hence requires a structure—a theory to give the functional form. If the effect were expected to be a linear decrease in the entire activity level with time, then the rate of decrease could be found by a series of experiments over several weeks at one temperature level. On the other hand, if the effect of time were expected to change the sensitivity of the phosphor to temperature, then a more elaborate investigation would be required, involving several levels of temperature at each of several times, the exact design of the experiment and analysis of the data depending rather strongly on the functional forms contained in the underlying physical theory.

In both the structure and the prediction problem, we are interested not only in fitting a line, but also in estimating the variation in the data which is not explained by the line, and especially in breaking that variation up into as many assignable components as seem justified by the theory and data available. The more complete the breakdown, the more thoroughly we shall understand what is really going on, and the better we will be able to cope with the prediction problems into which these components of variance must inevitably enter.

MATHEMATICAL MODELS

To break down the variability of our data into its smallest possible classifications, we must have a reasonable mathematical model and our algebraic techniques will depend directly on what particular model we adopt. I propose to discuss some of the models which seem to occur in the physical sciences with some frequency in order to emphasize their variety, if not to offer a pat formula for treating each one in detail. The treatment depends, as we have already said, not only on the model but also on the particular data, so detailed discussion of some points is possible only when we are confronted with data from a real experiment, and faced with the necessity of answering pertinent questions. This takes more space than is available, and I shall limit my discussion to the simpler models, showing, where possible, the sorts of questions which might naturally be asked.

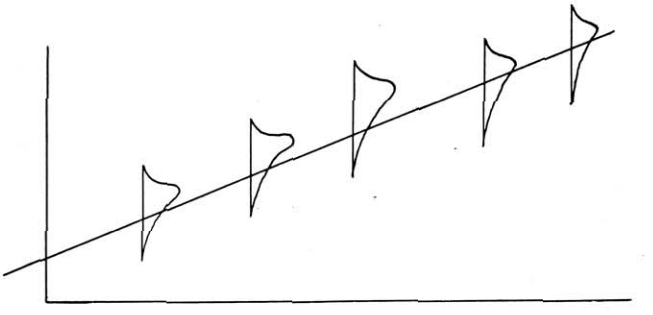

Figure 1

x without error. The most commonly discussed model involving a straight line is the one in which one variable is measured "without error," while the other is assumed to exhibit all the variability. This is the classical case treated by Gauss, known to physicists, and used by everybody else. If we sample repeatedly at one fixed value of x, we obtain a distribution of y values which is shown pictorially in Figure 1. The distributions are sketched in perspective, and—if we are lucky—look somewhat like the famous normal one. If we are still luckier, all of these distributions are identical, merely being located at different places near the line. Fitting a line to this underlying model is the ancient and honorable application of Least Squares, but the extraction of information about the residual variability goes far beyond such a mechanistic attitude. Later we shall have more to say about even this well-known model.

x and y both in error. When we consider that x and y are both measured erroneously, we face much more complicated situations. Mathematically simplest among these is the one espoused of late by Berkson where x is "set" at predetermined values, which are recorded, and y is then measured. Actually the errors in x do not enter into the numbers, X, recorded; they are uniformly and unimaginatively the same—being

the values to which the experimenter fondly hopes he has set his instrument. Because of this, the situation is actually one-variate-in-error, and, although we may admit that x is not precisely known, there is nothing the statistician can do about it, since the information is not present in the data. The graph displays the measured variable y and the idealized variable X, and thus the line shows the relation between these two. The real x is never measured. (Two centuries ago Bishop Berkley proclaimed a philosophy which denied the existence of an oak tree when no one bothered to observe it; two years ago Physician Berkson did the same for errors.)

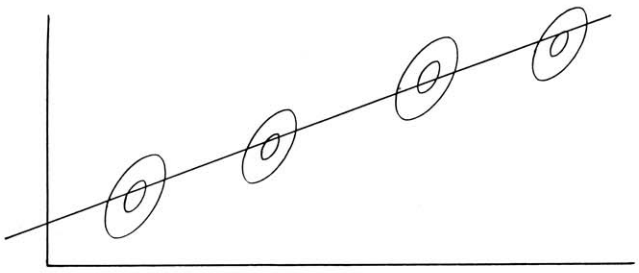

Figure 2

The true two-variates-in-error model hypothesizes a line along which the experimental system may be varied. If we set the system at a particular unknown point on this unknown line and repeatedly measure x and y, we generate a small bivariate distribution (which we hope is sort of normal) and if we move to a new point we generate a new distribution there. If we are complete optimists, we assume that the shapes of these several distributions are identical, differing only in the locations of their means. Like the one dimensional case, they may be located *on* the line, or merely near it. A fairly optimistic picture is shown in Figure 2, where we note that the error distribution contours show the x and y errors to be positively correlated—the ellipses are not parallel to the x and y axes. (They are intended to be geometrically similar, however.)

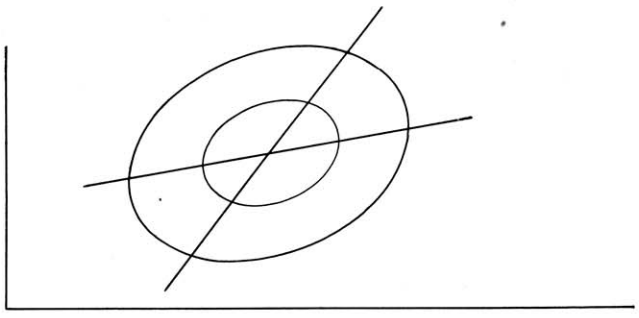

Figure 3

Still another two-variables-in-error model is the single bivariate distribution from which random or stratified samples are drawn—the heights of fathers versus the heights of sons, for example. This is not a particularly common model in the nonbiological sciences and so we introduce it here merely to say that we do not, in this paper, refer more than is absolutely necessary to this situation—one often confused with the structural fitting of interest to chemists and physicists. It is the model which gives rise to the two different lines, y on x and x on y, neither of which is the major axis of the ellipse. Figure 3 shows the intruder, and with this we shall try to dismiss him. Our attempts to get rid of our unwanted guest are foredoomed, however, as a glance at Figure 2 shows his children are already well established in our models there. We therefore make the best of a bad situation by discussing him only when we cannot do otherwise.

TYPES OF VARIABILITY

Whether our interest is structure or prediction, we need to know how the variability enters into our points before we can decide how stable our line may be or how close to it a new point is apt to fall. With the simpler models shown above there are two general sources of variability which can usefully be distinguished: (1) the amount by which points jump around within one of the elementary distributions, and (2) the amount by which the distribution means jump around the best fitted line.

The first of these types of variability is the variance of the residual "error"—it is measured by the Mean Squares Within (the finest subgrouping), to use the terminology of Analysis of Variance. The second type, in physical experiments, is apt to represent the effect of some uncontrolled variable which is nearly constant for each distribution separately, but varies from one to the next; or it may merely reflect the nonlinearity in the system, not accounted for by the theory. If we have more complicated models, with more categories of points (points in distribution I measured on Tuesday, by operator A as opposed to operator B on Wednesday), then we will get more categories of variability, some of which may cause subdivision of the two categories already mentioned, and some of which may occur in new ways—by rotating the error distributions, to mention a horrible example.

A third category of variability, which turns up in all the formalism of Analysis of Variance, is that attributable to the line itself. Since we are usually quite aware of this line, to regard it as a source of variability appears artificial, but it is done for a consistent mathematical treatment and need not trouble those who object to it. Just ignore that particular interpretation.

In order to estimate the first and finest category of variance, the Sums of Squares Within the error distributions, it is clearly necessary to have at least two points from each distribution. We then get independent estimates of the variance, σ_u^2, σ_v^2, and the covariance $\rho\sigma_u\sigma_v$ from each cluster, which may then be pooled into better estimates if we feel justified in assuming the shapes of all of these distributions to be alike. These remarks apply equally to the models with one or both variates in error. Although the SSW cannot generally be estimated by taking only one point from a distri-

bution, in one fairly common one-dimensional model, however, we can apparently estimate the SSW even with one point for each value of x, provided we are willing to assume that all the distributions are identical except for the location of their means and further that these means all fall on a true line (thereby denying the second category of variability mentioned above). In this special case, we fit a best line and then (conceptually) project all our points parallel to the fitted line down onto the y axis. Such a projection piles up all the separate similar distributions on top of one another, giving a single distribution exactly like each one of the originals. We now find ourselves with several points from one distribution and can estimate the variance quite easily. This sum of squares is merely a sum of squared deviations from the line, and is quite familiar to anybody who has fitted the simplest examples. An analogous technique is clearly not available with two dimensional error distributions unless we know where on the line the centers of these distributions are supposed to be—a most unlikely piece of information.

When we compare the no-error-in-x models with the two-variables-in-error ones, we see many similarities, but a few striking differences. In one dimensional error there is little question about which points belong to what distribution. Either we have several y values at a particular value of x, or we do not. The means of these distributions are easily found if there are several points, or it is merely the point itself if only one datum is given. When we examine the two-dimensional picture, we see that if the sample distributions lie close together it may not be at all clear to which one any particular point may belong, unless the experimenter has been able to label them. We do not know whether several points all come from one distribution, each point estimating the same "true point" (the mean of the distribution), or whether each point came from a separate though near-by distribution such as would occur if the true value of either x or y really changed a small amount between observations.

Still a further difficulty besets us when we try to decide how far a distribution mean is displaced from the line. In the one dimensional case the only displacements are vertical, but with two dimensional errors we have to decide just what point on the line is "nearest" the distribution mean, in the sense of being the most likely. Figure 4 shows a hypothetical error distribution and a line, indicating the point which is "nearest in probability." Note that this point depends on both the shape of the distribution and the slope of the line. Thus we have to know something about the general shape of the little error distributions before a line can be fitted and before the variation can be broken down to its component parts. (If we may assume bivariate normal distributions, the contours are ellipses and we must know both their eccentricity and their angle of tilt. This is frequently expressed as the equivalent requirement of knowing ρ_{uv} and σ_v/σ_u.)

If the shape of the individual error distributions is to be found from the SSW (*i. e.*, from the way a point tends to reproduce itself on repeating an experiment), then analogously the slope of the physical law is determined chiefly by the Sums of Squares Between, the amount by which the various clusters are separated in both x and y. The further apart our points, the better the slope can usually be determined. Sometimes however, the spread of our data is not under our control;

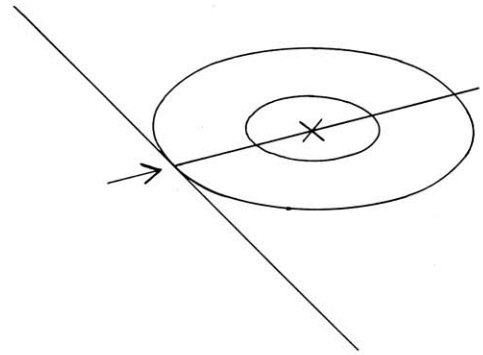

Figure 4

the variables may only be measurable in certain ranges, or perhaps they are not of human origin. In this case the problem arises that in addition to the little distributions we also have a distribution applied to x itself, and hence we can either take all the values we can get—bunched or widely distributed—or we can artificially separate them into several groups, a procedure which is intuitively distasteful.

NUMERICAL EXAMPLE

In order to illustrate some of our points, we take a set of data containing an underlying linear relation between two variables, both of which were repeatedly measured with error. To avoid losing sight of the main argument in a mass of experimental detail, we have manufactured these data to our own specifications. This has the further advantage that we know the underlying model, and can see how thoroughly we recover from the data the information we so carefully put in. The basic error distribution is assumed to be a bivariate normal with $\sigma_x^2 = 0.83$, $\sigma_y^2 = 0.56$, and $\rho\sigma_x\sigma_y = -0.186$. The underlying line is $y = 1.0 + 0.4x$, and ten measurements were made at each of the (true) points $(-4, -0.6)$, $(0, 1.0)$, $(2, 1.8)$, and $(6, 3.40)$. The values obtained are given in the table.

x	y	x	y	x	y	x	y
−5.88	−0.26	−1.62	1.97	1.40	1.65	6.11	3.30
−5.52	−0.99	−0.58	1.16	1.61	1.65	6.74	2.98
−4.59	−1.48	0	1.35	1.72	0.65	6.80	2.90
−3.68	−0.73	0.95	1.90	1.88	2.09	7.63	3.45
−3.76	0.06	1.18	−0.14	4.05	0.20	8.80	3.16
−5.22	−0.27	−0.92	2.15	0.05	3.35	7.27	2.70
−4.10	0.20	0.94	1.46	0.67	2.39	5.30	4.55
−4.87	−0.30	1.47	0.90	0.91	1.98	5.13	4.03
−4.04	−0.70	0.25	0.62	1.10	1.80	4.37	3.65
−3.18	−0.53	0.26	1.15	1.84	1.07	5.00	3.24

If we now pretend that we are a desperate statistician faced with these data and seeking a slope, knowing that a line is contained therein but otherwise unaware of the model, we might choose several courses:

(1) We might rashly decide to take x as accurately measured, ascribe all error to y, assume each error to be normally distributed fit by Least Squares, and set limits on the slope by the standard t-test.

(2) We might assume errors to be present in both x and y, use the maximum likelihood equation to determine a best slope, and wring our hands for lack of confidence limits.

(3) We might use the techniques of components of variance in regression, given by Tukey,[2] to find both a slope and some approximate limits for it.

Since we have no desire to compare the detailed mechanisms of these fitting procedures, we merely give a breakdown of the standard sums of squares and the final slopes as computed by the methods mentioned. The data are presented merely to allow the curious reader to try his own pet schemes for comparison.

Using a computational form analogous to matrix notation, we find the Sums of Squares for the data in the table to be:

Total SS	635.16168	199.06321 93.33410	$\equiv \begin{Vmatrix} Sxx & Sxy \\ & Syy \end{Vmatrix}$
SS between groups	591.96957	213.21946 76.82669	with 3 degrees of freedom
SS within groups	43.19211	−14.15625 16.50741	with 36 degrees of freedom

If we treat the x as known accurately, the slope, b, is 0.313 and five per cent limits on each side are 0.241 and 0.386. This incorrect model gives limits which don't even *include* the correct slope.

Maximum likelihood fitting gives a slope of 0.360, which is fairly good.

Regression components lead to 0.364 with 90 per cent confidence limits of 0.291 and 0.430.

If we now discard all our x observations, replacing them with the "true" values −4, 0, 2, 6, we create a Berksonian example which, however, is not comparable with the one above because it utilizes exact knowledge about x. Here the line is y-observed versus x-recorded and hence only y contains errors. The slope of this line is 0.383 with 90 per cent confidence limits of 0.323 and 0.442, given by the classical t ratio.

SOME PHYSICAL SYSTEMS

In order to give physical meaning to the mathematical models we shall mention briefly some experiments, rather idealized to be sure, which might be expected to display some of the types of variation we have been discussing. Simplest among these would be the determination of the elastic constant of a spring which obeyed Hooke's law. Here known weights, x, are applied to the spring and the elongation, y, is measured. The weights are presumed to be known accurately, so we obtain the classical model of error in y only. Repeated measurements with each of several weights would be possible so the assumption that the error distributions were independent of x could be experimentally checked. (If the error turns out to be proportional to the weight, a plot of log y versus log x will restore us to the more pleasant homoscedastic model). If we should perform our Hooke's law experiment using rocks for weights, these rocks then being weighed on kitchen scales to get x, we have a two-error model and again the measurements are repeatable if we think the spring is not going to be deformed by any particular weighing, and individual distributions may be generated. In view of the frequency with which Hooke's law is a bad approximation, the pertinent questions probably include a determination of the slope of the line with confidence limits, inquiring about whether a straight line is justified or whether a parabola is needed, and perhaps even a question about whether the later measurements show a permanent deformation or not. (This last question might be detected by fitting two lines—the second one to a second complete set of points—and then inquiring whether or not these lines could have arisen from the same population.)

A more interesting experiment is the measurement of the thickness of a zinc coating on sheet steel. This can be found by a nondestructive repeatable test with a magnigage or destructively by dissolving off the coating, measuring the thickness before and after. If we cut up a sheet of metal into several specimens and perform these tests on each one, we are sampling from some kind of a population of coating thicknesses and the picture is further obscured by the errors of measurement. Perhaps Figure 2 is a reasonable model here, but with no control over what values of x we will obtain, that variation being due to the fluctuations in the thickness of the coating from specimen to specimen. The further complication that there can be several magnigage (y) readings for one stripping (x) measurement means that no unique (x, y) point can be found, rather the mean of the y values has to be paired with the one x value, and the individual y readings used to estimate part of the variation within the finest subdivision, that imperfectly known bivariate error distribution. This is an admittedly messy model.

A Berksonian model is found in the calibration of a refractometer by weighing out several preassigned amounts, x_i, of sugar in a balance, dissolving these in equal volumes of water and measuring the refractive index, y_i. Even if we weigh out several "one gram" samples with error, we have the classical model with error in y only, as we clearly cannot estimate errors in x if we consistently record x *equals one* in our book. The nice part about all this is that for determining the slope of the line connecting y and x, this omission is unimportant, and the line obtained is unbiased. Note that in using this calibration for predictive purposes y would normally be measured, and x estimated from the curve. The pertinent questions include confidence

[2] TUKEY, J. W., "Components in Regression," *Biometrics*, **7**, 33–69 (1951).

limits for the line as well as an investigation of the errors attributable to the volumes of water used and perhaps even the operator who carried out the calibrations.

Another Berksonian experiment might test a coefficient of expansion by heating a standard length metal bar until several standard elongations, y_i, were observed (with unrecorded error) and the temperatures, x_i, measured. Here the calibration is apt to be used in either direction, and we will have to worry about how far in error we might reasonably be if we used this curve either way.

If we measure the disturbances in radio transmission caused by sunspot activity, we have an example of an experimental law (linear, we hope) in which we cannot control the variables. Here not only are the variables random (the number of sunspots and the coefficient of transmissivity) but they are both measured with errors which are themselves different random variables. Clearly the analysis of this type of data might require different techniques from the models already mentioned.

Finally, let us return to several sheets of zinc-coated steel with different nominal thicknesses, u_i. Again magnigage and stripping would yield estimates of the thickness, and thus we would obtain several clusters of values (I can scarcely say points). If we plotted the mean magnigage reading versus the stripping value, we then get a point for each specimen, and a cluster of points for each nominal thickness. This is a quite complicated model, but certainly no worse than occurs in many other experimental situations.

Some models are easy to analyze; some are very difficult. It behooves the experimenter to ponder carefully the models he believes to correspond to his reality, then to design his experiment so as to take advantage of the easier analyses. Rarely can the statistician, from the data alone, decide what model is appropriate. To reveal the model in this way requires impracticable amounts of data. But though small numbers of points cannot prescribe their own model, nevertheless a properly chosen model will extract more meaningful information from those few data than would an improper one. Thus it is necessary for the man who knows the physical system, the experimenter, to think carefully about the model from the theoretical viewpoint, as well as in the light of his experience, *before he undertakes any measurements.* If the experimenter will force himself to decide on the appropriate model before taking the data, he will avoid some of those embarrassing situations where, at the conclusion of the experiment, the information sought is not present in the numbers recorded.

In conclusion, I should like to reemphasize a point which will need to be repeated many times: the line of best fit is, by itself, of very little use in prediction if we do not know *how good* that best really is. Without such knowledge almost any line will do. The eyefitted line used regularly by the engineer is not bad and need not be apologized for—*as a line.* It is only because this technique of fitting tends to throw away much useful information that we decry it. I reiterate: it is not the Line of Best Fit we want; on the contrary, the Lines of Worst Plausible Fit are needed for prediction, those lines which bound the region in which the unknown correct line might reasonably be found. With these we obtain a picture showing our degree of confidence, here we have a quantitative measure showing our degree of uncertainty, our safety factor if you will. The wider the plausible region, the less assurance we have that our use of the fitted line will give reliable results. The narrower the region, the more confident we feel. This is one of the purposes of the statistical theory, to make quantitative our confidence and hence prevent undue optimism about our knowledge in the face of a variable universe. The other main purpose is to examine the fluctuations called errors, and by examining them to improve our knowledge about the way experimental systems operate. In the examination of those errors, the line is an obscuring gross fluctuation to be removed and removed efficiently. For this purpose statistical techniques, frequently using sums of squared residuals, have been evolved and are very important. Their purpose is not so much to fit a best line as to exorcise it, thus uncovering the finer structure of the data and revealing the subtler information which is often hidden inside, unknown even to the experimenter.

When the experimental men understand these two purposes in statistically analyzing straight line data, the cult of Least Squares will be dead. If the experimenter can be persuaded to ask, "What errors probably enter into my experimental system, and how may they be described in a simple algebraic model?" then the major battle will be won. The useful techniques of statistics will then be applied by the man who is best suited to apply them, by the experimenter. The information extracted will repay the effort well, and the thought required by the analysis will lead to improved experimental designs. Gauss will no longer be honored as a Prophet, but rather as a scientist.

ERRATUM

On page 49 of the January issue is a review of "Vinyl and Related Polymers" by Schildknecht. The title is erroneously listed as "Vinyl and Related Plastics." We apologize to the author and publisher.