Prediction, Parsimony and Noise

A model can be more accurate than the data used to build it because it amplifies hidden patterns and discards unwanted noise

Hugh G. Gauch, Jr.

A major purpose of scientific investigations is to describe reality through models. A model's worth depends in large part on its accuracy. An equivalent gain in accuracy can be achieved by increasing the size of a data set or by applying more sophisticated modeling to existing data. Collecting more data is usually costly in dollars. Modeling is costly as well, but the cost of arithmetic operations has shrunk dramatically. Several decades ago, the cost of performing a billion arithmetic steps would have been outrageous, but a present-day microcomputer can achieve billions of calculations with little cost in time or money. Computing power can be combined with advanced statistical procedures, an opportunity that often makes modeling more attractive than collecting additional data to improve accuracy. Data are costly; today, calculations are cheap.

Investigators commonly believe that a model can be no more accurate than the data it uses. But is this so? The answer depends on three matters: the precise question being asked of the model, the design of the experiment and the quantity and accuracy of the available data. For many scientific questions, numerous experimental designs and a variety of data sets, a model can be more accurate than its data.

Hugh G. Gauch, Jr., is a senior research specialist in the Department of Soil, Crop and Atmospheric Sciences at Cornell University. He received a B.S. in botany from the University of Maryland in 1964 and an M.S. in plant genetics from Cornell University in 1966. He has written books on ecological and agricultural applications of multivariate statistics. The focus of his current research is extracting more information from agricultural yield trials so that researchers can increase crop production more rapidly. Address: Department of Soil, Crop and Atmospheric Sciences, Cornell University, 1021 Bradfield Hall, Ithaca, NY 14853.

For a well-known example, consider the work of Gregor Mendel, an Augustinian monk. In 1856 Mendel began experimenting with heredity in garden peas. In one set of experiments, he crossed a pure-breeding tall pea plant—a plant with two genes for tall—with a pure-breeding short pea plant—a plant with two genes for short. The hybrid offspring had one gene for tall and one for short. This made the offspring tall because the gene for tall is dominant over the gene for short. When these hybrid plants were self-fertilized, they produced 787 tall pea plants and 277 short pea plants—a ratio of 2.84:1. Mendel completed six related experiments on other dominant-recessive combinations. When he combined his data from all seven experiments, he found a ratio of 2.98:1 for the dominant and recessive traits. He modeled this as a 3:1 ratio, which also explained and predicted other experimental results. If Mendel had generated an infinite number of pea plants in the second generation of tall-short crosses, he would have obtained one-quarter with two tall genes, one-half with one tall and one short gene and one-quarter with two short genes. This would have produced a ratio of three tall plants to every short plant, so the 3:1 model duplicates the underlying principles of genetics more accurately than Mendel's data. Likewise, the 3:1 model is a better predictor for the outcome of future experiments than is the empirical 2.84:1 result.

Increased accuracy through modeling now extends far beyond Mendel's experiments with peas. He averaged seven numbers, whereas sophisticated models now require billions of calculations. I shall describe examples in mathematics, chemistry and agriculture in which models surpass their data's accuracy. In so doing, I hope to show that the underlying statistical principles are entirely general, so that the opportunity to gain accuracy through modeling pervades science. Aggressive statistical modeling can help scientists make better use of the data already at hand, design more cost-effective experimental programs and increase the returns from research investments.

Statistical Steps

Most experiments include a treatment design and an experimental design. The treatment design specifies the controlled variables included in an experiment. For example, a treatment design might involve the yields from four varieties of a plant. On the other hand, the experimental design specifies the allocation of experimental units to the treatments, usually through randomization and often including replication. The experimental units could be the yield plots in which the varieties are tested.

An experiment samples from an entire population, and the data are used to make inferences about this population. For instance, an epidemiologist might collect data on the incidence of cancer in a thousand representative people in order to draw inferences about the world population of people who have cancer. A variety of constraints, such as cost, force investigators to use a limited sample. The distinction between a sample and a population leads to the difference between postdiction and prediction.

In literal terms, postdiction describes the past, whereas prediction forecasts the future. More specifically, postdiction says what happened within the confines of a single data set, or a sample of an entire population. A model never beats its data for postdiction, be-

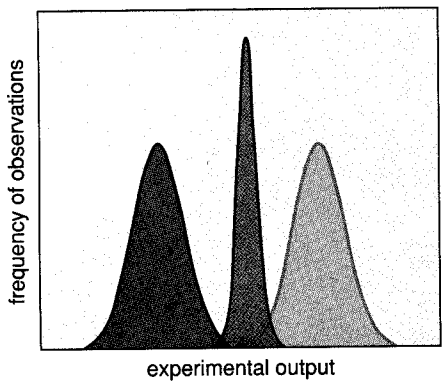

Figure 2. Variance is the spread of data around its mean. This graph reveals the frequency distribution for a hypothetical yield trial on three varieties of a plant; each curve shows the results for one variety. The variance can be divided into two categories: variance between groups and variance within groups. The variance between groups is a measure of the spread between the varieties. The variance within groups measures the spread of individual, replicated results within a variety. For instance, the variety on the right *(yellow)* and the variety on the left *(red)* have the same within-group variance, and both have more within-group variance than the variety in the middle *(blue)*.

cause the data are the actual measures of what did happen. Prediction, on the other hand, describes what is likely to happen in instances beyond the data at hand—in an entire population. Scientific applications are designed more often for prediction than for postdiction. A predictively accurate model can be closer to the true treatment effects than are its data.

A powerful model enhances prediction by amplifying the pattern in a data set and removing a large fraction of the noise. Noise arises from a host of measurement and sampling problems. Replication produces the most direct demonstration of noise. For example, a plant breeder may have four ostensibly identical replicates of a yield trial. But instead of getting four identical results, the yields might be 3,178, 2,754, 2,902 and 3,486 kilograms of crop per hectare. So, even without knowing exactly the true population mean, it is clear that the data are noisy.

The Concept of Variance

Statistical procedures are best explained through examples. Consider a hypothetical yield trial with five varieties of a crop tested in four environments, which gives the treatment design 20 treatments. The experimental design has three replications in a randomized complete-block design, which generates 60 observations.

Ronald A. Fisher, a British mathematical biologist, developed a statistical method called the analysis of variance. It is applicable to data from an experiment that has a treatment design and an experimental design. The total variation in the data is divided, or partitioned, into various sources. The first, most fundamental partition reveals two sources of variability: variance between treatments and variance within treatments. Variance between treatments arises from the treatment design; variance within treatments arises from the experimental design. Subsequent partitions reflect the organization of a given experiment. Analysis of variance determines whether the variability from a given source is large enough to be statistically significant, or else small enough to make chance the more probable explanation.

The total variation is expressed as the sum of squares of each datum minus the grand mean (the sum of all observations divided by the total number of observations), which is 167,438 in the present example. The total degrees of freedom is the number of categories minus 1, or 60 – 1 = 59 for this experiment. The total mean square is the grand mean divided by the degrees of freedom, or 167,438/59 = 2,837.93.

The first partition separates the treatment design from the experimental design. The 20 treatments have 19 degrees of freedom and a sum of squares of 147,438. The remaining degrees of freedom (59 – 19 = 40) are assigned to error, or effects of the experimental design, which has a sum of squares of 20,000. The sources of variation are indented in a table to show that the various subtotals add up to the totals. For instance, treatments and error add up to the total degrees of freedom (19 + 40 = 59) and the total sums of squares (147,438 + 20,000 = 167,438). The analysis of variance then continues by partitioning the treatments into a model and a residual, which will be explained below. Likewise, the error is partitioned into blocks and pure error. A complete block contains one replicate for all the genotypes in the experiment, and there are as many blocks as there are replications. The blocks are smaller than the entire experimental plot, and this tends to generate less variability in uncontrolled factors, such as soil, within a given block. When the block-to-block variation is large, partitioning this source from the error leaves a smaller pure error, which is the remaining variability within the blocks. Using this smaller error increases the significance of sources in F-tests.

The statistical significance of a source is judged by its F-ratio, named after Fisher. An F-ratio is calculated as the mean square of the source divided by the mean square for the appropriate error term. Then a table or a computer is used to derive the probability, or p, value. Roughly speaking, the p value is the probability that the observed result happened merely by chance; the comple-

variety

environment	a	b	c	d	e
1	352 298 331	249 295 251	266 226 219	231 254 196	182 199 216
2	268 271 232	253 270 218	243 215 205	199 205 217	234 184 221
3	201 170 169	165 214 206	154 206 153	157 176 144	171 159 195
4	125 121 102	180 133 158	159 146 124	170 180 139	121 138 164

Figure 3. Two-way factorial design generates a matrix of data. In this hypothetical yield trial, five varieties of a plant (a, b, c, d and e) were grown in four environments (1, 2, 3 and 4), and the experiment was replicated three times. The two factors, variety and environment, generate a data matrix with 20 cells—one for each variety-environment combination. Each cell holds the yields from the three replicates. Analysis of variance can be used to determine the sources of variability in the data set. (Adapted from Gauch 1992, p. 55.)

mentary value ($1 - p$) is the probability that the source caused a real effect. (James Berger and Donald Berry provide a more accurate description of the p value.) In most cases, investigators hope that the imposed treatments (different medicines, fertilizers or whatever) have a real effect. So a small p value is desired. The 0.05 and 0.01 (or 5 percent and 1 percent) significance levels are often used. In the hypothetical yield trial, the blocks have a mean square of 740 and their appropriate error term is pure error, which has a mean square of 487.37. This gives the blocks an F-ratio of 1.52, which yields a p value of 0.23206. This is not significant at even the 0.05 level, meaning that the blocks are not statistically significant in this experiment. The treatments, on the other hand, have a mean square of 7,759.89 and an F-ratio of 15.92. This gives a p value that is less than 0.00001, which is highly significant.

The amount of noise in a data set is quantified conveniently by the signal-to-noise ratio, or the signal variance divided by the noise variance. An important statistical goal is to minimize the deleterious impact of noise upon results and models. Most investigators partition the variance from the experimental design into blocks and error to increase significance levels and, in some cases, to increase the accuracy of treatment estimates. Nevertheless, investigators rarely partition the treatment variance to increase accuracy. This is unfortunate because experience shows that partitioning the treatment variance into a signal-rich model and a discarded, noise-rich residual is often several times as effective as analysis of the experimental design. Both strategies can be employed for optimal results. I consider "aggressive" statistical analysis to include partitioning of the variance in both the experimental and treatment designs. Although statistical modeling can be applied to both designs, here the term "modeling" is used primarily for analyzing the treatment design because the treatments (rather than the replications) are the entities of focal scientific interest, and their analysis generally offers greater gains in accuracy.

source	degrees of freedom	sum of squares	mean squares	F-ratio	p value
total	59	167,438	2,837.93		
treatments	19	147,438	7,759.89	15.92	0.00000
model	13	145,410	11,185.38	22.95	0.00000
genotypes	4	13,800	3,450.00	7.08	0.00023
environments	3	107,310	35,770.00	73.39	0.00000
IPCA 1	6	24,300	4050.00	8.31	0.00001
residual	6	2,028	338.00	0.69	0.65622
error	40	20,000	500.00		
blocks	2	1,480	740.00	1.52	0.23206
pure error	38	18,520	487.37		

Figure 4. Analysis-of-variance table partitions the hypothetical yield trial *(Figure 3)* into different sources of variability and judges their statistical significance. Source names are indented to highlight successive partitions. The first partition divides the variability that comes from the treatment design and the experimental design. The treatment design's degrees of freedom and sum of squares are then partitioned into an Additive Main effects and Multiplicative Interactions (AMMI) model and its residual. The model is further partitioned into a genotype effect, an environment effect and the first interaction-principal-component axis *(Figure 5)*. Likewise, the error from the experimental design is partitioned into blocks and pure error.

Comprehending Interactions

The hypothetical yield trial has a two-way factorial design with five genotypes and four environments. As mentioned above, the total degrees of freedom is 19. The simplest analysis of variance partitions the treatment variation into three sources: genotypes with four degrees of freedom, environments with three degrees of freedom and the genotype-environment interaction with 12 degrees of freedom.

The interaction is the non-additive variation that is left after removing the additive effects. For this example, the grand mean is 200, the deviation for the first genotype is 20 and the deviation for the first environment is 51. The average yield from the three replicates for the first genotype in the first environment is 327. The estimate from the additive model is 200 + 20 + 51 = 271. The interaction for this entry is 327 – 271 = 56. Note that the sum of the grand mean and the effects of the genotype, environment and genotype-environment interaction equal the experimental average. Analysis of variance finds the sums of squares for these effects to be 13,800, 107,310 and 26,328. The environmental effect is the largest, but all three are highly significant.

The additive effects involve one number for each genotype and environment, which makes them easy to understand. By sharp contrast, the interaction involves a matrix of numbers. One problem with the interaction is that most of the noise in the treatments goes into the interaction, decreasing its accuracy. Another problem is complexity. Given a real data set with 100 varieties and 30 locations, the interaction matrix has 3,000 entries. Such a matrix is not comprehended easily. It might contain complicated patterns of great importance that investigators cannot grasp by superficial examination. That challenge has generated a need for simplification. Statistical procedures for deriving parsimonious models from complex matrices can be given both geometric and algebraic explanations. I shall begin with a geometric explanation because it offers more intuitive appeal.

Around 1900, Karl Pearson of University College in London developed principal-components analysis. He visualized a matrix with r rows and c columns as r points in c-dimensional space (or the reverse). The goal of principal-components analysis is to project a

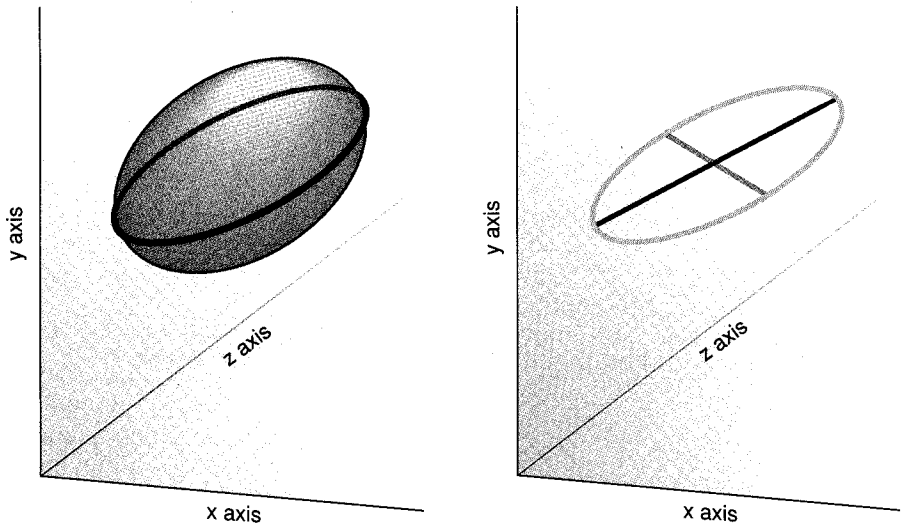

Figure 5. Principal-components analysis reduces the dimensionality of a data set. Its goal is enhanced simplicity, or parsimony, with a minimum distortion of the data. A data matrix with *r* rows and *c* columns can be conceived as *r* points in *c*-dimensional space. As a simplified example, a cloud of numerous points could occupy a three-dimensional shape similar to a football *(left)*. Principal-components analysis finds a sequence of new coordinate axes that pass through the points in the directions of greatest variation *(right)*. The first axis *(blue)* accounts for the most variation, and the second axis *(green)* accounts for the most remaining variation. In this way, principal-components analysis can reduce the original three-dimensional football to a two-dimensional ellipsoidal disk. Principal-components analysis can clarify patterns in a data set and improve the accuracy of a model.

cloud of points in high-dimensional space into relatively low-dimensional space, but maintaining the original configuration of points as faithfully as possible. More specifically, the first principal-components axis is the least-squares best-fitting line through the points. The first and second axes define the best plane through the points.

The first axis is most important. Axes can be added until their number equals the number of dimensions in the data set; the resulting full model fits the data exactly. The main goal, however, is dimensionality reduction—accepting a little inaccuracy for the sake of great gains in simplicity. In many cases, most of the variation in a large matrix, having even hundreds of dimensions, can be captured satisfactorily in just a few principal-components axes. The original high-dimensional data set may be nearly incomprehensible to investigators, but a principal-components graph, with only two or three dimensions, may reveal important patterns quite clearly. As a simplified example of dimensionality reduction, a three-dimensional cloud of points shaped like a football can be reduced to a two-dimensional elliptical disk.

Principal-components analysis can also be given an algebraic interpretation. Each row and column is given a score. A model's estimate for a given row-and-column combination is the product of their scores. This process is repeated for each axis in a model, and the results are summed. The additive model discussed earlier gives genotypes and environments deviations that are summed; the multiplicative principal-components model provides scores that are multiplied.

The Additive Main Effects and Multiplicative Interactions (AMMI) model combines analysis of variance and principal-components analysis. First, analysis of variance determines the additive, or main, effects of the factors, such as variety and environment. Then principal-components analysis is applied to the interaction (rather than to the original data). Accordingly, AMMI analysis produces some parameters that are added and others that are multiplied.

In the hypothetical yield trial, the additive effects have been given above. Recall that the grand mean is 200, the deviation for the first genotype is 20 and the deviation for the first environment is 51. The principal-components analysis of the interaction matrix gives the first genotype a score of 8 and the first environment a score of 7. Therefore, the AMMI model with one interaction-principal-components axis estimates the yield of the first genotype in the first environment as $200 + 20 + 51 + (8 \times 7) = 327$. This value happens to equal the actual yield. In general, the model is close but not equal to the data, thereby leaving some residual variance. From the sum of squares of 26,328 in the interaction, the first axis captures 24,300, leaving a residual of 2,028.

If a data set has r rows and c columns, the minimum number of axes is 0 and the maximum number is the smaller of two numbers: $(r - 1)$ and $(c - 1)$. The model with no interaction-principal-components axis is labeled AMMI0, that with one axis is AMMI1, and so on. The model that retains all axes is the full model (AMMIF). For example, if a data set has six rows and seven columns, the total number of axes is 5 ($[r - 1]$), and the AMMI family includes six members from AMMI0 to AMMI5.

The additive part of AMMI was invented by Fisher in 1918. Pearson invented the multiplicative part in 1901. In 1923 Fisher and Winifred Mackenzie of the Rothamsted Experimental Station, England, made separate applications of analysis of variance and principal-components analysis to data from a potato trial. They discovered that the multiplicative principal-components-analysis model fit the data better than the more popular additive model. Nevertheless, principal-components analysis was not accepted widely, largely because it requires as much as 100 times more calculations than analysis of variance. In 1952 analysis of variance and principal-components analysis were combined as AMMI by two groups, Evan Williams of the Commonwealth Scientific and Industrial Research Organization in Australia, and Eugene Pike and Thomas Silverberg of Raytheon Manufacturing Company. Another decade passed before computers made AMMI analysis feasible on large data sets.

The original motivation behind AMMI's development was to produce parsimonious summaries of large data matrices. AMMI provides a graph that shows the additive effects on one axis and the first interaction-principal-components scores on the other axis, using one type of point for genotypes and another for environments. This so-called biplot graph is remarkably informative, showing both additive and interaction effects for both genotypes and environments. These parameters of an AMMI model often capture more than 90 per-

cent of the entire treatment variation, giving an accurate and parsimonious presentation of data. For the hypothetical yield trial, an AMMI model captures 98.6 percent of the treatment variation.

Given a choice among the members of an AMMI family, or among some other group of competing models, investigators need to quantify a model's power. One useful measure is statistical efficiency—the number of replicates of actual data divided by the number of replicates needed for a model to produce equal accuracy. This value also equals the variance of the actual data divided by the variance of the model's estimates. For example, Richard Zobel of the USDA Agricultural Research Service and Cornell University, and I found in a soybean-yield trial that AMMI modeling based on two replicates is as accurate as the results from five replicates without modeling, which gives a statistical efficiency of 2.5. This shows that modeling offers a tremendously cost-effective means for gaining accuracy.

Choosing a Curve

To apply the concepts behind statistical modeling, I shall develop three examples: one from mathematics, one from chemistry and one from agriculture. The mathematical example models a particular cubic equation, which provides an exact basis for evaluating and comparing various models because the underlying truth is known precisely. Day-to-day research presents investigators with more difficult modeling problems, but progressively more-complicated issues will be addressed in the chemical and agricultural examples.

Scientific investigations usually begin by gathering data to search for an underlying relationship. Here I work backward: beginning with a precise underlying relationship, the cubic equation ($y = 12.00 - 3.50x + 1.17x^2 - 0.07x^3$), and then creating experimental data for integer values of x from one to seven. To mimic the collection of experimental data in the presence of noise, I add random Gaussian deviations with a variance of 0.2 times the variance of the cubic equation's data, which gives a signal-to-noise ratio of five.

The primary objective in this example is to compare the performance of various polynomials in fitting the noisy data. Such fitting involves three statistical goals. Estimation is finding the value of y at a level of x included in the experiment, here the integers from one to seven. Interpolation is finding the value of y at values of x between the experimental values, such as 1.5 or 6.8. Extrapolation is finding the value of y for a value of x outside the experimental range, such as 0.5 or 10. I shall emphasize estimation.

An infinite number of different high-degree polynomials can go through all the data points exactly. These various polynomials, however, give divergent and wild extrapolations and interpolations of y for new values of x. Although the higher-order polynomials always win the postdictive task of fitting the noisy experimental data, experience shows that parsimonious, lower-order polynomials win the predictive task of fitting new data. A parsimonious model is reduced to the simplest state that still reflects reality. Fisher recognized as early as 1921 that a parsimonious polynomial regression discards some of the noise in a data set.

As a first effort, I modeled the noisy data with a least-squares quadratic fit, which produced the equation $y = 7.95 + 1.13x + 0.06x^2$. Note that the quadratic curve is closer to the true cubic curve than are the data for six of the seven x integers. The variance of the noisy data around the values of the cubic equation is 3.524; the variance of the quadratic equation is only 1.678. This implies a statistical efficiency of 2.1 (3.524/1.678). So the model is more accurate than its data.

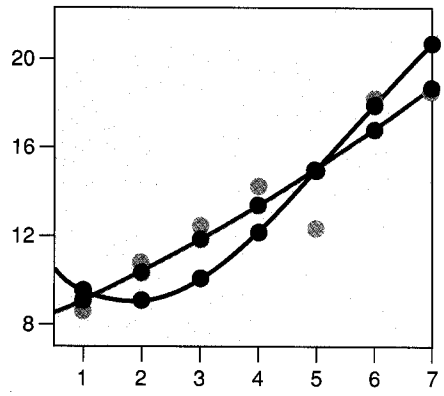

Figure 6. Cubic equation *(red)* is modeled with a quadratic equation *(blue)*. Noise is added to the cubic equation at x integers from one to seven to simulate data points *(green)*. Then the data points are fitted with a quadratic equation. At every x integer except six, the quadratic line is closer to the line from the cubic equation than the data points are. In other words, the quadratic model is a more accurate representation of the cubic equation than the data are.

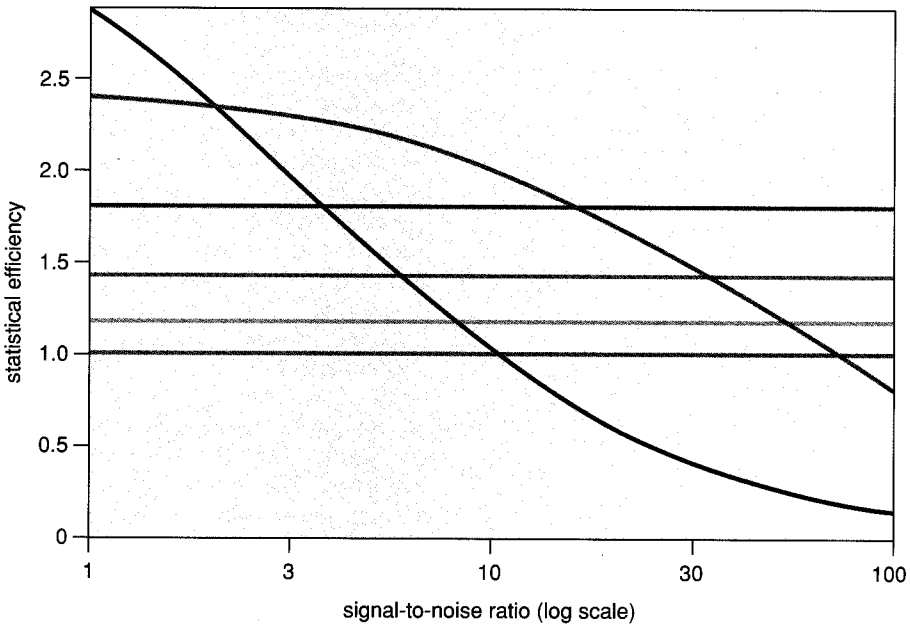

Figure 7. Noise affects a model's fit with data. The cubic equation (Figure 6) is degraded with noise. The noise is quantified with the signal-to-noise ratio—the ratio of signal variance to noise variance. Statistical efficiency is the number of real replications divided by the number of replications required for a model to produce the same accuracy. The noisy data are fit with polynomial equations of the first order *(red)*, second order *(purple)*, third order *(blue)*, fourth order *(green)*, fifth order *(yellow)* and sixth order *(orange)*. With high noise *(left)*, a first-order polynomial provides the highest statistical efficiency. A second-order polynomial gives the highest efficiency for data of intermediate noise *(middle)*. A third-order polynomial gives the best efficiency for data with little noise *(right)*. The higher-order models never give the best statistical efficiency.

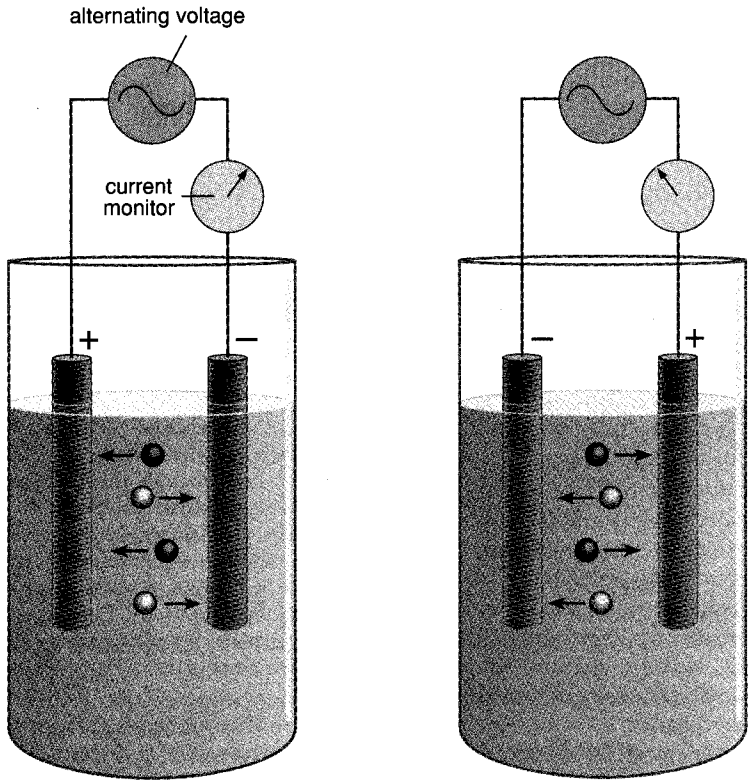

Figure 8. Electrolytic conductance can be measured by driving an alternating voltage through an ionic solution and measuring the current. Here the solution contains hydrochloric acid. The alternating voltage first makes the left-hand electrode positive *(left)*, driving the chloride ions *(red)* to the left and the hydrogen ions *(green)* to the right. When the voltage reverses *(right)*, the chloride ions move to the right and the hydrogen ions move to the left. The ionic movement carries a current.

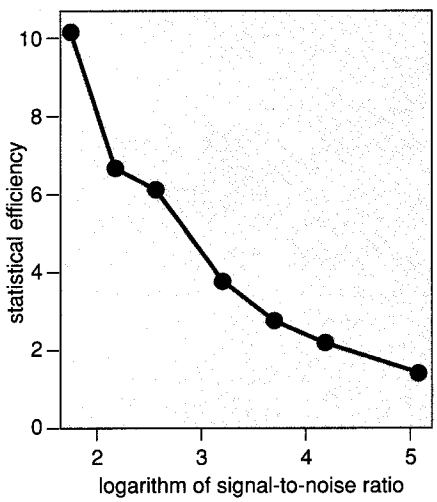

Figure 9. Electrolytic conductance was measured for hydrochloric acid of 19 concentrations at 13 temperatures. The data were degraded—made to appear more noisy—by rounding the data to less-accurate levels. The Additive Main Effects and Multiplicative Interactions (AMMI) method—a combination of analysis of variance and principal-components analysis—was used to model the data. The AMMI method produced the highest statistical efficiency for the noisiest data, those to the left side of the graph. Cleaner data *(right)* produced lower efficiencies.

The above example considers a single noisy data set. What is the average modeling performance over a variety of data sets? To answer that question, thousands of "experimental" data sets for the cubic equation were generated by using different random Gaussian deviates each time and averaging the results. The signal-to-noise ratios were varied from 1 to 100, and the model family included polynomials from the first order to the sixth order (with this last full model equivalent to the experimental data). The results show that the performance of the first-order and second-order polynomials varies with the signal-to-noise ratio, producing the highest statistical efficiency for noisier data. The higher-order polynomials provide the same statistical efficiency over all signal-to-noise ratios. Data sets with a signal-to-noise ratio below two are best fit with a first-order polynomial. A second-order polynomial best fits data sets with a signal-to-noise ratio between two and 16.6. Cleaner data, those with a signal-to-noise ratio above 16.6, are best fit with a third-order polynomial. No signal-to-noise ratio makes the fourth-, fifth- or sixth-order polynomial the best choice. In other words, lower-order models provide better predictive accuracy for noisy data sets, whereas cleaner data can support more parameters up to the true order of the cubic equation.

An Electrolytic Example

In the mathematical example, the true model was known by construction, but investigators are not afforded that luxury. Now I consider a more difficult case, the electrolytic conductance of hydrochloric acid. An electrolyte's conductance is measured by placing two electrodes in an ionic solution, connecting the electrodes to a rapidly alternating voltage and then measuring the current through the electrolyte. The conductance can be calculated from Ohm's law—voltage equals current divided by conductance. The units for equivalent conductance are centimeters squared divided by the ohm equivalent (where equivalent represents a volume of solution that contains one mole of negatively charged ions and one mole of positively charged ions).

The experiment being considered involved a two-way factorial design. One factor was temperature, which was varied from 0 to 55 degrees Celsius in increments of 5 degrees and included a final value of 65 degrees Celsius, for a total of 13 temperatures. The other factor was the concentration of hydrochloric acid, which was varied from 0.5 to 9.5 moles per liter in increments of 0.5 for a total of 19 concentrations. This data set includes 247 treatments, and the conductance ranged from 52.3 to 552.3 centimeters^2/ohm equivalent. (The raw data are available in Lide [1991, p. 5-94].)

To mimic the effect of using less-accurate measuring instruments, the data were degraded to seven levels of severity by rounding them to the nearest 1, 3, 5, 10, 20, 30 and 50. This generated signal-to-noise ratios from 130,000 to 57, with the data rounded to the nearest 1 having the highest signal-to-noise ratio and the data rounded to the nearest 50 having the lowest signal-to-noise ratio.

Then I fitted the seven noisy data sets with AMMI models. The AMMI4 model best fit the cleanest data—those rounded to the nearest 1; the AMMI2 model best fit the data rounded to the nearest 3; and the AMMI1 model best fit the data sets that were more severely degraded. Notice that the less accurate data support simpler AMMI models, just as shown in the mathematical example.

To determine the statistical efficiencies of the models, the variance of the degraded data around the original data was compared with the variance of the results from the AMMI models. For the noisiest data, those rounded to the nearest 50, the AMMI1 model achieved an impressive statistical efficiency of 10.15. For the most accurate data, those rounded to the nearest 1, the AMMI4 model's statistical efficiency was only 1.42—still worthwhile but not as impressive. So, it is possible to gain accuracy even when the underlying, true model is not known.

Figure 10. Soybean yield trial demands aggressive statistical analysis because the data are rather imprecise. Soybean varieties differ in many traits. For example, faster-maturing varieties turn yellow earlier in the fall. The most important trait is yield, and effective statistical analyses are required to determine accurately the yields from different varieties. Greater accuracy allows agricultural investigators to increase future yields more rapidly, even with less data. (Photograph courtesy of the author.)

Soybean Yields

The final example involves a soybean yield trial. The experiment employed four replications of the yields of seven soybean varieties in 10 environments. These data are rather imprecise, carrying only about one significant digit, which presents a distinctive challenge in modeling. Moreover, the true means for the treatments in this example are unknown, and they cannot be used as a standard for comparing a model's accuracy.

Consequently, these data demand a different approach for validating models. The data can be divided, using part for modeling and part for validation. For each of the 70 treatments—variety and environment combinations—three replicates are chosen at random to be used for modeling, and the remaining replicate, a total of 70 observations, is reserved for validation. This entire process is repeated many times with different randomizations, and the results averaged. Then the models AMMI0 through AMMI6 (or AMMIF) are each fitted to the modeling data to compute the expected values for each of the 70 treatments.

To generate a statistic for comparing models, I calculated the root-mean-square predictive difference. This value is a measure of the difference between the values predicted by a model and the data values reserved for validation. It is calculated by taking the difference be-

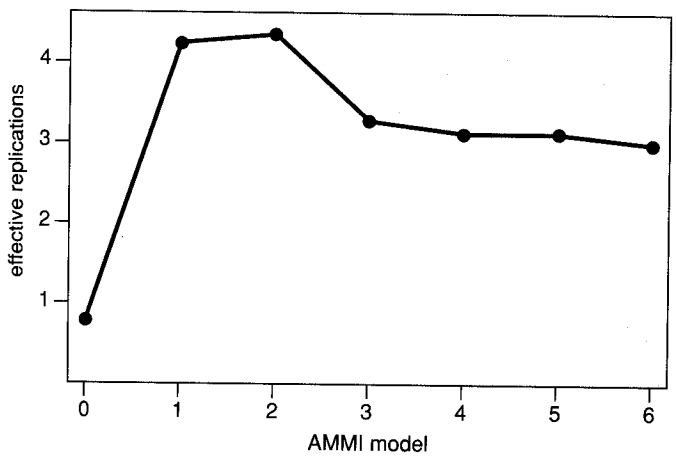

Figure 11. Less complex models better fit the soybean data. Effective replications is the number of replications needed with mere averaging to match the accuracy of a statistical model. Seven AMMI models, including zero to six principal-components axes, are applied to the yield data from the soybean trials. Three of the soybean replicates are used to generate models, and the fourth replicate is used to validate models. The AMMI model with zero axes produces less than one effective replication, so it is worse than the data. The models with one and two axes produce more than four effective replications, even though they are generated from just three actual replications. The more-complex models, with from three to six principal-components axes, provide about three effective replications, making them about equal to the data. The shape of this curve is called "Ockham's hill" after William of Ockham.

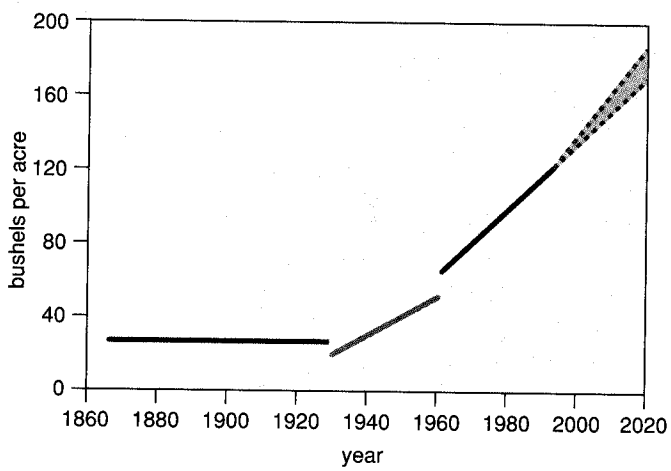

Figure 12. U.S. corn yields have increased with new technology. Before 1930 *(green)*, farmers relied on open pollination, selecting the best ears for future crops, which produced about 27 bushels per acre. Around 1930, investigators began using hybrids *(yellow)*, and the yields doubled. By the 1960s, better hybrids became available *(blue)*, raising the yields to about 120 bushels per acre. If the current trend continues *(blue dashes)* yields will be about 170 bushels per acre by the year 2020. But if investigators apply aggressive statistical analysis to selecting varieties *(red dashes)*, corn yields will likely exceed 180 bushels per acre by 2020. If aggressive data analysis generated similar yield increases in all major crops—such as corn, wheat, rice and soybeans—it would provide enough additional food to feed hundreds of millions of people. (Data from A. F. Troyer, Dekalb Plant Genetics.)

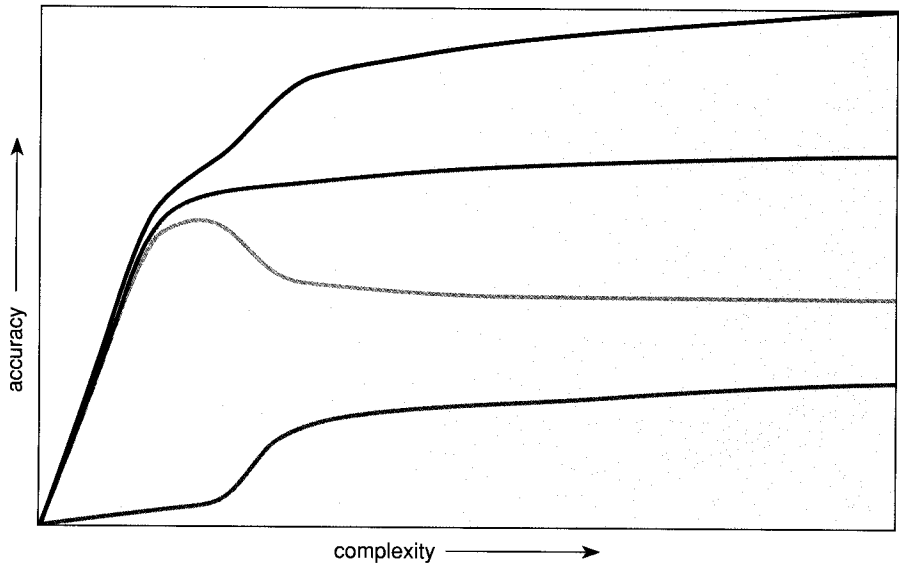

Figure 13. Accuracy and parsimony in combination determine modeling's benefits. Because pattern *(blue)* is determined by a few main factors, even simple models recover much of it in a data set. Because noise *(purple)* is idiosyncratic and complex, it is recovered more slowly as a model increases in complexity. A model's postdictive accuracy *(red)* develops from the sum of pattern and noise recovery. It increases quickly even in simple models, and then it tapers to a slower increase. By contrast, predictive accuracy *(green)* arises from pattern recovery minus noise recovery. It also increases quickly in simple models, but it then peaks on "Ockham's hill," and it later decreases in increasingly complex models. The goal of modeling is accuracy and parsimony. So the better models are located in the upper-left area in this graph.

tween a model's estimate and the validation observation for each treatment, squaring the result, summing the results across all treatments, dividing the sum by the number of validation observations and finally taking that number's square root. A low root-mean-square predictive difference is good, indicating that the model's expected values are close to the validation data. In this yield experiment, the AMMI2 model generated the most accurate results, with a root-mean-square predictive difference of 352.73 kilograms per hectare. The AMMI1 model was a close second, with a root-mean-square predictive difference of 353.69 kilograms per hectare.

Although the root-mean-square predictive difference offers the appeal of coming from empirical data, it is not the desired measure. Instead, a model should be assessed by its variance around the true means of the treatments, and not be penalized for imperfections in the validation data. Fortunately, this quantity can be estimated from the variance rule. In this case, the variance rule says: The variance of predictive differences equals the model's variance plus the variance of the validation data. The variance of the predictive difference for AMMI2 is just the square of the root-mean-square predictive difference, or $352.73^2 = 124{,}421$. The variance of the validation observations can be calculated from the data, and it is simply the error mean square, 101,161. Finally, the variance of the AMMI2 model is 124,421 – 101,161 = 23,260. Taking the square root of this number reveals the root-mean-square noise, which is 152.51 kilograms per hectare. This noise is the difference between a model's predicted value and the true mean of the population. So the estimates from the AMMI2 model are generally within about 152.51 kilograms per hectare of the true mean yields. These calculations show how a model's accuracy can be determined from a large number of validation observations.

These numbers can also be used to calculate a model's effective replications, which is the number of replications required when merely averaging the replicates in order to achieve the same accuracy that the model can with a (usually) smaller number of actual replications. This measure is calculated by dividing the variance in the validation observations by the variance of the model, or $101{,}161/23{,}260 = 4.35$. In other words, although the AMMI2 model is based on only three actual replications, it is as predictively accurate as the treatment means from 4.35 replications. Again, this model is better than its data. It provides 1.35 free replications, which in this case is equivalent to 94 free observations (about $2,000 worth of data).

The effective replications from the soybean-yield trial can be determined for the entire AMMI model family. Doing so shows that the AMMI0 model provides less than one effective replication, even though the data are based on three replications. Here the model is worse than the data because it underfits real patterns. The AMMI1 and AMMI2 models both generate more than four effective replications, and they are better than the data. All the higher models produce about three effective replications, and they are about equal to the data. Such a graph of effective replications versus increasingly complex models reveals what David MacKay of the Cavendish Laboratory, England, called "Ockham's hill," after William of Ockham, a 14th-century English philosopher who said that entities should not be multiplied without necessity. The model that provides the maximum effective replications, AMMI2 in this case, is at the peak of the hill. To the left of the peak, the curve falls sharply because the models underfit the pattern; to the right of the peak, the curve declines slowly where the models overfit the noise.

Economic Advantages

The increased accuracy from AMMI analysis provides plant breeders with two benefits: a more reliable determination of superior varieties and a quicker increase in crop yields. If AMMI modeling is used in most of the plant-breeding programs during the next several years, I estimate conservatively that the additional food production will be enough to feed several hundred million people. Moreover, the cost of the statistical analysis is trivial in comparison with the cost of collecting more data to achieve the same gain in accuracy.

The financial benefits of modeling are best presented through an example. Imagine that a yield trial is completed and that producing and collecting the data cost $500. If an overfit model is applied, it captures all the pattern in the data as well as all the noise, and it represents the average yields over the replications. Let us say that having the overfit model—the raw data—is worth $2,000. A good parsimonious model, with only a middling statistical efficiency of 2.5, would be worth $5,000—the product of $2,000 and 2.5. Underfit models would be worth very little, say $300, because they would capture little of the pattern.

What is the total value of these models? The full, overfit model costs $500 for data collection and is worth $2,000, so the experimental work returns $4 for every dollar invested. The cost of calculating the best model would be about $50, but it provides $3,000 worth of additional value. In other words, the best model returns $60 in value for every dollar invested in modeling.

This example reveals a typical relationship: Modeling provides an order of magnitude greater return on investment than does experimentation alone. The two phases of research, however, are not competitive because modeling requires data. Instead, experimentation and modeling are complementary.

This example concentrates on monetary advantages, but aggressive statistical analysis offers far greater benefits. More accurate modeling also produces better medical treatments, improved products, more food and a variety of other advances.

The Magnitude of Modeling

Most modeling efforts follow a simple philosophy: Enhance prediction by amplifying a pattern and by discarding noise. The examples from mathematics, chemistry and agriculture reveal specific applications of this approach, but the philosophy can be applied broadly. Throughout any scientific investigation, there is a subtle interplay between prediction, parsimony and noise.

This interplay appears if prediction, postdiction, pattern and noise are plotted together on a graph of model accuracy versus model complexity. Most of the pattern in a data set is recovered quickly with even relatively simple models. A pattern depends usually on just a few main causal factors, and thereby is relatively parsimonious and summarized readily. Noise, on the other hand, is recovered slowly as a model's complexity increases. Noise is idiosyncratic, complex and not easily summarized. In a set of related data, models can combine information to discriminate between pattern and noise.

The accuracy of both prediction and postdiction rise quickly as parameters are added in relatively simple models. Postdictive accuracy continues to increase with a model's complexity, but the accuracy of prediction peaks rather early and then decreases with increasingly complex models. Postdiction does not distinguish pattern from noise, so postdictive accuracy rises as the sum of pattern and noise. In contrast, prediction rewards the recovery of pattern but penalizes the recovery of noise. This causes predictive accuracy to increase as a function of pattern recovery minus noise recovery. Note that the distinction between postdiction and prediction generates two strikingly different views on parsimony, with only the latter exhibiting Ockham's hill.

Another benefit of modeling arises from the distinction between direct and indirect information. For example, using the least-squares quadratic fit to estimate the cubic equation's y value at $x = 3$ exploits the one experimental measure at that x integer, the direct information, and the six additional measures from other x integers, the indirect information. Likewise, in Mendel's experiments, he developed a 3:1 ratio for the tall-short crosses by combining the results from these crosses (direct information) with the results from six related crosses (indirect information). An individual piece of indirect information may not have much value, but combining all the indirect information can enhance the pattern and decrease the noise in a model.

Given this general picture of modeling, when does modeling offer the largest benefits? The answer involves three statistical considerations. First, modeling offers larger benefits as noise increases. Second, modeling accomplishes more for larger data sets. Third, modeling excels when a parsimonious, simple model captures most of the data's pattern. Two practical considerations must also be mentioned. Modeling is more advantageous when data-collection costs surpass the cost of statistical analysis. And modeling is extremely useful when gains in accuracy are urgent and the derived decisions are of great importance, such as in the diagnosis of an illness.

The conclusions reached here extend beyond the polynomial and the AMMI models. The general story about prediction, parsimony and noise applies to a wide class of models: multiple regression, moving averages, time-series analysis, factor analysis, canonical correlation analysis and countless others. For example, Gary Fick of Cornell and David Onstad of the Illinois Natural History Survey modeled an alfalfa yield trial with

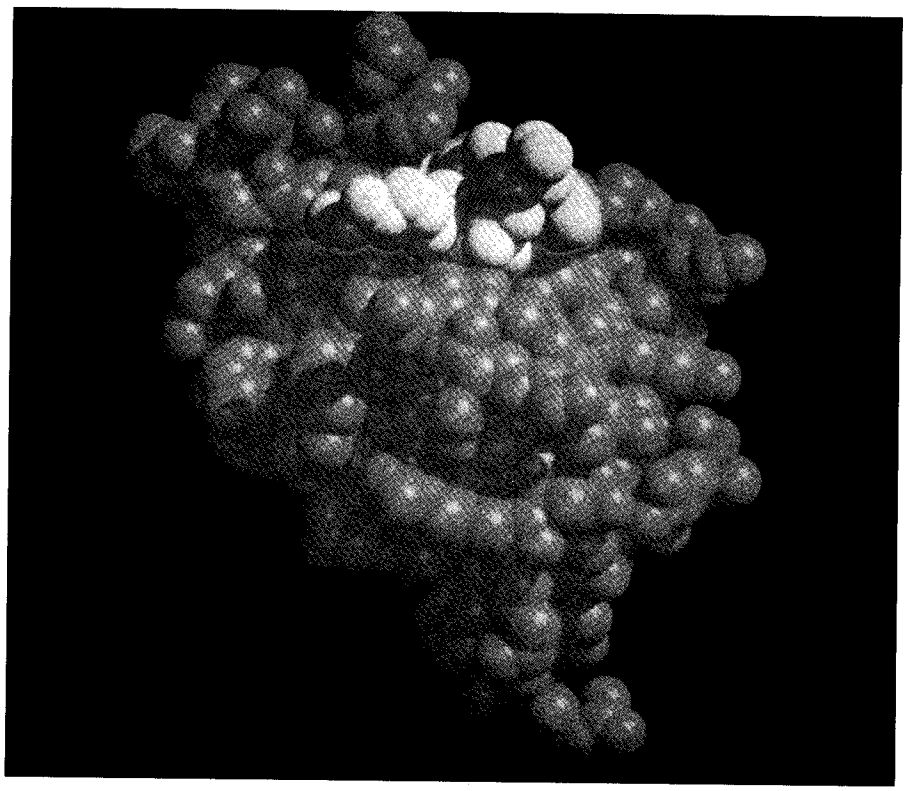

Figure 14. Protein shape is modeled with a statistical interpretation of x-ray-diffraction data. The shape of this immunosuppressive drug, FK506 *(yellow)*, bound to a protein, FKBP *(blue)*, in a T cell of the immune system was determined by using experimental x-ray-diffraction data, which alone are too noisy to determine an accurate structure. The diffraction data are combined with additional information on typical bond lengths and angles to determine the most-accurate structure of the protein. Image courtesy of Gregory Van Duyne of Yale University and Andrew Karplus and Jon Clardy of Cornell University.)

multiple regression, and they found that relatively parsimonious models were more predictively accurate.

At present, there are few exceptions to the generalization that investigators ignore the opportunity to gain accuracy through modeling. Nevertheless, principal-components analysis and other modeling systems have been used for many years in signal, radar and image processing to increase the signal-to-noise ratio and thereby sharpen images.

More recently, Axel Brünger of Yale University applied aggressive statistical modeling to the three-dimensional shape of proteins. A family of models was fitted to noisy x-ray diffraction data, reserving part of the data for validation to identify the most predictively accurate model. Choosing the model that optimally fits the data—neither overfitting nor underfitting it—allows a description of the protein shape that best reflects its true structure. Gregory Van Duyne of Yale University, Andrew Karplus and Jon Clardy of Cornell and Robert Standaert and Stuart Schreiber of Harvard University used Brünger's technique to describe the shape of a complex formed by an immunosuppressive drug, FK506, and a protein, FKBP, in T cells—the guardians in the immune system that detect and destroy foreign cells. This immunosuppressive drug is being prescribed for transplant patients to limit the chances of tissue rejection. Knowing the shape of the drug-protein complex may lead to the development of better immunosuppressants.

Scientific investigators have many needs and opportunities for models that are better than their data. Simplistic analyses often glean only a fraction of the information in a hard-won data set, whereas modeling extracts even the most subtle patterns. Failing to analyze data effectively is like a leaving an orange half-squeezed. As statisticians often say: Data worth collecting are also worth analyzing.

Acknowledgments

I appreciate suggestions on this manuscript from Millard Baublitz, John Chiment, Richard Furnas, Ruben Gabriel, Andrew Karplus, David MacKay, Nicholas Rescher, Frederick Suppe, Forrest Troyer and Richard Zobel. This work was supported by the Rhizobotany Project of the USDA Agricultural Research Service.

Bibliography

Berger, James O., and Donald A. Berry. 1988. Statistical analysis and the illusion of objectivity. *American Scientist* 76:159–165.

Bradu, Dan, and K. Ruben Gabriel. 1978. The biplot as a diagnostic tool for models of two-way tables. *Technometrics* 20:47–68.

Brünger, Axel T. 1993. Assessment of phase accuracy by cross validation: the free R value. *Acta Crystallographica, Section D* 49:24–36.

Fick, Gary W., and David W. Onstad. 1988. Statistical models for predicting alfalfa herbage quality from morphological or weather data. *Journal of Production Agriculture* 1:160–166.

Fisher, Ronald A. 1918. The correlation between relatives on the supposition of Mendelian inheritance. *Transactions of the Royal Society of Edinburgh* 52:399–433.

Fisher, Ronald A. 1921. Studies in crop variation. I. An examination of the yield of dressed grain from Broadbalk. *Journal of Agricultural Science, Cambridge* 11:107–135.

Fisher, Ronald A., and Winifred A. Mackenzie. 1923. Studies in crop variation. II. The manurial response of different potato varieties. *Journal of Agricultural Science, Cambridge* 13:311–320.

Gauch, Hugh G. 1992. *Statistical Analysis of Regional Yield Trials.* New York: Elsevier.

Gauch, Hugh G., and Richard E. Furnas. 1991. Statistical analysis of yield trials with MATMODEL. *Agronomy Journal* 83:916–920.

Gauch, Hugh G., and Richard W. Zobel. 1988. Predictive and postdictive success of statistical analyses of yield trials. *Theoretical and Applied Genetics* 76:1–10.

Jefferys, William H., and James O. Berger. 1992. Ockham's razor and Bayesian analysis. *American Scientist* 80:64–72.

Jeffreys, Harold. 1983. *Theory of Probability.* Third Edition. Oxford: Clarendon Press.

Lide, David R., ed. 1991. *CRC Handbook of Chemistry and Physics.* Boca Raton, Florida: CRC Press.

MacKay, David J. C. 1992. Bayesian interpolation. *Neural Computation* 4:415–447.

Mendel, Gregor. 1865. Versuche über Pflanzenhybriden. *Verhandlungen des Naturforschenden Vereins in Brünn* 4:3–47.

Pearson, Karl. 1901. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine Sixth Series* 2:559–572.

Pike, Eugene W., and Thomas R. Silverberg. 1952. Designing mechanical computers. *Machine Design* 24:131–137, 159–163.

Williams, Evan J. 1952. The interpretation of interactions in factorial experiments. *Biometrika* 39:65–81.

Next time will you be the guinea pig?

mentary value ($1 - p$) is the probability that the source caused a real effect. (James Berger and Donald Berry provide a more accurate description of the p value.) In most cases, investigators hope that the imposed treatments (different medicines, fertilizers or whatever) have a real effect. So a small p value is desired. The 0.05 and 0.01 (or 5 percent and 1 percent) significance levels are often used. In the hypothetical yield trial, the blocks have a mean square of 740 and their appropriate error term is pure error, which has a mean square of 487.37. This gives the blocks an F-ratio of 1.52, which yields a p value of 0.23206. This is not significant at even the 0.05 level, meaning that the blocks are not statistically significant in this experiment. The treatments, on the other hand, have a mean square of 7,759.89 and an F-ratio of 15.92. This gives a p value that is less than 0.00001, which is highly significant.

The amount of noise in a data set is quantified conveniently by the signal-to-noise ratio, or the signal variance divided by the noise variance. An important statistical goal is to minimize the deleterious impact of noise upon results and models. Most investigators partition the variance from the experimental design into blocks and error to increase significance levels and, in some cases, to increase the accuracy of treatment estimates. Nevertheless, investigators rarely partition the treatment variance to increase accuracy. This is unfortunate because experience shows that partitioning the treatment variance into a signal-rich model and a discarded, noise-rich residual is often several times as effective as analysis of the experimental design. Both strategies can be employed for optimal results. I consider "aggressive" statistical analysis to include partitioning of the variance in both the experimental and treatment designs. Although statistical modeling can be applied to both designs, here the term "modeling" is used primarily for analyzing the treatment design because the treatments (rather than the replications) are the entities of focal scientific interest, and their analysis generally offers greater gains in accuracy.

source	degrees of freedom	sum of squares	mean squares	F-ratio	p value
total	59	167,438	2,837.93		
treatments	19	147,438	7,759.89	15.92	0.00000
model	13	145,410	11,185.38	22.95	0.00000
genotypes	4	13,800	3,450.00	7.08	0.00023
environments	3	107,310	35,770.00	73.39	0.00000
IPCA 1	6	24,300	4050.00	8.31	0.00001
residual	6	2,028	338.00	0.69	0.65622
error	40	20,000	500.00		
blocks	2	1,480	740.00	1.52	0.23206
pure error	38	18,520	487.37		

Figure 4. Analysis-of-variance table partitions the hypothetical yield trial *(Figure 3)* into different sources of variability and judges their statistical significance. Source names are indented to highlight successive partitions. The first partition divides the variability that comes from the treatment design and the experimental design. The treatment design's degrees of freedom and sum of squares are then partitioned into an Additive Main effects and Multiplicative Interactions (AMMI) model and its residual. The model is further partitioned into a genotype effect, an environment effect and the first interaction-principal-component axis *(Figure 5)*. Likewise, the error from the experimental design is partitioned into blocks and pure error.

Comprehending Interactions

The hypothetical yield trial has a two-way factorial design with five genotypes and four environments. As mentioned above, the total degrees of freedom is 19. The simplest analysis of variance partitions the treatment variation into three sources: genotypes with four degrees of freedom, environments with three degrees of freedom and the genotype-environment interaction with 12 degrees of freedom.

The interaction is the non-additive variation that is left after removing the additive effects. For this example, the grand mean is 200, the deviation for the first genotype is 20 and the deviation for the first environment is 51. The average yield from the three replicates for the first genotype in the first environment is 327. The estimate from the additive model is 200 + 20 + 51 = 271. The interaction for this entry is 327 – 271 = 56. Note that the sum of the grand mean and the effects of the genotype, environment and genotype-environment interaction equal the experimental average. Analysis of variance finds the sums of squares for these effects to be 13,800, 107,310 and 26,328. The environmental effect is the largest, but all three are highly significant.

The additive effects involve one number for each genotype and environment, which makes them easy to understand. By sharp contrast, the interaction involves a matrix of numbers. One problem with the interaction is that most of the noise in the treatments goes into the interaction, decreasing its accuracy. Another problem is complexity. Given a real data set with 100 varieties and 30 locations, the interaction matrix has 3,000 entries. Such a matrix is not comprehended easily. It might contain complicated patterns of great importance that investigators cannot grasp by superficial examination. That challenge has generated a need for simplification. Statistical procedures for deriving parsimonious models from complex matrices can be given both geometric and algebraic explanations. I shall begin with a geometric explanation because it offers more intuitive appeal.

Around 1900, Karl Pearson of University College in London developed principal-components analysis. He visualized a matrix with r rows and c columns as r points in c-dimensional space (or the reverse). The goal of principal-components analysis is to project a