VOL. 81, No. 2 FEBRUARY 1974

Psychological Bulletin

LINEAR MODELS IN DECISION MAKING [1]

ROBYN M. DAWES [2] AND BERNARD CORRIGAN

University of Oregon and Oregon Research Institute, Eugene

Linear models are frequently used in situations in which decisions are made on the basis of multiple codable inputs. These models sometimes are used normatively (to aid the decision maker), sometimes contrasted with the decision maker (in the "clinical versus statistical" controversy), sometimes used to represent the decision maker ("paramorphically"), and sometimes used to "bootstrap" the decision maker (by replacing him with his representation). Linear models have been successfully employed in a variety of contexts. A review of these contexts indicates that they have common structural characteristics: (*a*) Each input variable has a conditionally monotone relationship with the output; (*b*) there is error of measurement; and (*c*) deviations from optimal weighting do not make much practical difference. These characteristics ensure the success of linear models. In fact linear models are so appropriate in such contexts that *random linear models* (i.e., models whose weights are randomly chosen except for sign) may perform quite well. Four examples involving the prediction of such codable output variables as grade point average and psychiatric diagnosis are analyzed in detail. In all four, random linear models yield predictions that are superior to those of human judges.

To the best of our knowledge, the first use of linear models in decision making was proposed by Benjamin Franklin (in Bigelow, 1887) in a letter to his friend, Joseph Priestly, on September 19, 1772:

> I cannot, for want of sufficient premises, advise you *what* to determine, but if you please I will tell you *how*. . . . My way is to divide half a sheet of paper by a line into two columns; writing over the one *Pro,* and over the other *Con.* Then, doing three or four days' consideration, I put down under the different heads short hints of the different motives, that at different times occur to me *for* or *against* the measure. When I have thus got them all together in one view, I endeavor to estimate the respective weights . . . [to] find at length where the balance lies . . . And, though the weight of reasons cannot be taken with the precision of algebraic quantities, yet, when each is thus considered, separately and comparatively, and the whole matter lies before me, I think I can judge better, and am less liable to make a rash step; and in fact I have found great advantage for this kind of equation, in what may be called *moral* or *prudential algebra* [p. 522].

By estimating the respective weights of pro and con arguments and finding "where the balance lies," Franklin was in effect adding together the positive weights of the pro arguments with the negative weights of the con arguments and then deciding pro or con,

[1] This research was supported, in part, by National Institute of Mental Health Grants, MH-21216 and MH-12972 and by National Science Foundation Grant GS-32505. We would like to thank our Oregon Research Institute colleagues for their detailed comments on earlier versions and Nancy Wiggins for her generosity in sharing her data and for her joint effort in replicating the Wiggins and Kohen experiment at Oregon. We would also like to thank Kenneth R. MacCrimmon for acquainting us with the quotation from Benjamin Franklin. Much of the material in this article has been verbally presented previously at the 1971 and 1972 annual meetings of the Society of Multivariate Experimental Psychologists, at invited colloquia and talks at the University of South Carolina business school (October 1972), at the University of Chicago business school (November 1972), at the University of Illinois Psychology Department (May 1973), at the 1973 convention of the Western Psychological Association (April 1973, tutorial), and at the 1973 convention of the Operations Research Society of America (November 1973).

[2] Requests for reprints should be sent to Robyn M. Dawes, Oregon Research Institute, P.O. Box 3196, Eugene, Oregon 97403.

depending on whether the sum was positive or negative. Franklin's use of the linear model was normative—that is, the model is meant to aid the decision maker in reaching a good decision. Linear models can also be descriptive—that is, the model is meant to represent the decision maker's behavior. They are sometimes used to decide whether or not to do something (as above); they are sometimes used to rank or rate objects or alternatives. Types of linear models range from those in which optimal weights are obtained by least squares regression procedures to those in which intuitive weights are obtained (as above) to those in which unit weights are applied (i.e., in which variables are made comparable in some manner and then simply added together). This article reviews the use of linear models in various decision-making contexts and proposes reasons that they are so ubiquitous. The review leads to the conclusion that a wide variety of decision-making contexts have structural characteristics that make linear models appropriate. It then naturally follows that they be used to help make good decisions and, insofar as a decision maker is behaving appropriately, they may be used to describe his decisions. Indeed linear models are so appropriate in some contexts that those with randomly chosen weights outperform expert judges.

In this article, four examples of linear models of decision making are discussed in detail. All involve a comparison of the accuracy of five modes of decision making: intuitive judgment, linear models based on intuitive judgment, linear models with optimal weights, linear models with randomly chosen weights, and linear models with unit weights.

Example 1

A pool of approximately 1,200 psychiatric patients took the Minnesota Multiphasic Personality Inventory (MMPI) in various hospitals; they were later categorized as "neurotic" or "psychotic" on the basis of more extensive information. The MMPI results are in the form of a personality profile of 11 scores, each of which represents the degree to which the respondent answers questions in a manner similar to patients suffering from a well-defined form of psychopathology. Thus a set of 11 scores is associated with each patient, and the problem is to predict whether a later diagnosis will be psychotic (coded 1) or neurotic (coded zero).

Example 2

Ninety first-year graduate students in the psychology department at the University of Illinois were evaluated on 10 variables that are predictive of academic success. These variables included aptitude test scores, college grade point average, various peer ratings (e.g., extroversion), and various self-ratings (e.g., conscientiousness). A first-year graduate grade point average was computed for all these students. The problem was to predict this grade point average from the 10 variables.

Example 3

Graduate students in the psychology department at the University of Oregon, who had been there from two to five years (or who would have been had they not dropped out), were evaluated on a 5-point rating scale by faculty members who knew them. The problem was to predict the average faculty rating from three variables available to the admissions committee at the time these students applied: scores on the Graduate Record Exam (GRE), undergraduate grade point average, and an approximate rating of the quality of the institution at which the grade point average was obtained.

Example 4

Experimenters assigned values to ellipses presented to subjects on the basis of each figure's size, eccentricity, and grayness; the formula used was $ij + jk + ik$, where i, j, and k refer to values on the three dimensions just mentioned. Subjects in this experiment were asked to estimate the value of each ellipse and were presented with outcome feedback at the end of each trial. The problem was to predict the true (i.e., experimenter assigned) value of each ellipse on the basis of its size, eccentricity, and grayness.

Clinical versus Statistical Prediction

One of the first areas to be investigated by clinical psychologists, as the profession grew rapidly after World War II, was the degree

to which human judgment could be used in the prediction of variables such as patient response to treatment, recidivism, or academic success (Sarbin, 1943). What could such judgment add to prediction that could be made on a purely statistical basis by, for example, developing linear regression equations? The statistical analysis was thought to provide a floor to which the judgment of the experienced clinician could be compared.

The floor turned out to be a ceiling. Meehl (1954) reviewed approximately 20 studies in which actuarial methods were pitted against the judgments of the clinician; in all cases the actuarial method won the contest or the two methods tied. Since the publication of Meehl's book, there has been a plethora of additional studies directed toward the question of whether clinical judgment is inferior to actuarial prediction (Sawyer, 1966), and some of these studies have been quite extensive (Goldberg, 1965). But Meehl (1965) was able to conclude, some 10 years after his book was published, that there was only a single example in the literature showing clinical judgment to be superior, and this conclusion was immediately disputed by Goldberg (1968a) on the grounds that even that example did not show such superiority. We know of no examples after that (within the standard limitations) that have purported to show the superiority of clinical judgment.

The first of these limitations is that comparative validity has always been evaluated by comparing the correlation between the criterion and the judges' predictions with the cross-validated correlation between the criterion and the predictions of the actuarial model, usually a regression equation. But no one has proposed an alternative way of comparing predictability, and correlation, because it is a good index of the degree to which two variables are rank ordered in a similar fashion, is a reasonable measure for assessing the prediction of such variables as success in graduate school or response to therapy.

The second limitation is that both the clinical predictions and those of the actuarial model are made on the basis of the *same codable input*. (Naturally, one cannot perform a linear regression analysis or a Bayesian analysis on uncoded variables.) Clinical judges may be superior in contexts in which they have access to variables that are not clearly codable or to variables that are codable, but cannot be assessed without the clinician's presence—for example, his feeling of liking or disliking a patient or potential graduate student. In fact there has been one recent investigation in which football experts predicted point spread better than did a linear prediction equation (Pankoff, 1967), but these judges may well have had access to information other than that fed into the equation. (This second limitation was laid down as one of the "ground rules" for the clinical versus statistical controversy by Meehl in 1954.)

A few authors, rather than investigating clinical versus statistical prediction, have attempted to synthesize actuarial and clinical prediction (Pankoff & Roberts, 1968; Sawyer, 1966). Such syntheses themselves may be classified as either clinical or statistical. In a clinical synthesis, an expert decision maker is given the outcome of the statistical prediction and then asked to improve upon it, whereas in a statistical synthesis the judgment of the expert is treated as an additional variable in the actuarial prediction system. Although such procedures are very appealing on purely logical grounds, empirical evidence concerning their success is not very encouraging. Goldberg (1968b) reported a study of clinical synthesis in which judges were given "actual values of the optimal formula for each [MMPI] profile," with the result that "the accuracy of these judges' diagnoses was not as high as would have been achieved by simply using the formula itself [p. 493]." Einhorn (1972) reported a study of statistical synthesis in which four medical experts on hodgkin's disease rated nine characteristics of biopsies from some 200 patients and also made an overall rating of the severity of the disease process. All the patients later died, and Einhorn was able to relate their longevity to both the nine characteristics and the overall judgments. He built and cross-validated two linear models for each doctor—one including the overall rating and one excluding it. For two of his four doctors, the model that included the overall rating had a

higher cross-validated correlation than did the model that excluded it; for two, the cross-validated correlation was lower.

In the examples discussed in this article, linear models with optimal coefficients have a higher cross-validated correlation than do human judgments.

Example 1

Twenty-nine clinical psychologists were asked to predict, on the basis of MMPI profiles, whether the patients were diagnosed as neurotic or psychotic; they made their predictions using a forced–normal distribution. The correlation between their ratings and the criterion ranged from .14 to .39, with a mean of .28; the cross-validated correlation of the weighting scheme derived from regression analysis was .46 (Goldberg, 1965). Moreover, the partial correlation between judgments and criterion, partialling out the predictions of the optimal linear model, averaged only .05; hence, the regression weights for such judgments in a linear synthesis (weighting) of clinical and actuarial predictions would be virtually zero (Hays, 1963, p. 575).

Example 2

Eighty University of Illinois students were asked to predict the grade point averages of the 90 first-year students who were evaluated on the 10 variables listed earlier; the correlations between predicted and obtained grade point average ranged from .07 to .48, with an average of .33; the cross-validated correlation resulting from regression analysis was .57 (Wiggins & Kohen, 1971). The predictions of 41 graduate students at the University of Oregon had correlations ranging from .14 to .48, with an average of .37, which again is less than that obtained from the regression analysis. As in Example 1, the average partial correlation between clinical judgment and criterion partialling out prediction of the optimal model was virtually zero (.01).

Example 3

The files of the Oregon students who were later rated by the faculty were searched to obtain an average rating from the admissions committee that evaluated them before they were selected; this average rating correlated .19 with the later faculty ratings, whereas the cross-validated correlation based on regression analysis was .38 (Dawes; 1971).

Example 4

The average correlation between judges' estimates and the assigned values in the ellipse experiment was .84, whereas the value predicted from equal weighting (which is optimal) correlated .97 with the assigned values (Yntema & Torgerson, 1961).

There are a number of reasons why linear models perform so well. First, in these contexts each variable has a *conditionally monotone* relationship to the criterion. That is, the variables can be scaled in such a way that higher values on each predict higher values on the criterion, independently of the values of the remaining variables. As pointed out by Amos Tversky (personal communication, 1971), this condition is the combination of two more fundamental measurement conditions: (*a*) independence (the ordinal relationship between each variable and the criterion is independent of the values of the remaining variables) and (*b*) monotonicity (the ordinal relationship is one that is monotone). (See Krantz, 1972; Krantz, Luce, Suppes, & Tversky, 1971.) And linear models are good approximations to all multivariate models that are conditionally monotone in each predictor variable. Rorer (1971) and Dawes (1968) have jointly explored this degree of approximation by computer simulation. Using correlations between the output of various models that were nonlinear (but conditionally monotone) in each variable and the output of the linear approximations to these models, Rorer (1971) and Dawes (1968) discovered a high degree of fit between models and linear approximations. Even hierarchical models and models involving multiple cut procedures were well approximated by linear models (as evaluated by correlation coefficients).

One reason then that linear models perform so well is that they have been investigated in contexts in which true relationships, whatever they are, tend to be conditionally monotone. No matter how psychiatric patients score on other variables, they are more

likely to be psychotic the higher they score on the schizophrenia scale, the higher they score on the paranoia scale, and the lower they score on the psychasthenia scale. No matter how graduate students score on other variables, they are more likely to do better the higher they score on the GRE, and so on. Moreover, variables that do not have a conditionally monotone relationship to the criterion variable tend to have a single peak relationship that is easily converted to a monotone relationship by changing from raw units to units of worth or predictability. For example, the job of custodian may require a certain amount of intelligence, but high levels of intelligence may result in poor performance because of boredom. An intelligence test may then be rescored in terms of the absolute distance from 100—that is, rescaled to measure "intellectual mediocrity." (It has, in fact, recently been suggested that such a variable may not only be relevant to selection of custodians, but to Supreme Court Justices as well.)

Second, the relative weights derived from a linear regression analysis are not affected by "error" in the criterion variable. Such error reduces the expected values of all these weights by the same constant amount and hence reduces the absolute value of the predicted criterion variable by that same amount. This linear transformation on the predicted value does not affect its correlation with the true score value. It does, of course, affect the correlation between predicted value and observed value.[3]

Third, error in the measurement of the independent variables tends to make optimal functions more linear—that is, curves separating values on the dependent variable tend to become flatter. In conjunction with Gold,[4] the present authors demonstrated this effect by considering the two-dimensional conditionally monotone function that is least well approximated by a linear function; this function is a conjunctive step function. When the variables are measured without error, this function consists of a rectangular contour separating high values from low values. As the independent variables are measured with an increasing amount of error, this contour becomes increasingly curved—eventually approximating a straight line. This curvature is demonstrated in Figure 1. The same effect was demonstrated earlier by Lord (1962), who proved that when a conjunctive step function (multiple cut) is appropriate under errorless measurement conditions, a sufficient amount of error dictates the use of a linear approximation in its place. (The reader who does not follow this brief description is referred to Lord's article.)

To summarize, linear functions are good approximations to conditionally monotone functions; the relative values of the weights are not affected by error in the criterion variable, and conditionally monotone functions tend to become more linear in the presence of increasing error in the predictor variables. Such models fit, then, because the contexts in which they are evaluated tend to be conditionally monotone contexts in which there is much error.

This conclusion—that linear models are often good approximations in many decision-making situations that psychologists study—is not original with this article. In discussing the evaluation of job applications, Thorndike (1918) wrote:

> The setting up of an equation of prophecy from an equation of status will usually be very complex,

[3] This conclusion is easily demonstrated algebraically when the variables are in standard score form. For $\beta = R^{-1}\mathbf{v}$, where β is a column vector of beta weights, R is the matrix of intercorrelations between predictor variables, and $\mathbf{v}$ is the column vector of "validities"—that is, intercorrelations between the predictors and the criterion. The intercorrelations in R are unaffected by the existence or nonexistence of error in the measurement of the criterion variable. What that error does, however, is to affect all the correlations in $\mathbf{v}$. Specifically let r'_i be the correlation that would be found between predictor i and criterion, if the criterion were measured without any error. Because correlation is equal to the covariance divided by the geometric mean of the variances, the actual correlation (r_i) when the criterion is measured with error is equal to $\alpha r'_i$, where α equals the square root of the ratio of true score variance in the criterion to total variance. Hence $\mathbf{v} = \alpha\mathbf{v}'$, where $\mathbf{v}'$ is the vector of validities that would be obtained were there no error. And it follows that $\beta = \alpha\beta'$, where β' is the vector of beta weights that would be obtained were there no error.

[4] E. Mark Gold was a contributing mathematician to the study.

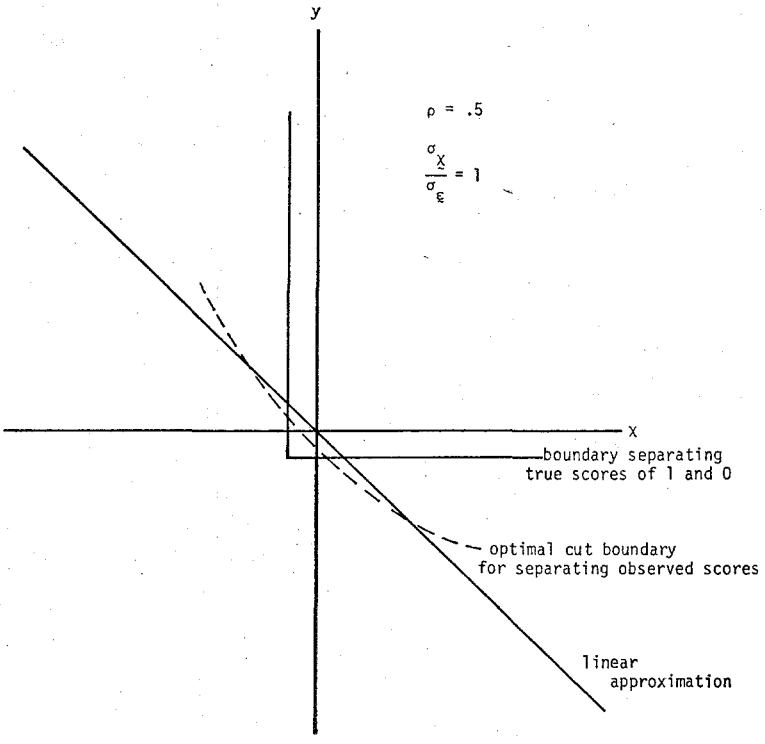

FIGURE 1. Conjunctive true score region, optimal cut boundary, and linear approximation.

but a rough [linear] approximation, if sound in principle, will often give excellent results. In so far as the lines of relation, interrelation, and dependency are rectilinear, the technique is greatly simplified; *and a rough approximation to this is probably often the case* [italics added; p. 75].

PARAMORPHIC LINEAR REPRESENTATION

In 1923, Henry A. Wallace (former Vice-President under Roosevelt) proposed that one method of determining "what is on the corn judge's mind" is to build a linear model of the judge by regressing his ratings of corn quality on various characteristics of the corn that he rates. Wallace's idea of analyzing the expert decision maker by constructing such a model apparently did not excite many readers at that time. Thirty-seven years later, Hoffman (1960) independently proposed that linear models could be used to represent expert judgment, and his proposal received a great deal of attention. Hoffman termed the linear model that he used to predict an expert's judgment a "paramorphic representation" of such judgment. The term was chosen because Hoffman did not mean to imply that the actual psychological process involved in making the judgment was that of weighing various variables, but rather that this process could be simulated by such a weighting. There were many cases in which the simulation was clearly inappropriate in that it predicted qualitative aspects of the judgment process that were not, in fact, discovered; the simulation was regarded as good paramorphic representation, however, if the output of the linear model corresponded to the output of the judge. Such linear models have been shown to be quite good paramorphic representations or, as some authors put it, quite good at "capturing the policy" of judges (Anderson, 1968; Beach, 1967; Christal, 1968; Dudycha & Naylor, 1966; Goldberg, 1968b; Hammond, Hursch, & Todd, 1964; Hoffman, Slovic, & Rorer, 1968; Hursch, Hammond, & Hursch, 1964; Naylor & Wherry, 1965; Nystedt & Magnusson, 1972; Schenck & Naylor,

1968; Slovic, 1969; Tucker, 1964; Wherry & Naylor, 1966; Wiggins & Hoffman, 1968). See Slovic and Lichtenstein (1971) for a recent review.

Does the success of such models indicate that the judges are nothing more than "linear machines"? The answer to this question hinges on whether or not the discrepancies between the actual judgments and those predicted by the linear model are reliable. If these discrepancies have no reliability, then the decision maker is behaving like a linear machine with an error component. If, on the other hand, these discrepancies can be shown to be reliable, then the decision maker is behaving in a consistently nonlinear way. Rorer and Slovic (1966) discovered that such deviations can be reliable, although completely unrelated to the criterion that the judge is attempting to predict! The reliability of the nonlinear component may also be assessed by comparing the correlation between the predictions of the model and the judge with the overall reliability of the judge. If the decision maker is acting like the linear model with an error component, the correlation between the model and the actual judgments should be equal to the square root of the reliability of these judgments. In conjunction with Winter,[5] the present author asked three judges from the admissions committee at the University of Oregon's psychology department to rerate 90 applicants, who had previously been rated on a 6-point scale to assess their suitability for entering the graduate program. The reliabilities of the three judges were .62, .69, and .68. Linear models of these judges' behavior were constructed on the basis of three predictor variables: undergraduate grade point average, GRE scores, and a rating of the undergraduate institution that the applicant attended. The judges' correlations with their linear models were .50, .75, and .79. Although these correlations did not approach the square root of the reliabilities, the study was marred in that the ratings were based on all the information in the applicant's folder and the linear models were not. Perhaps linear models based on more variables would lead to the conclusion that these judges behaved like linear machines.

[5] Ben Winter was a contributing mathematician to the study.

Bootstrapping

When there are actual criterion values against which the predictions of both the judge and the linear model of the judge can be compared, the paramorphic linear model often does a better job than does the judge himself. That is, the correlation between output of the model and criterion is often higher than the correlation between the decision maker's judgment and criterion, even though the model is based on the behavior of the decision maker. This "intriguing possibility" was first suggested by Yntema and Torgerson (1961). It was later demonstrated in a business context by Bowman (1963) and was eventually termed *bootstrapping*. Bootstrapping has turned out to be a rather pervasive phenomenon. For example, in the Wiggins and Kohen (1971) study, *the linear model of every one of their 80 University of Illinois judges did a better job than did the judges themselves in predicting actual grade point averages*. This result has been replicated for 40 of 41 University of Oregon judges making the same judgments (in a study conducted in conjunction with Wiggins, Gregory, & Diller[6]). Goldberg (1970) demonstrated it for 26 of 29 clinical psychology judges, and Dawes (1971) found it in the evaluation of graduate applicants at the University of Oregon.

Why does bootstrapping work? In 1963, Bowman wrote:

> It seems useful to attempt an explanation of why decision rules derived from management's own average behavior might yield better results than the aggregate behavior itself. Man seems to respond to selective cues in his environment—particular things seem to catch his attention at times (the last telephone call). . . . [These random and particularistic components can be eliminated] through the use of decision rules incorporating coefficients derived from management's own recurrent behavior [p. 316].

Working entirely independently on the prediction of neurosis and psychosis from MMPI profiles, Goldberg (1970) wrote:

[6] Nancy Wiggins, Sundra Gregory, and Richard Diller were contributing psychologists to the study.

TABLE 1

CORRELATIONS BETWEEN PREDICTIONS AND CRITERION VALUES

Example	Average validity of judge	Average validity of judge's model	Average validity of random model	Validity of equal weighting model	Cross-validated validity of regression analysis	Validity of optimal linear model
Prediction of neurosis versus psychosis	.28	.31	.30	.34	.46	.46
Illinois students' prediction of grade point average	.33	.50	.51	.60	.57	.69
Oregon students' prediction of grade point average	.37	.43	.51	.60	.57	.69
Prediction of later faculty ratings at Oregon	.19	.25	.39	.48	.38	.54
Yntema & Torgerson (1961) experiment	.84	.89	.84	.97	—	.97

For the clinician is not a machine. While he possesses his full share of human learning and hypothesis-generating skills, he lacks a machine's reliability. He "has his days": Boredom, fatigue, illness, situational and interpersonal distractions all plague him, with the result that his repeated judgments of the exact same stimulus configuration are not identical. . . . If we could remove some of this human unreliability by eliminating the random error in his judgments, we should thereby increase the validity of the resulting predictions. The problem, then, may be reformulated: Can the clinician's judgmental unreliability be separated from his—hopefully, somewhat valid—judgmental strategy [p. 423]?

Goldberg's answer was yes; the means of separation was by constructing a linear paramorphic representation of the judge.

In 1971, Dawes wrote:

A mathematical model, by its very nature, is an abstraction of the process it models; hence, if the decision maker's behavior involves following valid principles but following them poorly, these valid principles will be abstracted by the model—as long as the deviations from these principles are not systematically related to the variables the decision maker is considering [p. 182].

AN END TO BOOTSTRAPPING: RANDOM LINEAR MODELS

Belief in the efficacy of bootstrapping was based on a comparison of the validity of the linear model of the judge with the validity of his (or her) judgments themselves. That was only one of two logically possible comparisons. The other is between the validity of the linear model of the judge and the validity of linear models in general. That is, to demonstrate that bootstrapping works because the linear model catches the essence of a judge's expertise and at the same time eliminates unreliability, it is necessary to demonstrate that the weights obtained from an analysis of the judge's behavior are superior to those that might be obtained in other ways—for example, obtained randomly. In the four examples discussed in this article, there is no evidence of such superiority.

In each example, the authors constructed random linear models to predict the criterion. The sign of each predictor variable was determined on an a priori basis so that it would have a positive relationship to the criterion. Then a normal deviate was selected at random from a normal distribution with unit variance, and the absolute value of this deviate was used as a weight for the variable. Ten thousand such models were constructed for each example. The results are presented in Table 1 along with the earlier results. On the average, correlations between the criteria and the output predicted from the random models were higher than those obtained from the judges' models. The present authors also investigated equal weighting and, of course, discovered that such weighting was even better.[7] (In two of the four examples, the

[7] This result follows from a simple inequality: If several standardized predictor variables all have a positive correlation with a criterion variable, the correlation between the average of the predictor variables and the criterion will be higher than the average correlation between predictor and criterion (see Ghiselli, 1964). Here an equal weighting scheme gives the same output as does the average of all random models—all of which have positive validity. Hence it has a correlation higher than the average

TABLE 2

CORRELATIONS BETWEEN PREDICTIONS AND LINEAR MODELS

Example	Average r of judge with optimal linear model	Average r of judge's model with optimal linear model	Average r of random model with optimal linear model	r of equal weighting model with optimal linear model	r of split composite with optimal linear model
Prediction of neurosis versus psychosis	.53[a]	.67	.65	.74	1.00
Illinois students' prediction of grade point average	—	.72	.74	.87	.83
Oregon students' prediction of grade point average	.53[a]	.62	.74	.87	.83
Prediction of later faculty ratings at Oregon	—	.46	.72	.89	.70
Yntema & Torgerson (1961) experiment	—	.92	.87	1.00	—

[a] Empirically derived.

equal weighting scheme had a higher correlation with the criterion than did the cross-validated optimal weighting scheme. This anomalous result is explained by the fact that the ratio of observations of variables was too low to obtain stable beta weights in the actuarial analysis; in practice stepwise regression would be used and fewer variables weighted).

Essentially the same results were obtained when the weights were selected from a rectangular distribution. Why? Because linear models are robust not only in the three ways described earlier in this article, but they are robust over deviations from optimal weighting as well. In other words, the bootstrapping finding may be simply a reaffirmation of the earlier finding that linear models are superior to human judgments—the weights derived from judges' behavior being sufficiently close to the optimal weights so that the outputs of the models are highly similar. In other words, the solution to the problem of obtaining optimal weights is one that—in terms of von Winterfeldt and Edwards [8]—has a "flat maximum." Weights that are *near* to optimal lead to almost the same output as do optimal beta weights. The behavior of the expert judge, because he (or she) knows at least something about the direction of the variables, yields weights near optimal. (But note that in all cases equal weighting is superior to the models based on judges' behavior.)

This explanation for the efficacy of the models is illustrated in Table 2, which presents the correlation between the models and the optimal linear model (not cross-validated).[9] The table also presents the correlation between judges' predictions and those from the optimal linear models' predictions. These correlations also yield partial correlations approaching zero; it follows, as noted earlier, that a linear synthesis of the optimal linear model with the judges' estimates would not improve on the optimal linear model.

As von Winterfeldt and Edwards (see Footnote 5) pointed out, the questions of "What is flat?" and "How flat is flat?" are not well defined mathematically. Here, however, we wish to point out that even a linear model based on a single predictor has a peak that

correlation of the random models. Also see Dawes (1970).

[8] D. von Winterfeldt and W. Edwards. Costs and Payoffs in Perceptual Research. Unpublished manuscript, University of Michigan (Engineering Psychology Laboratory), 1973.

[9] The correlations are uniformly high. These correlations can be derived directly by a comparison of the validity of the model with the validity of the optimal linear model. If we were to predict the criterion from a linear composite of the optimal linear model and the nonoptimal linear model, the beta weight given to the nonoptimal model would be zero because it would be impossible to improve on the linear prediction from the optimal model. Hence the partial correlation between nonoptimal model and criterion partialling out the optimal linear model must also be zero (Hays, 1963, p. 575). The correlation between actual and optimal may then be computed by setting the numerator of the formula for the partial correlation coefficient equal to zero.

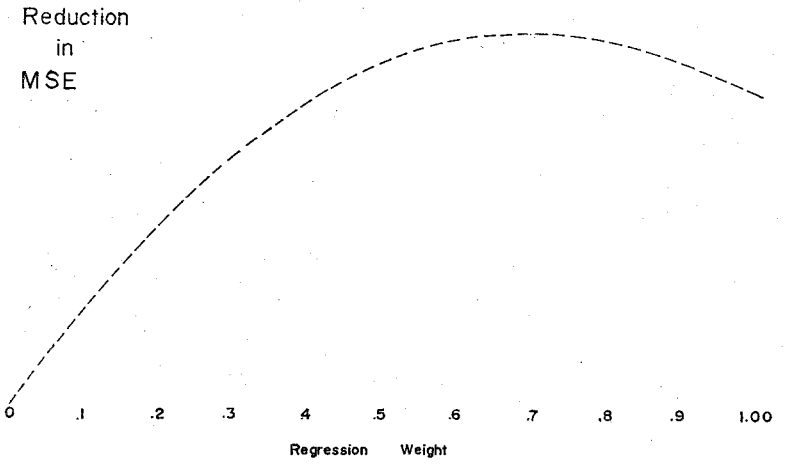

FIGURE 2. Reduction in mean square error (MSE) as a function of the believed correlation coefficient.

might generally be regarded as flat. Suppose that this single predictor variable correlates .71 with the criterion variable; our best prediction is, therefore, that the standard score on the criterion variable equals .71 times the standard score on the predictor variable, and the mean square error of prediction is given by $1 - r^2 = .50$.

Now suppose that the correlation between predictor and criterion is believed to be a rather than r. In such a case, the prediction is now that the standard score of the criterion variable is equal to a times the standard score of the predictor variable, where $a \neq r$ but $a = r + c$. The new mean square error of prediction is equal to $1 - r^2 + c^2$, which is equal to only .60 if c is .30 (i.e., if the correlation of .71 is believed to be 1.01 or .41). So a rather grievous error in estimating the correlation coefficient results in an increase in mean square error of prediction of only 20%. Figure 2 presents reduction in mean square error as a function of the believed correlation coefficient when the true correlation coefficient is .71. The maximum appears rather flat.

UNIT RATING

In the four examples discussed in this article, unit weighting did extremely well in predicting the criterion values. Many past investigators have also found that unit weighting does well in a variety of contexts. It is accepted as almost axiomatic that items forming a scale should be given unit weighting rather than be weighted by validities or covariances (Berdie & Campbell, 1968; Wang & Stanley, 1970). Unit weighting has also been advocated in situations in which populations change from time to time—as in evaluating officer candidates in New Zealand during World War II (Wrigley, personal communication, 1972). And such advocacy has been supported by empirical studies (Lawshe & Schucker, 1959; Trattner, 1963; Wesman & Bennett, 1959), all showing that equal weighting does as well as optimal weighting when the weights are applied to a new sample.

Recently, Schmidt (1971) has shown that equal weighting may be superior to optimal weighting schemes even when the cross-validation is performed on samples from the same (theoretical) population. In his simulation studies, Schmidt found that in the presence of suppressor variables the ratio of observations to predictors should be approximately 15 to 1 before optimally derived weights are superior to unit weights in cross-validation and, in the absence of suppressors, this ratio should be 25 to 1. In a similar study, Marks [10] found that a ratio of approximately 20 to 1 was necessary. (Marks's simulations had the

[10] M. R. Marks. Two Kinds of Regression Weights which are Better than Betas in Cross Samples. Paper presented at the annual meeting of the American Psychological Association, New York, September 1966.

specific property that the partial correlation between any two predictors partialling out the criterion variable was zero.)

In short, given the fact that in many contexts equal weights yield predictions very highly correlated with those obtained from optimal weights, equal weights may be superior. In contrast (Meehl, personal communication, 1972), beta coefficients are extremely unstable and most extrapolations are to samples from populations that differ somewhat from those on which the betas are estimated. Meehl (personal communication, 1972) concluded "in most practical situations an unweighted sum of a small number of 'big' variables will, on the average, be preferable to regression equations." [11]

Conclusion

Linear models work because the situations in which they have been investigated are those in which: (*a*) The predictor variables have conditionally monotone relationships to criteria (or may easily be rescaled to have such a relationship); (*b*) there is error in the dependent variable; (*c*) there is error in the independent variables; and (*d*) deviations from optimal weighting do not make much practical difference. These situations abound. (It is always better to be smarter, more beautiful, closer to age 29, closer to blood pressure 120 over 80, etc.) Thus the situation demands decision-making behavior approximately like that of a linear model if the decision making is to be appropriate—in other words, an analysis of the task faced by the decision maker (Edwards, 1971; Simon, 1969) leads to the conclusion that linear models work well. It is, therefore, not surprising that linear models outperform intuitive judgment. Nor is it surprising that decision makers (insofar as they are behaving appropriately) are paramorphically well represented by linear models. Again, to quote Thorndike (1918):

> There is a prevalent myth that the expert judge of men succeeds by some mystery of divination. Of course, this is nonsense. He succeeds because he makes smaller errors in the facts or in the way he weights them. Sufficient insight and investigation should enable us to secure all the advantages of the impressionistic judgment (except its speed and convenience) without any of its defects [p. 76].

The whole trick is to decide what variables to look at and then to know how to add.

[11] Trites and Sells (1955) also found equal weighting appropriate for estimating factor scores.

References

Anderson, N. H. A simple model for information integration. In R. P. Abelson, E. Aronson, W. J. McGuire, T. M. Newcomb, M. J. Rosenberg, & P. H. Tannenbaum (Eds.), *Theories of cognitive consistency: A sourcebook*. Chicago: Rand McNally, 1968.

Beach, L. R. Multiple regression as a model for human information utilization. *Organizational Behavior and Human Performance,* 1967, **2**, 274-289.

Berdie, R. F., & Campbell, D. P. Measurement of interest. In D. K. Whitla (Ed.), *Handbook of measurement and assessment in behavioral sciences*. Reading, Mass.: Addison-Wesley, 1968.

Bigelow, J. (Ed.) *The complete works of Benjamin Franklin*. Vol. 4. New York: Putnam, 1887.

Bowman, E. H. Consistency and optimality in managerial decision making. *Management Science,* 1963, **9**, 310–321

Christal, R. E. Selecting a harem—and other applications of the policy-capturing model. *Journal of Experimental Education,* 1968, **36**, 35–41.

Dawes, R. M. Algebraic models of cognition. In C. A. J. Vlek (Ed.), *Algebraic models in psychology: Proceedings of the NUFFIC international summer session in science*. The Hague: Netherlands Universities Foundation for International Cooperation, 1968.

Dawes, R. M. An inequality concerning correlation of composites vs. composites of correlations. *Oregon Research Institute Methodological Note,* 1970, **Vol. 1, No. 1.**

Dawes, R. M. A case study of graduate admissions: Application of three principles of human decision making. *American Psychologist,* 1971, **26**, 180–188.

Dudycha, A. L., & Naylor, J. C. The effect of variations in the cue R-matrix upon the obtained policy equations of judges. *Educational and Psychological Measurement,* 1966, **26**, 583–603.

Edwards, W. Bayesian and regression models in human information processing—a myopic perspective. *Organizational Behavior and Human Performance,* 1971, **6**, 639–648.

Einhorn, H. J. Expert measurement and mechanical combination. *Organizational Behavior and Human Performance,* 1972, **7**, 86–106.

Ghiselli, E. E. *Theory of psychological measurement*. New York: McGraw-Hill, 1964.

Goldberg, L. R. Diagnosticians vs. diagnostic signs: The diagnosis of psychosis vs. neurosis from the MMPI. *Psychological Monographs,* 1965, **79** (9, Whole No. 602).

Goldberg, L. R. Seer over sign: The first "good" example? *Journal of Experimental Research in Personality,* 1968, **3**, 168–171. (a)

GOLDBERG, L. R. Simple models or simple processes? Some research on clinical judgments. *American Psychologist,* 1968, **23**, 483–496. (b)

GOLDBERG, L. R. Man versus model of man: A rationale, plus some evidence, for a method of improving on clinical inferences. *Psychological Bulletin,* 1970, **73**, 422–432.

HAMMOND, K. R., HURSCH, C. J., & TODD, F. J. Analyzing the components of clinical inferences. *Psychological Review,* 1964, **71**, 438–456.

HAYS, W. L. *Statistics for psychologists.* New York: Holt, Rinehart & Winston, 1963.

HOFFMAN, P. J. The paramorphic representation of clinical judgment. *Psychological Bulletin,* 1960, **57**, 116–131.

HOFFMAN, P. J., SLOVIC, P., & RORER, L. G. An analysis-of-variance model for the assessment of configural cue utilization in clinical judgment. *Psychological Bulletin,* 1968, **69**, 338–349.

HURSCH, C. J., HAMMOND, K. R., & HURSCH, J. L. Some methodological considerations in multiple-cue probability studies. *Psychological Review,* 1964, **71**, 42–60.

KRANTZ, D. H. Measurement structures and psychological laws. *Science,* 1972, **175**, 1427–1435.

KRANTZ, D. H., LUCE, R. D., SUPPES, P., & TVERSKY, A. *Foundations of measurement.* Vol. 1. New York: Academic Press, 1971.

LAWSHE, C. H., & SCHUCKER, R. E. The relative efficiency of four test weighting methods in multiple prediction. *Educational and Psychological Measurement,* 1959, **19**, 103–114.

LORD, F. M. Cutting scores and errors of measurement. *Psychometrika,* 1962, **27**, 19–30.

MEEHL, P. E. *Clinical versus statistical prediction: A theoretical analysis and review of the literature.* Minneapolis: University of Minnesota Press, 1954.

MEEHL, P. E. *Clinical versus statistical prediction:* ple. *Journal of Experimental Research in Personality,* 1965, **1**, 27–32.

NAYLOR, J. C., & WHERRY, R. J. The use of simulated stimuli and the "JAN" technique to capture and cluster the policies of raters. *Educational and Psychological Measurement,* 1965, **25**, 969–986.

NYSTEDT, L., & MAGNUSSON, D. Predictive efficiency as a function of amount of information. *Multivariate Behavioral Research,* 1972, **7**, 441–450.

PANKOFF, L. A quantification of judgment: A case study. Unpublished doctoral dissertation, University of Chicago, 1967.

PANKOFF, L. B., & ROBERTS, H. V. Bayesian synthesis of clinical and statistical prediction. *Psychological Bulletin,* 1968, **70**, 762–773.

RORER, L. G. A circuitous route to bootstrapping. In H. B. Haley, A. G. D'Costa, & A. M. Schafer (Eds.), *Conference on personality measurement in medical education.* Washington, D.C.: Association of American Medical Colleges, 1971.

RORER, L. G., & SLOVIC, P. The measurement of changes in judgmental strategy. *American Psychologist,* 1966, **21**, 641–642. (Abstract)

SARBIN, T. R. Contribution to the study of actuarial and individual methods of prediction. *American Journal of Sociology,* 1943, **48**, 593–602.

SAWYER, J. Measurement and prediction, clinical *and* statistical. *Psychological Bulletin,* 1966, **66**, 178–200.

SCHENCK, E. A., & NAYLOR, J. C. A cautionary note concerning the use of regression analysis for capturing the strategies of people. *Educational and Psychological Measurement,* 1968, **28**, 3–7.

SCHMIDT, F. L. The relevant efficiency of regression in simple unit predictor weights in applied differential psychology. *Educational and Psychological Measurement,* 1971, **31**, 699–714.

SIMON, H. A. *The sciences of the artificial.* Cambridge, Mass.: MIT Press, 1969.

SLOVIC, P. Analyzing the expert judge: A descriptive study of a stockbroker's decision processes. *Journal of Applied Psychology,* 1969, **53**, 255–263.

SLOVIC, P., & LICHTENSTEIN, S. Comparison of Bayesian and regression approaches to the study of information processing in judgment. *Organizational Behavior and Human Performance,* 1971, **6**, 649–744.

THORNDIKE, E. L. Fundamental theorems in judging men. *Journal of Applied Psychology,* 1918, **2**, 67–76.

TRATTNER, M. H. Comparison of three methods for assembling aptitude test battery. *Personnel Psychology,* 1963, **16**, 221–232.

TRITES, D. K., & SELLS, S. B. A note on alternative methods for estimating factor scores. *Journal of Applied Psychology,* 1955, **39**, 455–456.

TUCKER, L. R. A suggested alternative formulation in the developments by Hursch, Hammond, and Hursch, and by Hammond, Hursch, and Todd. *Psychological Review,* 1964, **71**, 528–530.

WALLACE, H. A. What is in the corn judge's mind? *Journal of the American Society of Agronomy,* 1923, **15**, 300–304.

WANG, M. D., & STANLEY, J. C. Differential weighting: A review of methods in empirical studies. *Review of Educational Research,* 1970, **40**, 663–705.

WESMAN, A. G., & BENNETT, G. K. Multiple regression versus simple addition of scores in prediction of college grades. *Educational and Psychological Measurement,* 1959, **19**, 243–246.

WHERRY, R. J., & NAYLOR, J. C. Comparison of two approaches—JAN and PROF—for capturing rater strategies. *Educational and Psychological Measurement,* 1966, **26**, 267–286.

WIGGINS, N., & HOFFMAN, P. J. Three models of clinical judgment. *Journal of Abnormal Psychology,* 1968, **73**, 70–77.

WIGGINS, N., & KOHEN, E. S. Man vs. model of man revisited: The forecasting of graduate school success. *Journal of Personality and Social Psychology,* 1971, **19**, 100–106.

YNTEMA, D. B., & TORGERSON, W. S. Man-computer cooperation in decisions requiring common sense. *IRE Transactions of the Professional Group on Human Factors in Electronics,* 1961, HFE-2(1), 20–26.

(Received August 27, 1973)