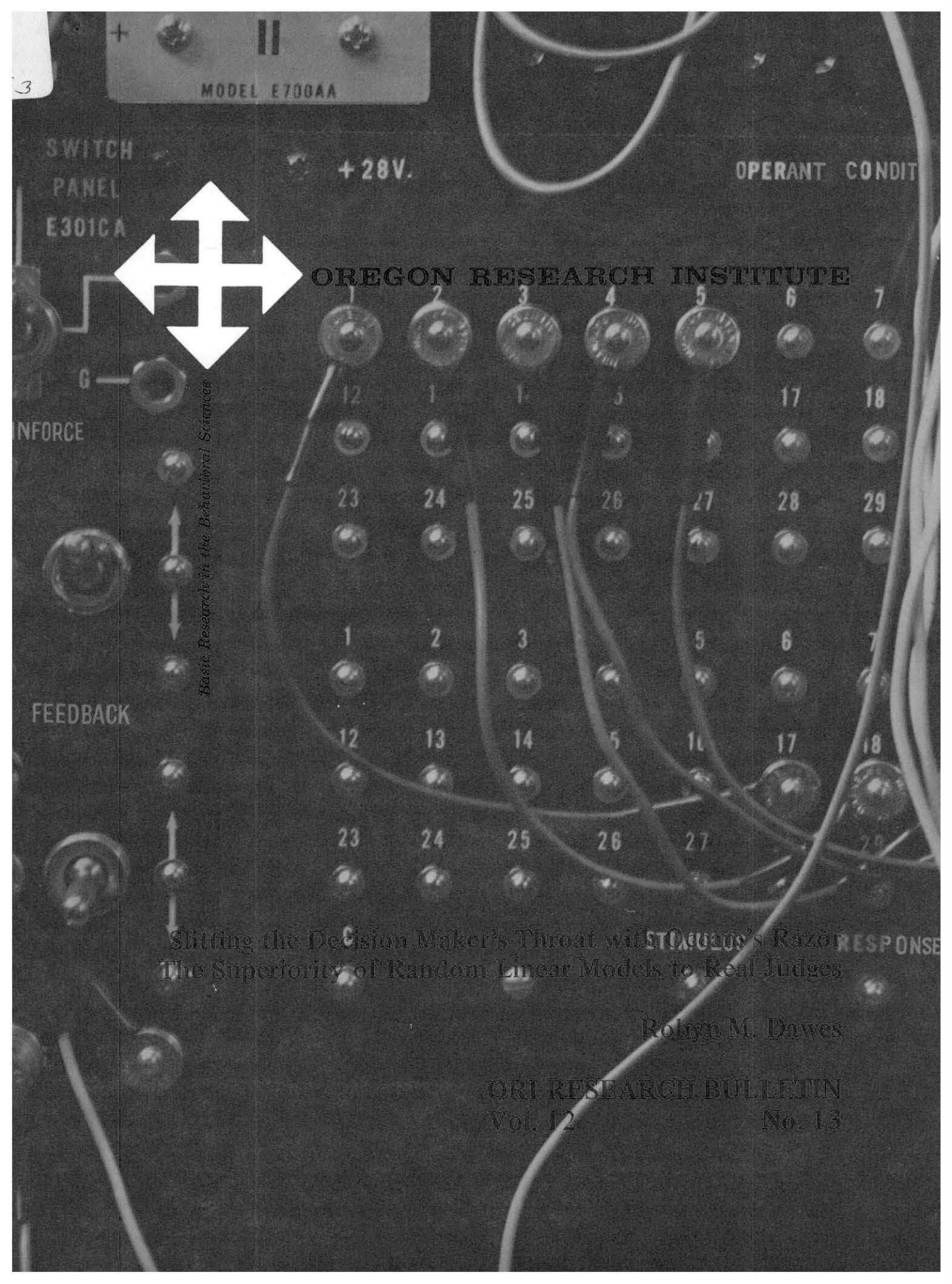

Basic Research in the Behavioral Sciences

Slitting the Decision Maker's Throat with Occam's Razor
The Superiority of Random Linear Models to Real Judges

Robyn M. Dawes

ORI RESEARCH BULLETIN
Vol. 12 No. 13

OREGON RESEARCH INSTITUTE

Slitting the Decision Maker's Throat with Occam's Razor

The Superiority of Random Linear Models to Real Judges

Robyn M. Dawes

University of Oregon and Oregon Research Institute

Research Bulletin
Vol. 12 No. 13
December, 1972

SLITTING THE DECISION MAKER'S THROAT WITH OCCAM'S RAZOR--
THE SUPERIORITY OF RANDOM LINEAR MODELS TO REAL JUDGES[1]

Robyn M. Dawes
University of Oregon and Oregon Research Institute

Abstract

Two types of linear models have been found to be superior to human judges in predicting codable criterion variables from codable predictor variables: actuarial models (based on the regression of the criterion in the predictors), and bootstrapping models (based on the regression of the judges' predictions on the predictors). In both types of models, the predictors are coded (or recoded) so that each bears a monotonic relationship to the criterion; further, in both models, nonlinear relationships and nonmonotone interactions between variables are ignored. Perhaps these linear models are superior because the expected validity of any linear model whose weights are in the appropriate direction is superior to that of human judges in this context. Present research reviews a variety of situations in which linear models whose weights are in the appropriate direction but otherwise chosen at random do better than do judges, even bootstrapped judges.

SLITTING THE DECISION MAKER'S THROAT WITH OCCAM'S RAZOR--
THE SUPERIORITY OF RANDOM LINEAR MODELS TO REAL JUDGES[1]

Robyn M. Dawes
University of Oregon and Oregon Research Institute

This talk will be concerned with the specific problem of predicting a codable output from multivariate codable input. The emphasis will be on socially interesting output predicted from socially interesting input--specifically, on the type of prediction most often made by a human judge. The central question to be asked is: what role should the human judge play in this prediction if it is to be made in an optimal manner? The answer will be: no role whatsoever, or at least no role other than that of deciding which variables should be included in the input.

Let me give four examples of such prediction; these will be discussed at greater length later in this paper.

Example 1. A pool of roughly 1200 psychiatric patients took the Minnesota Multiphasic Personality Inventory (MMPI) at various hospitals and later were categorized as "neurotic" or "psychotic" on the basis of more extensive information. ("Neurotic" refers to someone who is emotionally disturbed but who is basically in contact with external social reality when he or she wishes to be--and has an emotional investment in it; "psychotic" refers to someone who has withdrawn investment from the social world and has lost contact with it.) The MMPI results are in the form of a personality "profile" of eleven 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. (Most patients present a "mixed" pattern.) Thus, a vector of eleven scores is associated with each patient, and the problem is to predict whether a later diagnosis will be psychotic (coded 1) or neurotic (coded 0).

The second example concerns a prediction of first year graduate GPA of students admitted to the Psychology Department at the University of Illinois. Again, there is a profile associated with each student--this time consisting of 10 scores on such variables as aptitude test performance, college GPA, peer ratings of extroversion, self-ratings of conscientiousness and achievement motivation, and so on. And again, the problem is to predict the univariate output (GPA) from the multivariate input (the profiles).

The third example also concerns prediction of success in graduate school; the success is, however, in a different location--the Oregon Psychology Department instead of the Illinois one; it is evaluated in a different way--by faculty ratings of students who had been in the program from two to five years rather than by GPA; and the number of predictor variables is reduced to the three available to the Admissions Committee: scores on the Graduate Record Exam (GRE), undergraduate GPA, and an approximate rating of the quality of the institution at which the GPA was obtained.

Example 4. In a laboratory situation, experimenters assigned scores (values) to ellipses they had drawn on the basis of each figure's size, eccentricity, and grayness. (The actual formula used was $ij + jk + ik$ where I, J, K refer to 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 is to predict the true (i.e., experimenter assigned) value of each ellipse.

How best can these predictions be made? To answer this question, it is necessary to distinguish between two situations: that in which we have a "reasonable" sample of cases in which the output values are known, and that in which we do not know the output values.

Output Values Known

If a reasonable sample of cases exists for which the output values are known, the best way to make the predictions is to estimate beta weights for the input variables on the basis of multiple regression; human judges should be ignored. I state this answer with a great deal of confidence, because it is based on a great deal of research.

One of the first areas to be investigated by clinical psychologists--when clinical psychology became a profession after World War II--was the degree to which human "clinical" judgment about patients could be used in the prediction of such variables as response to treatment, recidivism, and so on. The basic idea was to see what such clinical judgment added to prediction that could be made on a purely statistical basis by developing regression equations or Bayesian prediction schemes; in other words, the actuarial analysis was thought to provide a floor to which the judgment of the experienced clinician could be compared.

It very quickly turned out that the floor was a ceiling. In his influential 1954 book, Paul Meehl reviewed roughly 20 studies in which actuarial methods were pitted against the judgment of the clinician, and in all cases the actuarial method either won the contest, or both methods tied. Since that book was published, there has been a plethora of additional studies oriented toward the question of whether clinical judgment is superior to the actuarial combination (Sawyer, 1966), and some of these studies have been quite extensive

(Goldberg, 1965). But Meehl was able to conclude over 10 years after his book was published that there was only a single example in the literature showing clinical judgment to be superior (Meehl, 1965), and this conclusion was immediately disputed by Goldberg (1968a) on the grounds that even that example did not show such superiority. I know of no examples after that that have purported to show the superiority of clinical judgment.

Two constraints in all these studies must be noted. First, the clinical judgment and the actuarial prediction are made on the basis of the same codable input. A clinical judge may have access to variables that are not clearly codable, or he may have access to variables that are, but that cannot be assessed without his presence--e.g., his feeling of liking or disliking the patient. Such variables are not included in these studies. Second, the question of validity has always been settled by comparison of correlation coefficients--one of the correlation between judge's prediction and reality, the other the cross-validated correlation between the predictions of the regression equation or Bayesian analysis and reality. (The weights are derived on the basis of a multiple regression analysis of one sample, and they are applied to a second sample to form the linear composite of the predictor variables; the correlation of this composite with the criterion variable may then be interpreted as a simple Pearson coefficient.) But no one has proposed an alternative way of comparing predictability, and correlation--since it is a good representation of the degree to which two variables are rank ordered similarly--seems quite reasonable for situations such as assessing graduate students or predicting which patients are best candidates for which types of therapy.

The examples mentioned earlier in this paper are no exception. Twenty-nine experienced judges (Ph.D.'s or advanced students in clinical psychology) were asked to predict whether the patients were neurotic or psychotic; they

were asked to rate the patients in a forced normal distribution. The average correlation coefficient of their rating with the binary criterion was .28, while the cross-validated correlation of the weighting scheme derived from regression analysis was .46 (Goldberg, 1965). Eighty University of Illinois students were asked to predict the GPA's of 90 other first-year students in psychology; the average correlation between predicted and obtained GPA was .33, while the cross-validated correlation resulting from regression analysis was .57 (Wiggins & Kohen, 1971). Forty-one graduate students at the University of Oregon making the prediction had an average correlation of .37, which is again less than that obtained from the regression analysis. The files of the Oregon students who were later rated by the faculty were searched to obtain--where possible--an average rating from the Admissions Committee that evaluated them before they were selected; this average rating correlated .19 with the later faculty ratings, while the cross-validated correlation based on regression analysis was .38 (Dawes, 1971). Finally, the average correlation between judge's estimate and assigned value in the ellipse experiment was .84, while the value predicted from an equal weighting scheme is .97 (Yntema & Torgerson, 1961).

Output Values Unknown

Even where there is not a reasonable sample of output values from which to derive regression weights, raw human judgment may be improved upon. The method is to build a linear "paramorphic" representation of the human judge (Wallace, 1923; Hoffman, 1960) and then to use this representation in place of the judge (Goldberg, 1970; Dawes, 1971; Wiggins & Kohen, 1971). Such linear representations have been shown to be quite good at "capturing the policy of the judge" (Hammond, Hursch & Todd, 1964; Tucker, 1964; Hursch, Hammond &

Hirsch, 1964; Naylor & Wherry, 1965; Dudycha & Naylor, 1966; Wherry & Naylor, 1966; Beach, 1967; Schenck & Naylor, 1968; Goldberg, 1968b; Hoffman, Slovic & Rorer, 1968; Wiggins & Hoffman, 1968; Anderson, 1968; Christal, 1968; Slovic, 1969). The reason that such models should be superior to the judges themselves will be discussed at greater length later in this paper.

The Ubiquity of the Linear Model

In a paper presented at the fall '71 meetings of the Society of Multivariate Experimental Psychologists, I proposed that linear models work so well primarily because of three factors.

First, linear models are good approximations of any multivariate models that are conditionally monotone in each variable. Len Rorer and I have explored the 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 approximation to these models, we discovered a high degree of fit between models and linear approximations. One reason, then, that linear models do so well in the context in which they have been investigated is that these contexts are ones in which the true relationship--whatever it is--tends 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 depression 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.

Second, the relative weights derived from a linear regression analysis are impervious to "error" in the criterion variable; that is, the expected values of the beta weights are merely reduced by a constant factor when the criterion variable is measured with error, and since the multiple correlation is simply the Pearson Product Moment correlation between predicted and observed criterion values, the correlation between predicted criterion values and true score values on the criterion is unchanged by the fact that the predicted values are reduced by a constant factor. What happens is that error in the dependent variable has no effect on the matrix of intercorrelations between independent variables, but reduces all the correlations between independent and dependent variables by a constant amount, and hence the beta weights by a constant amount. (Of course, error in the dependent variable results in greater error in estimating these beta weights.)

Third, error in the independent variables tends to make optimal functions more linear; by this statement I mean that curves separating values on the dependent variables tend to become flatter. Mark Gold and I demonstrated this effect by considering the most nonlinear conditionally monotone function possible in two dimensions--a conjunctive step function. When the variables are measured without error, this function consists simply of a rectangular contour separating ones from zeros. As the independent variables are measured with an increasing amount of error, this contour becomes increasingly curved--eventually approximating a straight line. See Figure 1. We discovered later that this same effect had been demonstrated earlier by Frederic Lord (1962).

Insert Figure 1 about here

To summarize, linear functions are good approximations to conditionally monotone functions; they are impervious to error in the criterion variable,

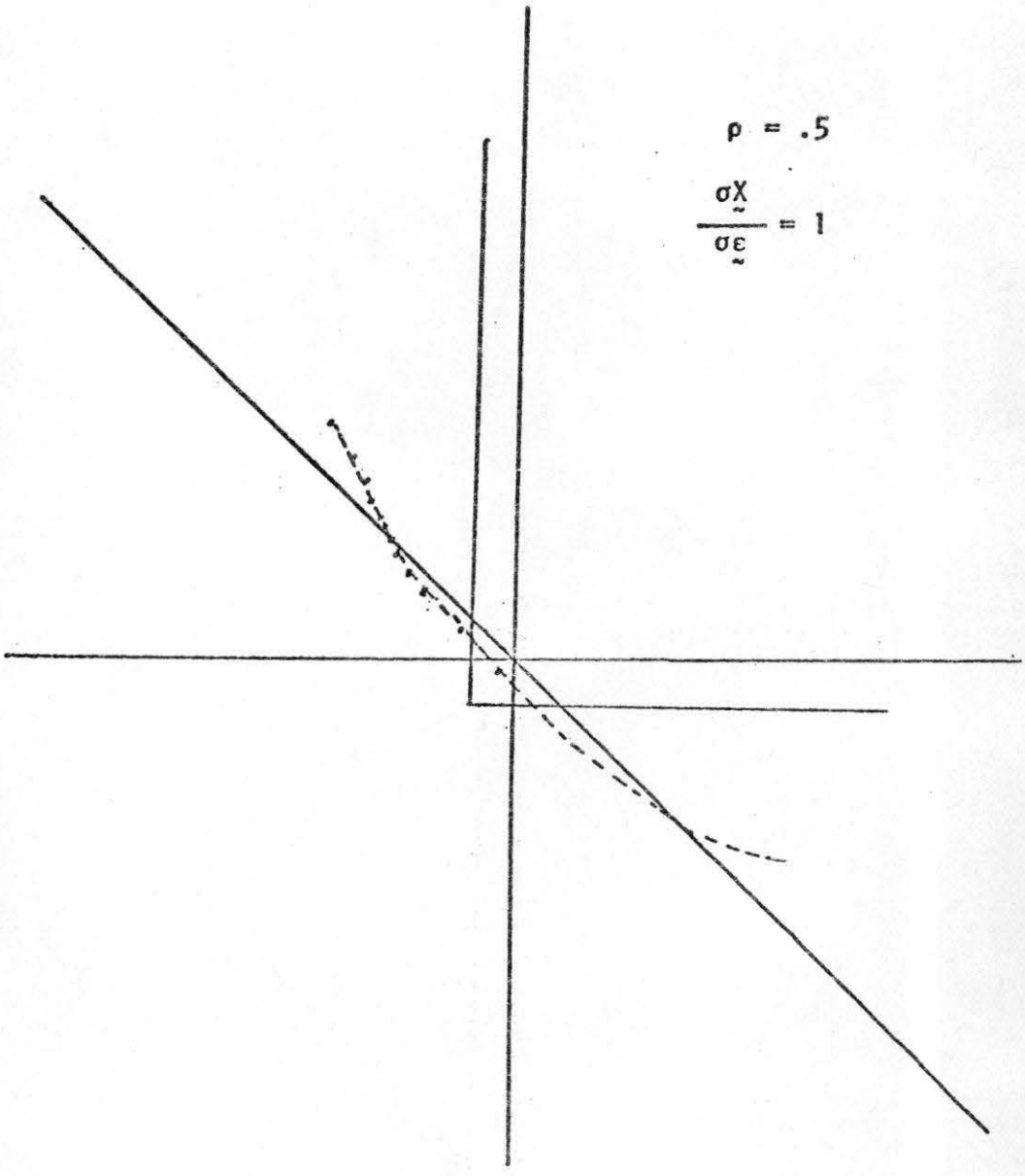

Figure 1

Conjunctive true score region, optimal cut boundary, and linear approximation

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 we evaluate them tend to be conditionally monotone contexts in which there is a lot of error.

And why should the linear representation of the judge do better than does the judge himself? This "intriguing possibility," which was first suggested by Yntema & Torgerson (1961) and later termed "bootstrapping," has turned out to be rather pervasive. For example, in the Wiggins & Kohen (1971) study the linear model they constructed of every single one of their 80 University of Illinois judges did a better job than did the judges themselves in predicting actual GPA's. Wiggins, Gregory, Diller, and I replicated the effect for 40 of 41 University of Oregon judges; Goldberg (1970) demonstrated it for 26 of 29 clinical psychology judges, and I (Dawes, 1971) found it in the evaluation of graduate applicants at the University of Oregon.

Why does bootstrapping work? In 1971 I wrote "the answer is that 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. For example, a decision maker may be weighting aptitude, past performance, and motivation correctly in predicting performance in graduate school and beyond, yet he may be influenced by such things as fatigue, headaches, boredom, and so on; in addition, he will be influenced by whether the most recent applications he has seen are particularly strong or weak. A paramorphic representation of his behavior would not be affected by these extraneous variables." (Page 182)

I was wrong. In the back of my mind there was a nagging question, but one that seemed so absurd to me at the time that I never bothered to answer it until recently. Could it be that bootstrapping worked because linear models in general are superior to human judges? And that it does not matter that the weights of the linear models were derived from an expert judge? What would happen if I selected weights at random--although in the appropriate direction--and constructed linear models based on these weights? One day, when I couldn't think of any better usage of his time, I asked one of my research assistants talented in computer programming to do just that. For each predictor variable he selected a normal deviate at random from a normal distribution with unit variance and then used the absolute value of this deviate as the weight for the variable. Each variable had previously been scaled on an a priori basis so that it had a positive relationship to the output.

The results surprised hell out of me. In the examples discussed in this paper, linear models whose weights were randomly selected in the manner just described on the average outperformed the linear models whose weights were selected on the basis of the experts' judgments. That is, the correlations between predicted output and obtained output were on the average higher. In each case, we constructed 10,000 such random models--so the superiority is quite well established. And, of course, we discovered that equal weighting schemes were better yet.[2] These results are presented in Table 1, along with the earlier results. (In two of the four examples, the equal weighting scheme even 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 to variables was too low to obtain stable beta weights in the actuarial analysis; in practice we would use stepwise regression and weight fewer variables.)

Insert Table 1 about here

Essentially the same results were obtained when beta weights were selected from a rectangular distribution.

Why? I believe the answer is that linear models are robust not only in the three ways described earlier in this paper, but that 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 judges--the weights derived from expert judges being sufficiently close to the optimal weights that the outputs of the models are highly similar. (I say may, because I am not quite yet willing to give up that beautiful Pythagorean idea that by building a mathematical model of the expert judge we "distill the essence" of his expertise.) In other words, the linear regression problem may be one--in terms of von Winterfeldt & Edwards (1972)--that has a "flat maximum." Beta weights that are near to optimal lead to almost the same output as do optimum beta weights; the behavior of the expert judge--because he or she knows at least something about the direction and magnitude of the variables--yields beta weights near the optimal ones; the result is a linear model that makes predictions highly similar to those of the optimal linear model.

As von Winterfeldt & Edwards point out (1972), the questions of "what is flat?" and "how flat is flat?" are not well defined mathematically. Here, I would just like to present some quasi-mathematical arguments that lead to the conclusion that slightly incorrect beta weights don't matter very much practically.

Consider a linear model based on a single predictor variable. Let us suppose that this variable correlates .71 with the criterion variable, and

Table 1

	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
Prediction of Neurosis vs. Psychosis	.28	.31	.30	.34	.46
Illinois Students' Prediction of GPA	.33	.50	.51	.60	.57
Oregon Students' Prediction of GPA	.37	.43	.51	.60	.57
Prediction of Later Faculty Ratings at Oregon	.19	.25	.39	.48	.38
Yntema & Torgerson Experiment	.84	.89	.84	.97	.97

that our best prediction is therefore that the standard score on the criterion variable equals .71 times the standard score on the predictor variable. The mean square error of prediction is given by

$$1 - r^2$$

or

.50.

Now let us suppose that an error is made in estimating the correlation between predictor and criterion (or that the correlation is based upon some fallible judgment, rather than data), and let us suppose further that the correlation coefficient is misestimated by as much as .3. In such a case, where the prediction is now that the standard score of the criterion variable is equal to r + c times the standard score of the predictor variable, the new mean square error of the prediction is equal to

$$1 - r^2 + c^2$$

which is equal to .60 if c is as much as .3 (i.e., if the correlation of .71 is believed to be 1.01 or .41). So it appears to be that a rather grievous error in estimating a correlation coefficient results in an increase in mean square error of prediction of only 20 percent. Figure 2 presents reduction in square error in prediction as a function of believed correlation coefficient when the true correlation coefficient is .7. The maximum appears rather flat to me.

Insert Figure 2 about here

Consider now the multivariate case, again by example. Suppose that a criterion variable is predicted from four predictors, the first two positive and the second two negative. Suppose, further, that the optimum weighting scheme involves giving twice as much weight to the first and third variables,

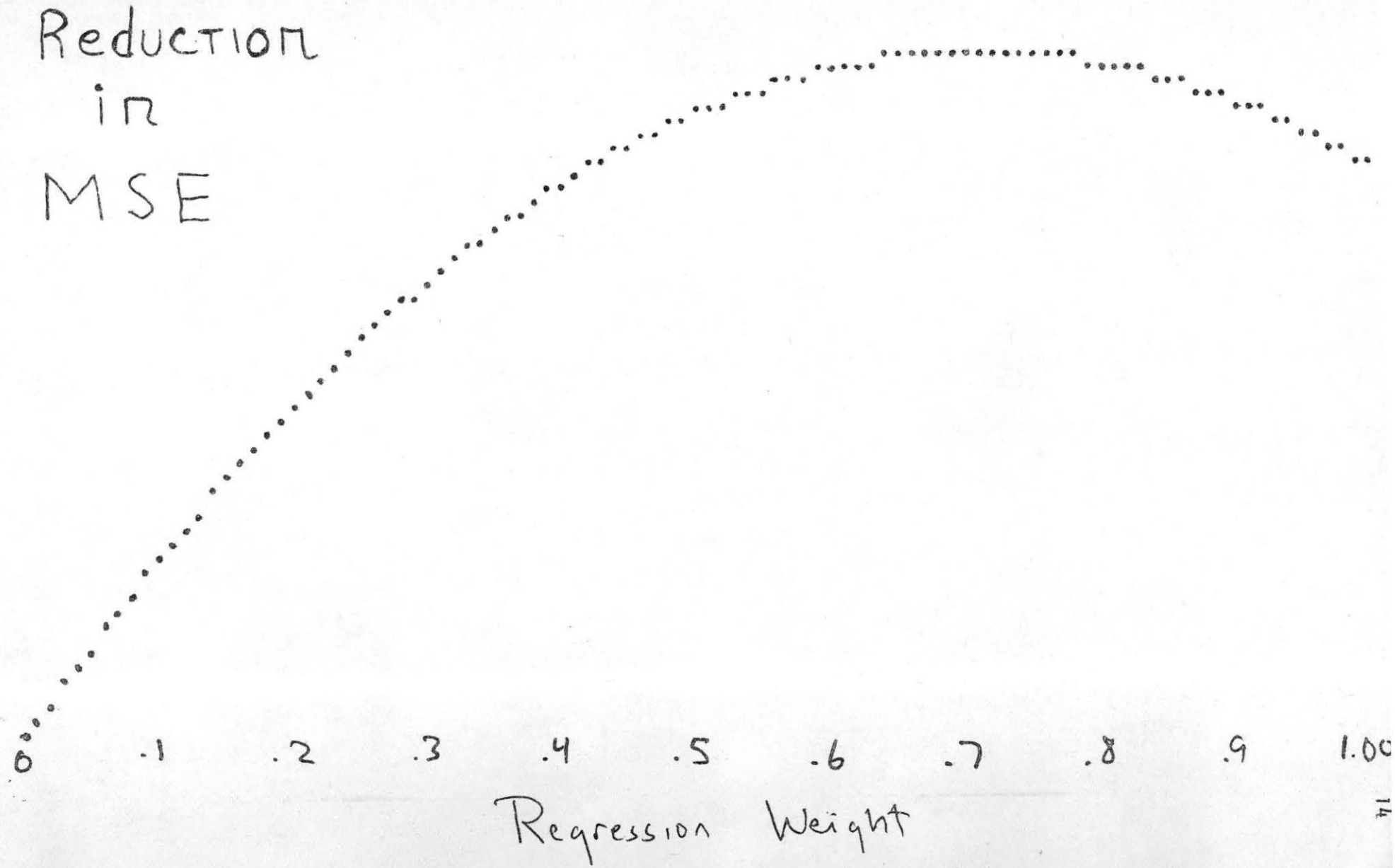

Figure 2.

but that we mistakenly give twice as much weight to the second and fourth variables. In other words, the optimum weighting scheme is:

$$2X_1 + X_2 - 2X_3 - X_4$$

but it is believed to be

$$X_1 + 2X_2 - X_3 - 2X_4.$$

If X_1, X_2, X_3, and X_4 are uncorrelated, then the output of the two models will be correlated .80. Intercorrelations between the positive predictors on the one hand and between the negative predictors on the other will yield an even greater correlation between the output of the two models. Such an intercorrelation is often found in the situations discussed in this paper--e.g., intercorrelations between undergraduate GPA's and GRE scores.

The implications of these conclusions--if they are valid--are clear. First, linear models are extremely robust, hence have great practical value in decision making. Second, the other side of the coin is that they are lousy models of human judgment--because they are too easily fit by the data. Third, the practical implication is that the important problem in the sort of prediction situations discussed in this paper is not that of optimally weighting variables, but rather that of discovering variables to be combined in more or less optimal ways.

The next question that I plan to study is that of when we should stop adding variables in a prediction situation; in other words, how weak a relationship must a predictor have to a criterion, or how redundant must it be with other predictors, before adding it and giving it equal weight actually hurts prediction rather than improves it.

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.

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. A case study of graduate admissions: Application of three principles of human decision making. American Psychologist, 1971, 26, 180-188.

Dawes, R. M. An inequality concerning correlation of composites vs. composites of correlations. Oregon Research Institute Methodological Note, 1(1), 1970.

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.

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.

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.

Lord, F. M. Cutting scores and errors of measurement. Psychometrika, 1967, 27, 19-30.

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

Naylor, J. C. & Wherry, R. J., Sr. 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.

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.

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

Strauss, R. W. The MMPI and types of hospital discharge. Doctoral dissertation, University of Michigan, 1967.

Tucker, L. R. A suggested alternative formulation in the developments by Hirsch, Hammond, & Hirsch, and by Hammond, Hirsch, & Todd. Psychological Review, 1964, 71, 528-530.

von Winterfeldt, D. & Edwards, W. Costs and payoffs in perceptual research. Preprint of article to appear in Handbook of Perception, E. C. Carterette & M. P. Friedman, Eds., 1972.

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

Wherry, R. J., Sr., & Naylor, J. C. Comparison of two approaches--JAN and PROF--for capturing rarer 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.

Footnotes

1. Paper presented at Chicago School of Business, November 17, 1972; at the Seminar on Multiple Criteria Decision Making, Columbia, South Carolina, October 26, 1972; and at the Annual Meeting of the Society of Multivariate Experimental Psychologists, Fort Worth, Texas, November 1972.

2. 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; Dawes, 1970.) Here, an equal weighting scheme gives the same output as the average of all random models--all of which have positive validity; hence, it has a correlation higher than the average correlation of the random models.