Psychological Bulletin
1976, Vol. 83, No. 2, 213–217

Estimating Coefficients in Linear Models: It Don't Make No Nevermind

Howard Wainer
Department of Behavioral Science, University of Chicago

It is proved that under very general circumstances coefficients in multiple regression models can be replaced with equal weights with almost no loss in accuracy on the original data sample. It is then shown that these equal weights will have greater robustness than least squares regression coefficients. The implications for problems of prediction are discussed.

In the two decades since Meehl's (1954) book on the respective accuracy of clinical versus clerical prediction, little practical consequence has been observed. Diagnoses are still made by clinicians, not by clerks; college admissions are still done by committee, not by computer. This is true despite the considerable strength of Meehl's argument that humans are very poor at combining information optimally and that regression models evidently combine information rather well. These points were underlined in some recent work by Dawes and Corrigan (1974), in which they found again that human predictors do poorly when compared with regression models. Strikingly, they found that for some reason, linear models with random regression weights also do better than do humans. Even more striking, when all regression weights were set equal to one another they found still higher correlation with criterion on a validating sample. The obvious question here is Why? Is it because humans are so terrible at combining information that almost any rule works better, or is it some artifact of linear regression?

Support for the latter interpretation is seen in an early paper by Wilks (1938, p. 27, Theorem I), in which he proved, under a set of reasonably general conditions, that with a sufficiently large number of intercorrelated predictor variables, *virtually any combination* of them will yield the same prediction! Or, as Henry Kaiser (1970) once said about a similar problem in factor analysis, "It don't make no nevermind" (p. 403). The conditions for this surprising result to hold are essentially those which anyone attempting to build a prediction model would accede to readily: (a) All predictor variables should be oriented properly (if you don't know what direction the criterion variable lies with respect to a predictor, that predictor shouldn't be used); and (b) the predictor variables should be intercorrelated positively.

In the further exploration of this area, Einhorn and Hogarth (1975) showed that *equal* regression weights would be a reasonable choice in Wilks' weight set and that they have a number of attractive side benefits. They are easy to estimate, and they do not "use up" any degrees of freedom in their estimation. They are insensitive to outliers, and nonnormality in the original sample does not perturb their values in a way that impairs their accuracy on a validating sample. And, perhaps because of these characteristics, equal weights are very robust, often giving higher correlation with criterion on validation than least squares estimates of the regression coefficients and almost never giving drastically inferior results.

Summarizing then, we have the following results:

1. Humans are inferior to linear models in their ability to optimally combine information.

I wish to thank Bert F. Green, Jr., for his careful readings of this paper as well as his thoughtful comments. Of course, he should not be held responsible for the extreme stance taken herein; indeed, he warned against it.

Requests for reprints should be sent to Howard Wainer, Committee on Methodology of Behavioral Research, Green Hall, 5848 South University Avenue, University of Chicago, Chicago, Illinois 60637.

2. Humans are inferior to linear models even when the regression coefficients are chosen in a very crude way (e.g., set them equal).

3. Equally weighted linear models are not very inferior to least squares regression weights and indeed are frequently superior.

The question remains though, Why are equal regression weights so good? The hints obtained from Wilks (1938) and Einhorn and Hogarth (1975) indicate that it is some property of regression and not that humans are so inept that any consistent rule is superior. I think that this notion is correct and can be proved under reasonably general assumptions. To do this I borrow heavily from the recent work of Green (Note 1). Let me start by stating what will become the end of the argument in the form of a theorem.

Equal Weights Theorem

When k linearly independent predictor variables $\mathbf{x}_i$ ($i = 1, \ldots, k$) with zero mean and unit variance are used to predict a variable $\mathbf{y}$, which is also scaled to zero mean and unit variance, and when the population values of the standardized least squares regression coefficients are β_i ($i = 1, \ldots, k$), then the expected loss of variance explained using equal (.5) weights will be less than $k/96$ if all βs are uniformly distributed on the interval [.25, .75].

This expected loss is diminished considerably if the $\mathbf{x}_i$'s are not independent. In this case denote the variance–covariance matrix of the $\mathbf{x}_i$'s by R.

Proof

Before going to the multivariate case described in the above theorem, it is instructive to follow Green (Note 1) and examine the bivariate case. Let us define θ^2 to be the proportion of total variance which is not explained using a regression coefficient a, but which is explained using β. It will be helpful to further define the following:

$$\hat{y} = \beta x \qquad [1]$$

and

$$\bar{y} = ax. \qquad [2]$$

Note that since $y - \bar{y} = y - \hat{y} + \hat{y} - \bar{y}$ and variance of independent variables is additive, we have

$$\sigma_{y-\bar{y}}^2 = \sigma_{y-\hat{y}}^2 + \sigma_{\hat{y}-\bar{y}}^2.$$

The left side of the equation is the amount of variance left unexplained by the use of a. This is equal to the amount originally unexplained plus θ^2, or

$$(1 - r^2) + \theta^2 = \sigma_{y-\bar{y}}^2 + \sigma_{\hat{y}-\bar{y}}^2. \qquad [3]$$

But

$$\sigma_{y-\hat{y}}^2 = 1 - r^2,$$

so that Equation 3 reduces to

$$\theta^2 = \sigma_{\hat{y}-\bar{y}}^2.$$

Note from Equations 1 and 2 that $\hat{y} - \bar{y} = (\beta - a)x$, so that

$$\sigma_{\hat{y}-\bar{y}}^2 = (\beta - a)^2\sigma_x^2. \qquad [4]$$

Let us call the difference between the true regression coefficient β and the value being used the error, or $\gamma = \beta - a$, which implies that Equation 4 becomes

$$\theta^2 = \gamma^2\sigma_x^2.$$

But since $\sigma_x^2 = 1$, this yields the interesting result that

$$\theta^2 = \gamma^2. \qquad [5]$$

To more easily see the implications of this, let us examine the situation when $r = .9$. When the variables are standardized, this implies that the regression weight will also be .9. We have explained 81% of the variance, leaving only 19% unexplained. If we are interested in seeing how much of a deviation from the optimal regression weight of .9 is possible and still reduce our accuracy by no more than 1% of the total variance, we can use Equation 5. It tells us that if $\theta^2 = .01$, then γ^2 is also .01 or that $\gamma = \pm.1$. Or it tells us that $\hat{a}$ can range from .8 to 1 and still lose no more than 1% of the variance. In Green's (Note 1) words, this is an indication of the "flabbiness of regression." If a ranges freely within the interval .8 to 1, the expected loss is given by

$$\text{E (loss)} = 2\int_0^{.1} \gamma^2\, d\gamma$$

$$= 2\,(.1)^3/3 = .002/3 = .00067.$$

With the above demonstration available to clarify intuition, it is now possible to go forward to prove the theorem. The multivariate form of Equation 5 is given by

$$\gamma' R \gamma = \theta^2, \qquad [6]$$

where γ is a k-element vector of differences between the population value of the regression coefficients and the value a being used for them. Note that if $R = I$, as in the theorem, the total effect of these errors can be obtained by merely summing the effect of each of the errors. To follow the conditions of the theorem, we see that the expected loss for any element of a, if all a_i's are set equal to .5 and if all β_i's range uniformly in the interval [.25, .75], is given by

$$E\ (\text{loss}) = 2\int_0^{.25} \gamma^2 d\gamma = \frac{2(1/4)^3}{3} = \frac{1}{96}.$$

Therefore the expected loss for a k-predictor case in which all predictor variables are linearly independent is $k/96$. The theorem is thus proved. It is obvious that if $R \neq I$, then the expected loss will not merely be the sum of the individual losses but will instead be some smaller number. In fact, it will be related to the eigenvalues of R^{-1} in a manner described in detail elsewhere (Green, Note 1).

The implications of this theorem are obvious; even when the correct regression weights are known, the expected loss in accuracy caused by the use of equal weights is very modest indeed. The requirements of this theorem are fairly general in that requiring the weights to be in the range specified is no real restriction, and orienting predictor variables properly is a task of no great difficulty. A predictor whose relation to the criterion variable is unknown with respect to direction shouldn't be used. The same is true for variables whose relative influence is so small that their regression coefficients are very small. If you have a variable whose regression coefficient is greater than .75, the criterion of interest is very predictable and you probably don't need schemes like the one proposed here. Note that the theorem is easily generalized with respect to the allowable interval for the regression weights. The same result obtains for any interval so long as the difference between the largest and smallest regression weight is .5.

Let us go a bit further and ask why equal weighting schemes in regression do so well on validation relative to least squares weights. Once again I start the argument with a theorem, which can be thought of as a corollary to the previous theorem.

Generalizability of Fit Theorem

If the β_i's are not known and are estimated by the fallible values b_i $(i = 1, \ldots, k)$, which are uniformly distributed on the interval [.25, .75], then we can expect that equal weights will do no worse on a validation sample than they did on the original one relative to the b_i's. In fact, we expect that, relative to the b_i's, the performance of the equal weights will improve.

Proof

Before the above theorem can be proved, it is important to review why a shrinkage of accuracy is usually observed when a regression model whose parameters were estimated on one data sample is tried on a neutral or validation sample. This shrinkage is due to two possible factors, either of which can yield shrinkage but which usually occur in combination. First, it is usual to overfit the original data, thus fitting some of the noise. This results in an overestimate of the goodness of fit of the model and is termed *capitalization on chance*. Of course, the excess goodness of fit disappears when the model is tried on a neutral sample. A second factor is the presence of outliers (data points which deviate from multivariate normality) in the original sample. These points typically have an undue influence on the estimates of the parameter values. Once again, these outliers are not usually represented in the same way on the validation data and so a reduction of accuracy occurs.

The use of equal weights avoids both of these problems. First, since equal weights are not estimated with the data, there is little

likelihood of capitalization on chance. Second, the existence of outliers in the original data set has no influence on the estimates and so cannot possibly pull them away from the correct values. Thus, the proof of the generalizability of fit theorem is as follows: Two circumstances can have occurred which will reduce the accuracy of the least squares fit on the validation sample relative to the original sample. Neither of these circumstances can affect the accuracy of equal weights; therefore, we expect that their accuracy will not decrease on the validation sample. If neither of the two things which can reduce the accuracy of least squares weights occurs, then the relationship between the least squares weights and the equal weights which was observed on the original sample will hold on the validation sample. If, however, anything goes wrong, then the equal weight model will improve relative to the least squares weights on validation. Frequently this improvement is substantial. Just how substantial depends on the seriousness of the deviation from theoretical assumptions in the original sample.

We can summarize that if results as good as those demonstrated with the equal weights theorem are possible in the original sample, it is no wonder that on a neutral sample (when the least squares estimates of the regression weights no longer have the benefits of capitalization on chance) equal weights are more robust. The robustness of equal weights is especially striking when the least squares weights are perturbed because of sharp deviations from multivariate normality. Thus, Green's (Note 1) findings complement those of Einhorn and Hogarth (1975).

Note that at no time did I make any mention of sample size. In fact, in the equal weights theorem I assumed that the regression weights that were being ignored were the correct population values. Even in this case equal weights result in little loss. In this conclusion I differ from Einhorn and Hogarth (1975), who do not support equal weights when large samples are available which allow the rejection of the hypothesis of equal weights at some statistically significant level. If the least squares weights are not perfect, equal weights will do very well indeed relative to least squares estimates, but even when estimates of the β_i are perfect, equal weighting schemes do not yield a serious decrement in accuracy.

The conclusion to be drawn is very clear and coincides exactly with that of Dawes and Corrigan (1974). When you are interested solely in prediction, it is a very rare situation that calls for regression weights which are unequal. This is particularly true in the behavioral sciences, in which relative prediction is the most typical kind of problem. The solution is then: (a) orient all predictor variables in the proper fashion, discarding equivocal ones; (b) scale them all into standardized form; and (c) add them up. To avoid devilish repercussions, let me explicate these steps a bit. Step a is equivalent to assigning a weight of $+1$, 0, or -1 to each variable while keeping watch that the intercorrelations of the weighted variables are all positive. A way of doing this would be to calculate least squares regression weights, dropping variables with small weights that have low correlation with criterion. If any large negative weights appear, those variables should be changed in sign and the least squares regression repeated; if negative weights persist, there might be a suppression effect that needs to be examined.

Steps b and c are straightforward. The following of these steps does not, in any way, enter into the truth of the equal weights theorem; it only points a way for one to be sure that the conditions of the theorem are upheld in data. Lest the stance taken in this paper be viewed as revolutionary and extreme, let me quote Green's (Note 2) comment:

> test makers have been using this method ever since the Army Alpha. That is, all items are scored in the positive direction, poor items are discarded, and the test score is simply the sum of the item scores. Most attempts at differential item weighting show relatively little improvement over simple scoring.

The equal weights theorem merely proves what many have believed all along; that is,

that the resulting prediction is apt to be very close to the optimal one, were the optimal weights known, and often better than one which does not use optimal weights. Note also that this sort of scheme works well even when an operational criterion is not available.

An example of a possible use of this method is found in Wainer (1974), in which a linear model using equal weights is used to predict individual voting behavior of U.S. Senators.

REFERENCE NOTES

1. Green, B. F. *Parameter sensitivity in multivariate methods.* Unpublished manuscript, Johns Hopkins University, 1974.
2. Green, B. F. Personal communication, 1975.

REFERENCES

Dawes, R. M., & Corrigan, B. Linear models in decision making. *Psychological Bulletin,* 1974, *81,* 95–106.

Einhorn, H. J., & Hogarth, R. M. Unit weighting schemes for decision making. *Organizational behavior and human performance,* 1975, *13,* 171–192.

Kaiser, H. F. A second generation little jiffy. *Psychometrika,* 1970, *35,* 401–415.

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

Wainer, H. Predicting the outcome of the Senate trial of Richard M. Nixon. *Behavioral Science,* 1974, *19,* 404–406.

Wilks, S. S. Weighting systems for linear functions of correlated variables when there is no dependent variable. *Psychometrika,* 1938, *3,* 23–40.

(Received November 18, 1974)