USING STATISTICAL TECHNIQUES TO FIND PREDICTIVE RELATIONSHIPS BETWEEN VARIABLES

by

John H. Halton

Computer Sciences Technical Report #428

April 1981

USING STATISTICAL TECHNIQUES TO FIND PREDICTIVE RELATIONSHIPS BETWEEN VARIABLES

John H. Halton

ABSTRACT

This paper discusses from a practical standpoint the techniques to be used in finding a predictive model of the regression type, and investigates the various difficulties which may arise. Particular attention is given to models intended to simplify known deterministic procedures for predicting the value taken by a dependent variable y, given a vector x of independent variables.

USING STATISTICAL TECHNIQUES TO FIND PREDICTIVE RELATIONSHIPS BETWEEN VARIABLES

John H. Halton

INTRODUCTION

It is often important to find a relationship between a variable y and a set of variables $\boldsymbol{x} = (x_1, x_2, \ldots, x_m)$, in the form

$$y = f(\boldsymbol{x}) = f(x_1, x_2, \ldots, x_m), \tag{1}$$

from a set of n observed values of these variables. The occurrence of *errors of measurement*, together with imprecisions in the formula (1) and possible effects of *hidden variables* not included in $\boldsymbol{x}$ but affecting y, lead us to think of y (and often of $\boldsymbol{x}$ also) as *random variables* [r.v.], rather than deterministic ones. These r.v., defined as measurable functions on an appropriate probability space, are then denoted by Greek letters, η and ξ; and, instead of (1), we seek a relationship of the form

$$\mathrm{E}[\eta \mid \xi = \boldsymbol{x}] = f(\boldsymbol{x}), \tag{2}$$

where $\mathrm{E}[\eta \mid \xi = \boldsymbol{x}]$ is the *conditional expectation* of η, given that the r.v. ξ takes the value $\boldsymbol{x}$. The statistical technique used to determine an approximation to f from the observed values of η and ξ is called **multivariate regression analysis** and involves the procedures referred to as the *Analysis of Variance* [AV]: the possible functions f are restricted to a family of functions $\varphi(\boldsymbol{x}; \boldsymbol{a})$, where $\boldsymbol{a} = (\alpha_1, \alpha_2, \ldots, \alpha_g)$ is a vector of *parameters*; and this is usually, in turn, limited to *linear families* of

the form

$$\varphi(\boldsymbol{x};\ \boldsymbol{a}) = \sum_{h=1}^{g} \alpha_h\ \phi_h(\boldsymbol{x}). \tag{3}$$

The parameters $\boldsymbol{a}$ are then determined so as to minimize the *sum of squares*

$$S_n(\boldsymbol{a}) = \sum_{j=1}^{n} \left\{ \eta_j - \sum_{h=1}^{g} \alpha_h \phi_h(\xi_j) \right\}^2 W(\xi_j) = (\eta - \Phi^{\mathrm{T}}\boldsymbol{a})^{\mathrm{T}} \boldsymbol{W} (\eta - \Phi^{\mathrm{T}}\boldsymbol{a}), \tag{4}$$

where η is the column vector $(\eta_1, \eta_2, \ldots, \eta_n)$ and $\eta_j = \eta(s_j)$, Ξ is the $(n \times m)$ matrix with rows $\xi_1, \xi_2, \ldots, \xi_n$ and $\xi_j = \xi(s_j)$, $\boldsymbol{W}$ is the $(n \times n)$ diagonal matrix with components $(\boldsymbol{W})_{ij} = \delta_{ij}\ W(\xi_j)$, and $s_1, s_2, \ldots, s_n$ are the n outcomes, in our probability space, of repeated independent sampling. The minimum of $S_n(\boldsymbol{a})$ occurs uniquely when $\boldsymbol{a}$ satisfies the *normal equations*

$$\sum_{k=1}^{g} \left(\sum_{j=1}^{n} \phi_h(\xi_j)\phi_k(\xi_j)W(\xi_j) \right) \alpha_k = \sum_{j=1}^{n} \phi_h(\xi_j)\eta_j W(\xi_j), \tag{5}$$

which may be written

$$\Lambda \boldsymbol{a} = \lambda, \tag{6}$$

where

$$\Lambda = \Phi \boldsymbol{W} \Phi^{\mathrm{T}} \quad \text{and} \quad \lambda = \Phi \boldsymbol{W} \eta, \tag{7}$$

and Φ is the $(g \times n)$ matrix with components $(\Phi)_{hj} = \phi_h(\xi_j)$; uniqueness of solution being guaranteed if and only if [iff] the matrix $\Lambda = \Lambda(\eta, \Xi; \boldsymbol{a})$ is invertible (i.e., non-singular.)

If one further assumes that the *deviations from a true regression*

$$\zeta(\boldsymbol{x}) = \left\{ \eta - \sum_{h=1}^{g} \beta_h \phi_h(\boldsymbol{x}) \right\} \sqrt{W(\boldsymbol{x})} \tag{8}$$

are distributed with a *standard normal distribution* [mean 0, variance 1]

independent of x, with all observed deviations mutually statistically independent, then it follows that

$$E[a_{min}] = \beta \quad \text{and} \quad V[a_{min}] = \Lambda^{-1}, \tag{9}$$

where a_{min} is the solution of (5) - (7) for a given set of n observations and $V[...]$ denotes the variance-covariance matrix. What is more, if we choose to partition the parameter set a into a_1 and a_2, with e and $(g-e)$ components, respectively, and if we form corresponding submatrices Φ_1 and Φ_2, these yield a restricted normal equation

$$\Lambda_1 a_1 = \lambda_1, \text{ with } \Lambda_1 = \Phi_1 W \Phi_1^T \text{ and } \lambda_1 = \Phi_1 W \eta, \tag{10}$$

and a full normal equation

$$\Lambda a = \lambda, \text{ with } \Lambda = \begin{pmatrix} \Phi_1 W \Phi_1^T & \Phi_1 W \Phi_2^T \\ \Phi_2 W \Phi_1^T & \Phi_2 W \Phi_2^T \end{pmatrix} \text{ and } \lambda = \begin{pmatrix} \Phi_1 \\ \Phi_2 \end{pmatrix} W \eta \ . \tag{11}$$

If the respective solutions are $a_{min}^{(e)}$ and $a_{min}^{(g)}$, then, on the null-hypothesis that $\beta_2 = 0$, it follows that $S_n(a_{min}^{(g)})$ —— the residual sum of squares after allowing for the full sample-regression —— and $\Delta_{12} = S_n(a_{min}^{(e)}) - S_n(a_{min}^{(g)})$ —— the extra sum of squares accounted for by the full sample-regression, over the restricted sample-regression —— are *independently* distributed with the *standard chi-squared distribution*, with $(n-g)$ and $(g-e)$ *degrees of freedom* [d.f.], respectively. Thus we may compute the *variance-ratio*

$$F = \frac{S_n(a_{min}^{(g)})/(n - g)}{\Delta_{12}/(g - e)} \tag{12}$$

and test the significance of the contribution of the full sample-regression

over the restricted sample-regression by reference to the known distribution of F. **This is the gist of the application of analysis of variance to the step-by-step development of an effective regression formula.**

For a mathematical derivation of these results, with emphasis on the assumptions and choices required along the way, see the earlier paper on ***"Multivariate Regression Analysis"*** *by the present author.*

The purpose of this further paper is to discuss procedures for the successful development of an effective predictive regression model of the form embodied in (1) - (3), using these techniques.

INITIAL APPROACH TO THE PROBLEM

Suppose that we are given a set of n observations

$$(\eta_j;\ \xi_{1j},\ \xi_{2j},\ \ldots,\ \xi_{mj}) \quad \text{for} \quad j = 1,\ 2,\ \ldots,\ n; \qquad (13)$$

where the η_j are *averages* of many readings of a particular variable υ_j and ξ_j denotes a set of corresponding independent variables [e.g., υ may be the weight of an inventory item specified by the parameters ξ.] If, as seems plausible in many instances, we assume that the individual readings of υ are mutually independent, and that the actual readings are random variables [r.v.] identically distributed with a distribution of a somewhat normal form [e.g., binomial, Poisson, chi-squared], so that the Central Limit Theorem takes effect after a moderate number of repetitions; then we may conclude that the averages η are *normally distri-*

buted with mean $\mu(\boldsymbol{x}) = E[\upsilon | \xi = \boldsymbol{x}]$ *and variance* $\sigma^2(\boldsymbol{x}) = \text{var}[\upsilon | \xi = \boldsymbol{x}]/N$, when N is the number of observations of υ for $\xi = \boldsymbol{x}$, used to arrive at the average value η associated with the independent variables $\boldsymbol{x}$.

It is possible, of course, to attempt to verify that the η_j are normally distributed and to estimate the means $\mu(\xi_j)$ and the variances $\sigma^2(\xi_j)$ by standard statistical techniques (such as the estimation of the moments of the distribution.) As always, we can never finally verify the hypothesis; but only decide not to reject it on the evidence obtained. In many cases, such distributional studies are difficult to carry out, because of lack of detailed data. In the absence of such data, there is no way to estimate $\sigma^2(\boldsymbol{x})$ [which corresponds in our model to the function $1/W(\boldsymbol{x})$] and it is often, without any particular foundation, apart from invincible ignorance (!), assumed to be *constant*.

It will be relevant at several points in our discussion, to observe that *the particular variable* η *obtained is arbitrary*, since we have a choice of recording instead any function $\eta' = \Phi(\eta, \xi)$ we wish [for example, we could record the logarithm or the cube of η.] The validity of the assumption of constant variance is thus extremely precarious. One way to investigate this question would be to plot the (η_j, ξ_{ij}) data for $j = 1, 2, \ldots, n$ and various choices of i, to see if the spread of values of η seems indeed to be uniform over values of ξ_i: if not, it is possible that a preliminary transformation of the variable η may improve the situation.

Turning to the mean $\mu(\boldsymbol{x})$, we observe from (2) that $\mu(\boldsymbol{x}) = f(\boldsymbol{x})$, the regression function which we seek to determine. This leads us to consideration of the family of functions $\varphi(\boldsymbol{x}; \boldsymbol{a})$ from which we are to select an approximation to $f(\boldsymbol{x})$. In practice, we are compelled to limit ourselves to consideration of linear families of functions (3); and here again, it is sometimes possible to improve the effectiveness of such a model by suitably transforming the function η. For instance, if we suspect that the function $f(\boldsymbol{x})$ takes the form [with $m = 3$]

$$f(x_1, x_2, x_3) = \alpha_0 x_1^{\alpha_1} x_2^{\alpha_2} x_3^{\alpha_3}, \tag{14}$$

then we may be able to approximate the regression for $\log\eta$ more closely by a linear regression formula

$$f'(x_1, x_2, x_3) = \alpha_1 \log x_1 + \alpha_2 \log x_2 + \alpha_3 \log x_3 + \alpha_4, \tag{15}$$

than we could ever hope to do for η itself. However, we must note that, in general, $E[\log\eta] \neq \log E[\eta]$; so that a *bias* will probably be introduced into our model: this is least if the variance $\sigma^2(\boldsymbol{x})$ is small.

Note that the requirements of transforming η to linearize the parametric dependence of the regression family, and to make the variance of the deviations from regression constant, may well conflict; and then we must adjust our model accordingly, perhaps by introducing a weight-function $W(\boldsymbol{x})$ after all.

It must be made clear that not all dependences can possibly be linearized [e.g., consider $m = 1$ and $f(x) = \alpha_1 + \alpha_2 x + \alpha_3 x^{\alpha_4} + \alpha_5^{\,x}$, or even so simple a function as $f(x) = \alpha_1^{\,x} + \alpha_2^{\,x}$.] Such cases must be handled by

very specialized techniques, if they are amenable at all. Note, too, that, even if we can linearize, the solution we obtain is "least-squares" only for the transformed variable, not for the original one.

The reader is referred, e.g., to E1, chs. 3 and 5 [the references are to the bibliography of the earlier paper, *Multivariate Regression Analysis* (MRA, henceforth)] for further discussion of transformations of the dependent variable η, analysis of residuals, and linearization of regression families.

THE CHOICE OF THE MODEL

The discussion of linearization of the regression family (or **model**) for our analysis already introduces the general, and fundamental, question of choosing an appropriate family of functions (3). We must now consider this further. First of all, it must be emphasized that a firm grasp of the physical situation must be retained at all times. There is an uncountably infinite number of possible choices for the functions $\phi_h(\boldsymbol{x})$, even after we have limited ourselves to a linear family [the linearity here has to do with the *parameters* α_h, not with the functional dependence on $\boldsymbol{x}$], so that it is simply impossible to cover all possible functional dependences, or even a large proportion of them. Even quite a modest attempt at a "try to cover the waterfront" approach will exhaust the resources of the largest computer; and even if this were not the case, the inclusion of a large num-

ber of unsuitable functions may well lead to spurious fit of the data, which does not then stand up to use as a predictor for further data. It should be remembered that the number g of parameters in our model may not exceed the number n of observations; and, in practice, g should be several times smaller than n, to avoid spurious fits. In this connection, note the remarks on pp. 39 and 40 of **MRA** [under "(v)".] Draper & Smith [**E1**, p. 415] say: "The important point to remember is that the screening of variables should never be left to the sole discretion of any statistical procedure, including the [automatic] multiple regression procedures [of statistical computer packages]" and (p. 419) "In our experience, the 'let's try everything' approach to transformations on original input variables [corresponding to our choices of functions $\phi_h(\boldsymbol{x})$] rarely works well. Unless the underlying mechanics of the problem imply the usefulness of a transformed X-variable, applying one is likely to be a waste of time" and (p. 422) "Finally, no scientist should be persuaded to abandon his scientific insight and principles in favor of some computerized statistical screening procedure. The use of multiple regression techniques is a powerful tool only if it is applied with intelligence and caution."

In some situations, there is no clear understanding of the mechanisms involved, and one is then reduced to the kind of blind casting about for predictor variables which is discouraged by all experienced practitioners.

But if there is some suggestion of the kind of relationship which holds, it is strongly recommended that the first stage of model-building should consist of a careful study of the situation, leading to a listing of all the functions $\phi_h(\boldsymbol{x})$ which are likely to represent factors having an additive linear effect on η. Several questions now arise:

(i) *Are there any variables* (so-called "hidden variables") *in addition to those listed in* $\boldsymbol{x} = (x_1, x_2, \ldots, x_m)$, *which will affect the value of* η? This question is addressed by G. E. P. Box in *Technometrics*, vol. 8, 1966, pp. 625-629 (see also **E1**, p. 295), who points out that systematic changes in hidden variables may cause all sorts of biases and spurious fits to distort the predictive capability of our model. Often, such variables are unsuspected and unmeasured; so that there is no remedy for the lack of success of the model; but sometimes, some careful thought will reveal the source of the problem. One variable which often causes trouble is *time*, through seasonal fluctuations, as well as historical trends and changes (such as the adoption of new manufacturing techniques, or the unpredictable vagaries of public opinion and fashion.)

Another source of erroneous conclusions in the use of "unplanned data", such as are obtained as a by-product of a continuing practical operation, rather than from a carefully designed experiment, is the effect of controls intended to keep the operation practically useful and efficient, which keep certain variables or functions of variables deliberately nearly

constant, thereby veiling their effect on the dependent variable. [This may not matter, if indeed we are only interested in predicting the behavior of η for values of $\boldsymbol{x}$ satisfying these control criteria; but will lead to errors if we stray into other regions of the predictor-variable space.]

(ii) *Will any of the putative predictors* $\phi_h(\boldsymbol{x})$ *have so slight an influence on the values of* η *that they will be better left out of consideration, at lest initially, lest they introduce spurious fits and lead to poor predictive performance of our model?* This is a matter of the physical realities of the problem, which must always remain at the forefront of our considerations. It is not good to proliferate predictors, especially since this may easily lead to dependence or near-dependence between predictors, which is reflected in numerical ill-conditioning in the normal equations (5) - (7) whose solution is essential to the regression analysis.

(iii) *Is the situation such that the value of* η *is affected by the variables* $\boldsymbol{x}$ *in a different way, according to certain qualitative circumstances* [e.g. the cost of a process may depend on which of a set of alternative machines is used; or on the level of priority accorded to it]? If so, it is useful to introduce so-called "dummy variables", usually with possible values 0 or 1, to indicate which alternative choice occurs. Note that it is best to reduce all choices to dichotomies: if there are five ways of doing something, we avoid using a single 5-way variable (with the implication of effect proportional to the variable-value) and include five

{0, 1}-variables instead (noting that it is only possible for one of these to take the value 1, in any given instance.)

(iv) *Are all the influences of the independent variables* $\boldsymbol{x}$ *on the dependent variable* η *truly representable by a linear form?* This question has already been touched on on page 6 above. Another kind of non-linearity arises, for example, if η is the time required for a process, whose specification is dependent on $\boldsymbol{x}$. If the dependence consists of formulae for the time taken by successive sub-processes, all is well; but if certain of the sub-processes may be carried out simultaneously (and, in a mass-production situation, this may be a matter of carrying out one sub-process on one item, while another sub-process is performed on another, identical item), the total time becomes the time required by a so-called "critical path" through the process schedule, while other sub-processes do not add to the time required, and may even be idle for part of the time. Such a situation requires additional preliminary analysis to determine (perhaps with a number of iterations) which of the functions $\phi_h(\boldsymbol{x})$ actually will contribute to the value of η.

(v) *Are there any strong correlations among the dependent and predictor variables?* When we have answered questions (i) - (iv) and if no insurmountable problems have arisen, we must compute the *sample variance-covariance matrix* of the r.v. η, $\phi_1(\xi)$, $\phi_2(\xi)$, ..., $\phi_g(\xi)$, whose row-h/column-k component (taking index 0 for η) is given by

$$V_{hk} = T_{hkn}/(n - 1); \tag{16}$$

where

$$T_{00n} = \sum_{j=1}^{n} \eta_j^2 - \frac{1}{n} \sum_{i=1}^{n} \eta_i \sum_{j=1}^{n} \eta_j, \tag{17}$$

$$T_{0kn} = T_{k0n} = \sum_{j=1}^{n} \eta_j \phi_k(\xi_j) - \frac{1}{n} \sum_{i=1}^{n} \eta_i \sum_{j=1}^{n} \phi_k(\xi_j), \tag{18}$$

$$T_{hkn} = \sum_{j=1}^{n} \phi_h(\xi_j)\phi_k(\xi_j) - \frac{1}{n} \sum_{i=1}^{n} \phi_h(\xi_i) \sum_{j=1}^{n} \phi_k(\xi_j); \tag{19}$$

and these are computed by the recurrence relations [see **MRA**, pp. 7 - 8]

$$S_{00} = 0, \quad S_{0j} = S_{0(j-1)} + \eta_j, \tag{20}$$

$$S_{h0} = 0, \quad S_{hj} = S_{h(j-1)} + \phi_h(\xi_j), \tag{21}$$

$$T_{001} = 0, \quad T_{00j} = T_{00(j-1)} + \frac{j}{j - 1}\left(\eta_j - \frac{1}{j} S_{0j}\right)^2, \tag{22}$$

$$T_{h01} = 0, \quad T_{h0j} = T_{h0(j-1)} + \frac{j}{j - 1}\left(\phi_h(\xi_j) - \frac{1}{j} S_{hj}\right)\left(\eta_j - \frac{1}{j} S_{0j}\right), \tag{23}$$

$$T_{hk1} = 0, \quad T_{hkj} = T_{hk(j-1)} + \frac{j}{j - 1}\left(\phi_h(\xi_j) - \frac{1}{j} S_{hj}\right)\left(\phi_k(\xi_j) - \frac{1}{j} S_{kj}\right). \tag{24}$$

From this, we obtain the *sample correlation matrix* with (h, k)-component

$$R_{hk} = \frac{V_{hk}}{\sqrt{V_{hh}V_{kk}}} = \frac{T_{hkn}}{\sqrt{T_{hhn}T_{kkn}}}. \tag{25}$$

Two considerations now arise. First, if there are large values of R_{hk} [i.e., since, as a consequence of the Cauchy-Schwarz inequality, all correlations satisfy

$$-1 \leqslant R_{hk} \leqslant +1, \tag{26}$$

if $R_{hk} \approx \pm 1$], when h and k are both non-zero, then we need to orthogonalize the functions $\phi_h(x)$ and $\phi_k(x)$; that is we must use a Gram-Schmidt procedure [see, e.g., D2, p. 67, or D3, pp. 314 (§10.2), 488] to reduce the ill-conditioning of the matrix Λ. This amounts to defining a new function

$$\phi'_k(x) = \phi_k(x) - \frac{T_{hkn}}{T_{hhn}} \phi_h(x). \tag{27}$$

Secondly, if $R_{0h} = R_{h0} \approx \pm 1$ [note that $R_{hh} = 1$ exactly, for all h] with h non-zero, then the function $\phi_h(x)$ is *strongly correlated* with η, and therefore is a prime candidate for inclusion in a regression formula.

DEVELOPING AND TESTING THE REGRESSION MODEL

The questions (i) - (v) above having been resolved satisfactorily, we may now proceed with the regression analysis. *If* (as one is strongly tempted to do; through impulses arising from overwork, impatience, laziness, and ignorance ... the desire is almost universal, but *must* be resisted!) *one disregards the preliminaries and puts more-or-less raw data into an automatic statistical regression analysis via computer, the result is generally either spurious or inconclusive.* Even disregard of question (v) alone can lead (a) to numerically inaccurate results due to ill-conditioning, and (b) to *spurious fits* due to the multiplicity of available functions and parameters (i.e., too many degrees of freedom), leading to

a divergence between apparent (i.e., tabulated) and actual levels of significance of statistical criteria [see **MRA**, page 39, at (v).]

Since it is generally impractical to consider *all possible regressions* [for g predictor variables $\phi_h(\boldsymbol{x})$, there would be 2^g possible regression formulae], there are many ways of selecting which will be considered, even after the preliminary pruning has been carried out. One procedure (often used in automatic statistical computer packages) is

+ to seek, first, the single predictor yielding the largest sum of squares due to regression (or equivalently, the smallest residual sum of squares): this is equivalent to selecting the predictor most highly correlated with the dependent variable η. If this yields a statistically significant lessening of the sum of squares, we seek, next, the single predictor in the remaining set which yields the smallest residual sum of squares (after a bivariate regression on the first-selected predictor together with this

+ new one.) This stepwise procedure is now continued, until either all the proposed predictors have been incorporated with sufficient significance, or (more likely) no additional predictor yields significant improvement. [In this matter, there is much room for individual opinion, based (it is to be hoped) on experience and a close monitoring of the process: see, e.g., **MRA**, page 48, **C3**, §27.27, or **E1**, chs. 2 - 4, 6, and 8.]

The outcome of all this effort is a formula

$$E[\eta \mid \xi = \boldsymbol{x}] \approx \sum_{h=1}^{g} \alpha_h \phi_h(\boldsymbol{x}), \tag{28}$$

with a carefully selected and pruned set of predictor functions $\phi_h(\boldsymbol{x})$ and a significant reduction in the sum of squares (i.e., the sum of squares due to this regression formula significantly exceeds that due to the residuals.) A new question now arises: *How good is this formula for predicting the value to be observed for* η, *given the independent variables* $\boldsymbol{x}$?

(a) It is valuable to examine plots of the *residuals*,

$$\eta_j - \hat{\eta}_j = \eta_j - \sum_{h=1}^{g} \alpha_h \phi_h(\xi_j), \tag{29}$$

in themselves, against the values of $\hat{\eta}_j$, and against the predictors $\phi_h(\xi_j)$. It is necessary, of course, that the spread (i.e., more specifically, the sample variance) of the residuals should be approximately constant (or, in the general case, of the order of $1 / W(\xi_j)$.) [See **E1**, chs. 3, 6.]

(b) One should observe the reduction in the sum of squares due to the regression, taking into account the questions of actual (rather than tabulated) significance raised, e.g., in **MRA**, page 39 (at (v).)

(c) If the problem is what we have called *Type I* [see **MRA**, pages 36, 40]; that is, if the values of the independent variables observed are the only ones which will arise; then this suffices. However, if the problem is *Type II*; that is, if the observed values of $\boldsymbol{x}$ are randomly sampled,

from a distribution not yet specified, then yet another question arises:
✝ *How representative is our sample of* x*-values, and how stable is the formula over the whole range of possible* x *for which predictions of* η *are to be made?* First, we observe that the theoretical treatment of this situation, even when we make the (often grossly unwarranted) assumption that the ξ are normally distributed with the deviations from regression,
✝ is very complicated and limited in its utility. To obtain a satisfactory answer to the first part of our latest question, it is necessary to make sure that as large and widespread a selection of possible values of x as can be achieved should be obtained. On the other hand, in order to get
✝ some sense of the stability of the regression (28), one should split the sample into two or more sub-samples and compare the regressions obtained in each. The selection of sub-samples as to both size and distribution is important, but not really fully understood; so, once again, rules of thumb and matters of opinion prevail.

✝ One thing is clear: *If the model is to be useful in a Type II situation, it is necessary that it yield good predictions for* x*-values not included in the sample used to construct the model. Thus it is essential that this predictive capability be tested before a model, however internally consistent it may be, is accepted.* Failure of the model to perform satisfactorily in such a test is a clear indication that it is inadequate, and probably that any internal consistency observed is due to spurious fit, caused by an overabundance of degrees of freedom in the regression analysis.

ON MODELS INTENDED TO SIMPLIFY KNOWN DETERMINISTIC PROCEDURES

If we are dealing with a general scientific problem, it is a serious question, whether a regression model of any kind, involving only the given variables x can possibly yield acceptable predictions, even for some of the observed x-values: after all, there may be hidden variables at work, of which we have no knowledge at all; or the variability of the deviations from regression, even if there is a respectable regression relation, may be so great as to minimize the predictive value of the regression formula.

✢ In some situations, we have a *known deterministic procedure* (or possibly a statistical one, but of low variability and high complication) for accurately predicting η, given x. This enormously simplifies our task, which reduces to finding a *simple formula*, to do the job that we know that we *can* do, with considerable labor. Since we have precise information on a procedure which will give the required answers, there are *no* hidden variables and *no* unknown relationships. It is simply a matter of looking at the known formula and seeking to approximate it with a simpler one.

✢ It may, indeed, be possible to arrive at a simplified formulation of the predictive procedure in purely deterministic terms, by the use of *approximations* with absolute *ranges of variation* of the values involved.

Even if it seems, after all, that a statistical approach is more ✢ natural or straightforward; first, a close examination of the longer procedure should indicate whether a linearized or linear regression can

possibly be expected to yield sufficiently precise results. *Of course,*
✝ *by a well-tried process, we may find that the use of linear regressions over restricted ranges of values of* x *may work, when an overall model would not.* This is worth exploring. But if there is no such way and the answer is in the negative; then one may terminate the attempt at once; or, at least, resort to a search (necessarily with little hope of success) for special *ad hoc* methods.

If it appears that a linear regression should describe the situation adequately, then the analysis of the data should be relatively straight-
✝ forward, since the functional relationships established in the known procedure are those which should appear in the initial set of predictor functions from which a final set for analysis will be selected.

If the predictions are necessary in any case, and the problem is just one of reducing cost (in terms of time and/or effort), *then it seems sure that a simpler predictive model must be available: the only question is how much simpler it will be and what will be the resulting economy.*