CONIFEROUS FOREST BIOME

Ecosystem Analysis Studies

U.S./International Biological Program

Internal report

INTERNAL REPORT 1

SENSITIVITY ANALYSIS AS "PROPAGATION OF ERROR" AND MODEL VALIDATION

MODELING NOTEBOOK

Scott Overton
Oregon State University

Some applications of a model, or ways in which a model may be "exercised," fall into a general category that is commonly called "sensitivity analysis." Considerable confusion surrounds the subject, at least some of which arises because the purpose, and hence the execution and inferential import, of sensitivity analyses may be quite varied. One objective of this note is to delineate the major kinds of sensitivity analyses and to differentiate their meaning. The second objective is to phrase the subject in conventional statistical terms so that it may be recognized as an extension of conventional statistics.

Section I. Propagation of error.

For the present purpose, all of the essential structural features of the model may be subsumed into the expression,

$$y = h(X), \tag{1}$$

where y is the output variable (possibly vector valued) and X is the input variable (probably vector valued) and where h is defined either explicitly or implicitly. The elements of X may be state variables or parameters or environmental variables or driving variables; their definition is one of the ways that sensitivity analysis may vary in purpose, as we shall elaborate later. However, we shall see that procedurally it makes no difference and that, in fact, we can establish most of the important ideas while ignoring the definition of X and while treating y and X as univariate.

The basis of sensitivity analysis is the question: What difference does it make if X is perturbed by an increment ε? That is, we consider

$$y + \delta = h(X + \varepsilon) \tag{2}$$

and specifically are interested in the manner in which δ changes with respect to X, y, and ε. This is the central idea, but the way the relation is expressed is highly variable. We might look at δ/ε or $\frac{dy}{dx}$ or $\frac{\delta/y}{\varepsilon/X}$ or at any number of possible expressions of the "sensitivity" of y to a change in X. The important variants are incremental (as for

example δ/ε), differential ($\frac{dy}{dX} = \lim_{\varepsilon \to o} \frac{\delta}{\varepsilon}$), and statistical ($E(\delta^2)/E(\varepsilon^2)$).

Now if the function h is algebraically explicit, it may be possible to perform a sensitivity analysis analytically. If h exists only in the form of a complex non-algebraic algorithim, for example as a computer simulation program, then it will be necessary to use numerical methods. In either case, execution requires expertise from conventional mathematical analysis or statistical analysis, or both, but numerical evaluation may require additional expertise, particularly computer technology.

An incremental analysis requires only the calculation of $y = h(X)$ and $y + \delta = h(X + \varepsilon)$, if h is linear. If h is non-linear, this pair is repeated over a range of X, so as to study the behavior of δ over X. A numerical analysis of $\frac{dy}{dX}$ would follow essentially this same form, and again one would study the behavior of $\frac{dy}{dX}$ over some region of X. The technical aspects of this problem are simple; one has only to calculate y for a particular value of X and describe the relationship.

The problem of too many variables (which we will treat later) turns out to be important to both incremental and to differential analysis, but the extension can readily be made. As for the differential form, it is argued that this is useful primarily in the case of analytic functions and its utility is drastically reduced when the differential cannot be expressed in closed form. Further, both the incremental and differential forms are hardly distinguishable from a general study of model behavior, which should be treated as a separate topic.

General formulation of the statistical sensitivity question

Consider Equation 2 and let ε be randomly selected from some statistical population. For example, let $\varepsilon \sim N(0, 1)$. Then we ask, not what is the incremental perturbation of y for a particular perturbation of X, but rather what is the statistical distribution of δ given the specified statistical distribution of ε. This is a central question of statistical analysis and it is useful to cite some of the conventional results. For simplicity, we shall use y in place of $(y + \delta)$ and X in place of $(X + \varepsilon)$, so that now we implicitly define $\delta = y - E(y)$ and $\varepsilon = X - E(X)$.

Let the distribution function of X be F(X) and the computer model algorithm be $y = h(X)$. We can visualize the problem in terms of Figure 1.

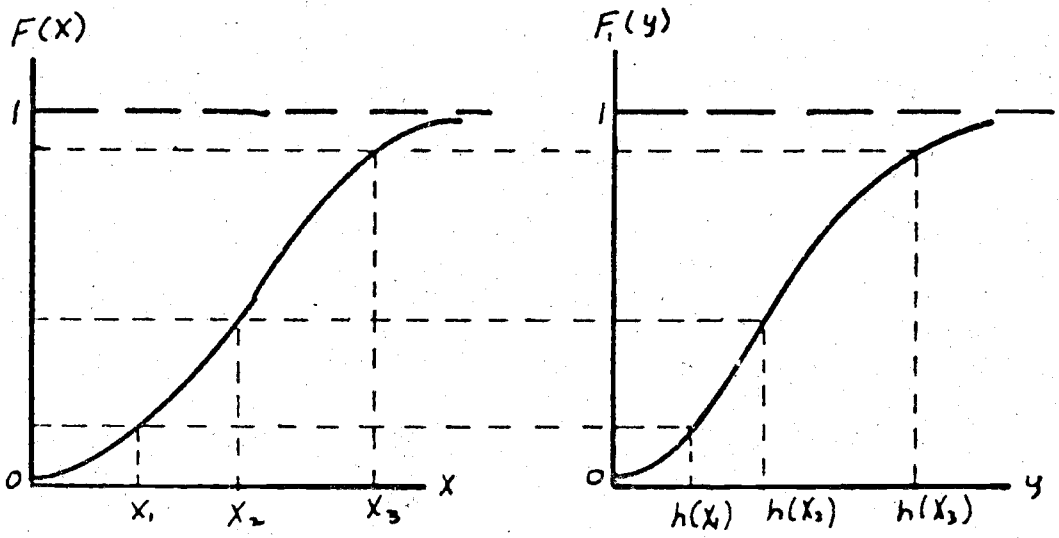

Figure 1. Distribution functions of X and $y = h(X)$ when h is monotone increasing.

If h is monotone increasing, write $F_1(h(X)) = F(X)$, and if h is monotone decreasing, $F_1(h(X)) = 1 - F(X)$. However, if h is not monotonic, the problem is more complex, as is illustrated in Figure 2, and for

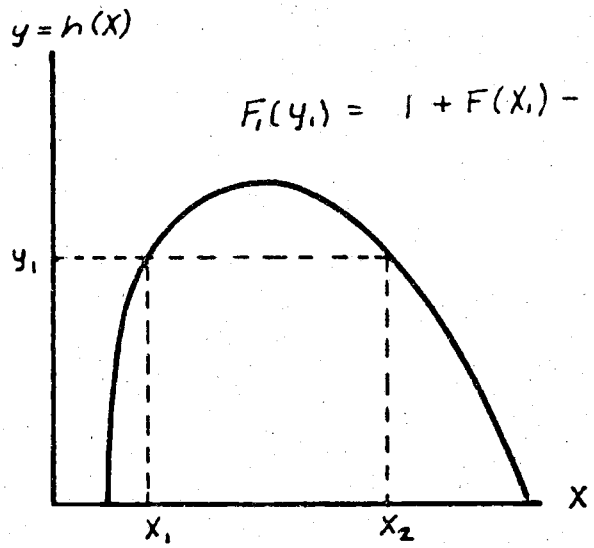

Figure 2. If y is a unimodal function of X, then the probability that Y is equal to or less than y is the sum of the probabilities 1) X is equal to or less than X_1 and 2) X is greater than X_2, where X_1 and X_2 are the solutions to the equation $y_1 = h(X)$.

arbitrary h, the region of integration may be very fragmented. It is apparent from this elementary consideration that construction of a good algorithm for direct numerical evaluation of F(y) under unconstrained h will be very difficult, and the problem is obviously one for Monte Carlo or indirect numerical techniques. Before elaboration of that, however, the problem of multiple X's should be considered. Specifically, lett $y = h(X_1, X_2)$ with (X_1, X_2) distributed according to $F(X_1,X_2) = F_1(X_1)F_2(X_2|X_1)$. This can be visualized best in terms of two contour maps in X_1 and X_2, one describing the y surface and the other describing the surface of $f(X_1,X_2) = \frac{d^2F}{dX_1dX_2}$.

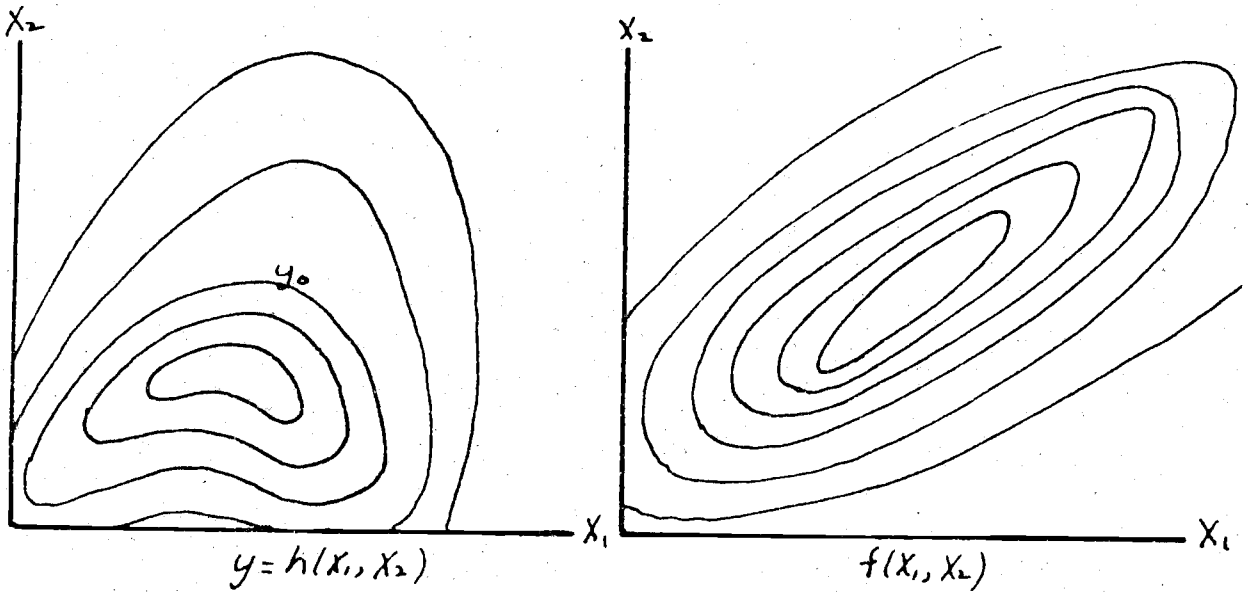

Figure 3. Contour (isopleth) representation of the functions $h(X_1,X_2)$ and $f(X_1,X_2)$.

Then, if y_0 is the value of y at a particular contour in the map of h, then determination of $F(y_0)$ involves the integration of $f(X_1,X_2)$ over the region defined by that contour. Again, this obviously will be very difficult to express as a general algorithm. Because we really do not need this much information, but rather need only variance, we will not pursue the specific distributional question any further.

Rather, the question considered is: If X is distributed as F(X), with variance V(X), what is V(y)? This question can be answered much more quickly and simply than can the more general distribution question, and Monte Carlo techniques

are quite suitable for complex cases. Before elaborating that, however, let us review the basic results from statistical analysis.

If $y = h(X)$ is a linear function of X, then the general case can be written

$$y = \underline{a}'\underline{X} \quad , \quad V(\underline{X}) = \Sigma \quad ,$$

where $\underline{X}$ is a column vector of input variables, $\underline{a}$ is a column vector of constants and $\Sigma = E[(\underline{X} - E(\underline{X}))(\underline{X} - E(\underline{X}))]$,

where Σ is called the variance-covariance matrix of $\underline{X}$.

Then $E(y) = \underline{a}'E(\underline{X})$

and $V(y) = \underline{a}'\Sigma\underline{a}$.

The nonlinear explicit function case also is treated in conventional statistical literature (c.f. Deming 1943, 1964). By Taylor's expansion, if $h(\underline{X})$ is well behaved, y can be expressed,

$$y = h(E(\underline{X})) + \sum_{i=1}^{k} (X_i - E(X_i)) \frac{\partial h}{\partial X_i}\bigg|_{E(\underline{X})} + \cdots\cdot$$

$$= h(E(\underline{X})) + \sum_{i=1}^{k} a_i \, (X_i - E(X_i)) + \cdots\cdot$$

and to the first order of approximation,

$$V(y) = E[(y - h(E(\underline{X}))^2] = \underline{a}'\Sigma\underline{a} \quad ,$$

where again $\underline{a}$ is the column vector of the a_i, $i = 1, \cdots, k$ and Σ is, as defined before, the variance-covariance matrix of the X_i.

The technique of approximating the variance in this manner is called "propagation of error" or the "delta technique", depending on the literature referenced. The first term would seem to apply to most sensitivity analyses of the statistical form, as they involve propagation of error of the output variable, given prescribed distributions of the input variables.

The general problem of concern here is the non-linear non-explicit function defined by a complex computer algorithm which generates the output y from the vector of input, $\underline{X}$, but the Taylor's expansion method requires an explicit well-behaved function. The Monte Carlo strategy is simple, and can be applied to the non-explicit function. It can be described as

follows. The variance of y is estimated by taking a sample of independent random values of y, where a random value of y is obtained by generating y from a random value of $\underline{X}$ selected from the joint distribution function modeled for the X's. That is, specify $F(X_1, X_2, \cdots, X_k)$

$= F_1(X_1)F_2(X_2|X_1)\cdots F_k(X_k|X_1, \cdots, X_{k-1})$ and select $\langle X_1, X_2, \cdots, X_k \rangle$ sequentially from this distribution using one of the several methods of selecting a random variable from a specified distribution. Then calculate $y = h(\underline{X})$ by the algorithm, repeat the entire process until n values of y are obtained, and estimate, $E(y) \hat{=} \bar{y}$ and $V(y) \hat{=} s_y^2 = \frac{1}{n-1}\{\sum y^2 - \frac{(\sum y)^2}{n}\}$. It follows from general theory that $\bar{y}$ and s_y^2 are unbiased for the mean and variance of the random variable Y.

If the algorithm for computing $y = h(\underline{X})$ is very complex, it may be costly to simulate a single value and prohibitive to repeat the simulation a large number of times. We are interested in obtaining the best estimate of V(y) with the fewest replications of the simulation process. Therefore, let us consider a numerical solution as an alternative to the Monte Carlo solution. For illustration, restrict X to the univariate case and consider the process of selecting a sample of X's. A single X is selected by first generating a random variable $u \sim U(0,1)$ and generating $X = F^{-1}(u)$, as illustrated in Figure 4.

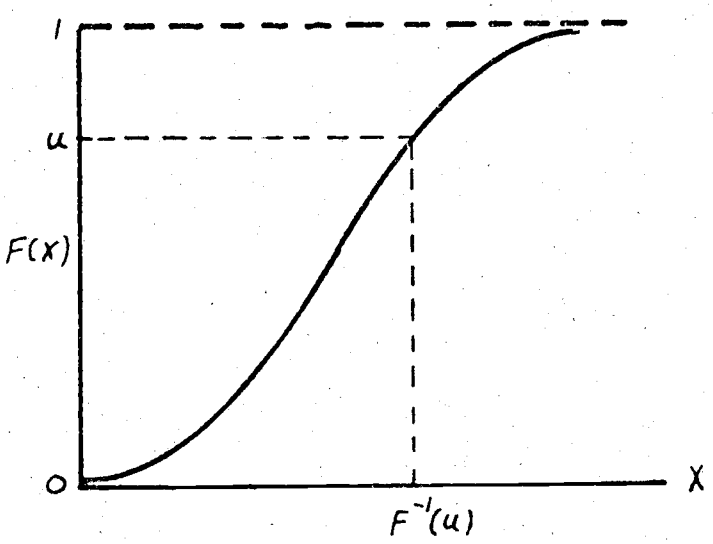

Figure 4. The "probability transformation," such that X is a function u and $u \sim U(0, 1)$.

If we select n independent values of u and repeat the translation process, then the "sample" of X's is independent and from the distribution F(X). However, we can sample "systematically" in the interval (0,1) and obtain good estimates of E(X) and σ_x^2 (and hence E(y) and σ_y^2) and in fact the systematic sample will be "better" in a very meaningful way. For example, if one wishes to estimate the mean from one observation, it is clearly best to set u = 1/2 than to allow u to vary randomly over the entire range (0,1). Similarly, a systematic sample of, say, size 10, arranged symmetrically about u = 1/2, will provide a better estimate of the mean and of the variance than will a random sample of the same size.

Some other questions might be raised here. For example, it is not clear that the conventional sample variance formula is best to use with the systematic sample, but to my knowledge this has not been investigated in the present context. The sample size question is also of interest--how big a sample is needed and what circumstances affect the decision?

Generalization to many variables

Suppose that we have chosen to use a systematic representation of the random variables $X_1, \ldots, X_k$ and that we maintain the general structure $F(X_1,\ldots,X_k) = F_1(X_1)\ F_2(X_2|X_1)\ldots\ F_k(X_k|X_1,\ldots,X_{k-1})$. Now, for each of the n points in X_1 chosen according to $F_1(X_1)$, we will select n points in X_2 according to $F_2(X_2|X_1)$, and for each pair of (X_1,X_2) values so chosen we will select n points in X_3 according to $F_3(X_3|X_1,X_2)$, etc. If $X_1, X_2,\ldots, X_k$ are independent, so that

$$F(X_1,\ldots,X_k) = F_1(X_1)\ F_2(X_2)\cdots\ F_k(X_k) \quad ,$$

then the selection process is simplified but the total number of simulations, n^k, is unchanged.

The problem of too many variables

The preceding result, that n^k points $\langle X_1\ldots, X_k\rangle$ are needed to adequately represent the distribution of $X_1,\ldots, X_k$, where n values of X_i are deemed necessary to represent the distribution of X_i, leads to the recognition of an essential problem of too many variables. If n = 10 (not unreasonable if one is concerned about non-linear

behavior) and if k = 10 (not large at all in terms of ecosystem models) then 10^{10} simulations are necessary to "adequately" characterize the distribution of y. This is obviously prohibitive, even if a simulation costs dollars, rather than tens of dollars. We might (some day) be willing to spend $10,000 for a critical sensitivity analysis, or perhaps even $100,000, but it is difficult at this stage to even imagine such a circumstance. Just for focus, let's establish 2^{10} as the greatest number of simulations that an analysis can require, and specify 2^6 as a practical limit. Few of our problems will qualify.

The pure Monte Carlo approach--to select n independent values $<X_1, \ldots, X_k>$--can be improved by specifying that these be a subset of the, say, 10^k, values required for the better numerical analysis. However, the gain is slight, and the value of consideration of this possibility lies in recognition that the Monte Carlo approach is approximated by a random reduction of these 10^k sets to some lower figure, say perhaps $10^{f/R}$. By random we emphasize that in pure Monte Carlo we exercise no pattern control over which are selected. This must be considered an unsatisfactory state of affairs. Clearly, it will be necessary to use Monte Carlo methods at some point in the analysis, but let's do so only as a last resort, and let's attempt to find ways to reduce the magnitude of the problem.[1]

Section II. Some specific ecosystem questions requiring propagation of error techniques.

In the beginning of this note, I promised to discuss the different ways in which sensitivity analysis arises in study of ecosystems and the validation of ecosystem models. Again, the emphasis will be on the statistical interpretation--all of the questions will be related to questions of conventional statistical analysis and the methods will be shown to be either central to or direct extensions of standard methodology.

Questions that require some kind of "sensitivity" analysis are classified below. In the discussion that follows, we shall see that

[1]This will be the subject of another modeling note. Several aspects of this problem are apparent. Erratic model behavior is rightly the problem of study of model behavior, and a good strategy should virtually eliminate this concern from a study of propagation of error.

for the purpose of propagation of error, they all resolve into question 1, except for question 6, and question 6 is identified as another topic, the study of model behavior.

1. What is the precision of a prediction?
2. What precision is needed in the estimation of a particular parameter or other "input" variable?
3. Is the model valid (under some criterion)?
4. How critical is some particular model feature or structure, as compared to some alternative feature or structure?
5. Which of two model forms is best?
6. What is the pattern or response over a (broad) range of a particular input variable? That is, what is the behavior of the system?

In an earlier section, we identified two primary types of sensitivity analysis. The first (A) may be characterized by a plot of dy/dx over a range of X and the second (B) as an evaluation of $V(y|\Sigma)$ where Σ is the variance-covariance matrix of the input variables. In the above list, questions 1-5 may be answered by type B and 3-6 by type A. The following discussion will deal only with type B, the propagation of error questions. Question 6, the only one not specifically addressed by Type B sensitivity studies, is properly a subject of discussion in its own right, and will be treated in another note.

Validity testing (question 3)

The term "validity testing" is tossed about casually, but a question regarding its meaning is more apt than not to bring some vague, imprecise, hand-waving answer such as, "Well, everyone knows it is necessary to validate the models before we accept them." Again, note that model testing is a central problem of statistics, and the following treatment is from that point of view. This is an essential aspect of model validation, if not its entirety.

First, it is useful to recognize two aspects of testing, absolute and relative. We may ask the validation question of a single model, and may decide that this model is not valid (i.e., not good enough) without any implication that there exists another model that is better. Conversely, we can test to see which of two or more possible models

is the better, without any implication that either is good enough. By this point, questions 3, 4, and 5 are all validation questions.

A second distinction that must be made is with regard to validation of model form (structure and specification of mathematical relationships) and validation of the values of the mathematical parameters that have been chosen. To illustrate, consider the distinction between testing the numerical value of a hypothetical regression coefficient in a linear model and testing between a linear model and a non-linear model with the same number of parameters. The standard procedures for testing between two forms of linear models are based on the identification of the parametric family to which all linear models belong, and on the error structure. Some non-linear models belong to recognizable parametric families (as for example, Turner's single process law), but in general, comparisons among models cannot be put into the parameter estimation framework.

There are other ways in which testing a model in a new circumstance differs from the usual questions of testing, but these are best brought out by a detailed description of the process. Suppose the model $h(X,Z,\Theta)$ has been constructed in study area A and is being validated (tested) in area B. It is necessary to elaborate the identification of model components in order to address the variety of validation questions:

$y = h(X, Z, \Theta)$,

X is the vector of state variables,

Z is the vector of driving variables,

Θ is the vector of parameters, and

h designates the form of the model.

At one level of validation, we may ask whether the model can be transported in its entirety, including estimated Θ's, from area A to area B. At another level, we may ask whether h can be extended from A to B if we go to the trouble of estimating new values of Θ. Lastly, if we have structured h hierarchically, it is possible to ask whether certain modules or hierarchical levels of h may be retained if others are constructed specifically for area B.

In yet another dimension, we may recognize that validation may be with regard to any of a number of selected outputs (y) and further that a decomposable model will allow validation to be made on submodels

so that a very broad pattern of validation activity is available to us.

Any single such evaluation, however, will necessarily be of the following form. Let $\hat{y}$ be the predicted value of the variable y and $\overset{*}{y}$ be the measured value against which $\hat{y}$ is contrasted, where

$$\hat{y} = h(\hat{X}, \hat{Z}, \hat{\Theta});$$

the hats denoting that all input variables are estimated by some means or another. The validation questions, then, can be simplified into the question: Is the discrepancy between $\hat{Y}$ and $\overset{*}{y}$ greater than can be accounted for in terms of $V(\hat{y})$ and $V(\overset{*}{y})$: That is, if

$$|\hat{y} - \overset{*}{y}| \leq z + (V(\hat{y}) + V(\overset{*}{y}))^{\frac{1}{2}}, \text{ validate}$$

$$> z + (V(\hat{y}) + V(\overset{*}{y}))^{\frac{1}{2}}, \text{ invalidate,}$$

where z is the normal deviate.

This will be recognized as a simple statistical test of the prediction $\hat{y}$ against the estimated value $\overset{*}{y}$ and invalidation is equated to rejection of the null hypothesis. In this context, it is obvious that a model cannot be validated in the sense of being proven true. It can be validated only in the sense of being shown to be adequate, where adequate is specified in choosing the value z of the normal deviate.

At this point, we can discontinue our discussion of validation, as such, because we have now recognized that it is, in the final analysis, a form of prediction testing. Conversely, we can always interpret prediction testing as model validation and the essential unanswered question is with regard to appropriate specification of $V(\hat{y})$, which is question 1.

The sample size question

Question 2, relating to the necessary precision of an estimated parameter, or state variable or driving variable, is a generalization of the classical sample size problem and can be viewed to advantage in that structure. The problem requires two elements for solution: a function V relating the variance of $\hat{y}$ to the effort, c, expended

in determination (estimation) of X and a constraint either on V or on c. In the simplest form, write

$$V(\hat{y}) = g(c).$$

Thus, if c is specified, one solves directly for the precision that will be attained and if V is prescribed, one solves for the required c,

$$c = g^{-1}(V).$$

In the present context, it is desirable to reorient this a bit because (1) g is nonexplicit and (2) it is easier to conceptualize the problem in terms of variance of input variables. This partitions out the problem of achieving a prescribed precision for input variables, which is a sampling problem. Therefore, we consider the specification

$$V(y) = g[V(X)],$$

a formulation that has been addressed in the section on precision of prediction. The inverse problem, of determining the necessary V(X) for a prescribed V(y),

$$V(X) = g^{-1}[V(y)],$$

can conceptually be solved in two ways:

1) Calculate V(y) for each member of a selected set of V(X), using numerical techniques to determine the optimum V(X). [Costs can easily be entered into the problem at this point by constraining the selection of V(X)].
2) Construct an <u>inverse</u> program[2] that sends y into X and simulate V(X) according to V(y). [The basic variance of prediction problem]. Some additional problems pose themselves here:
   a. The answer is not generally unique and some criterion must be established for restriction of the solution.
   b. If more than one y is involved, then the answers are not necessarily consistent and some compromise strategy must be reached.

---

[2] I am not aware of any attempts to accomplish this, but the idea is basic and probably has been tried. This clearly can be done for some models, but the limitations are not obvious on casual contemplation.

Thus, question 2 also resolves into a strategy involving application of question 1. Again, as was discovered for question 3, the strategy is definable in terms of simple (say linear) problems, and the only way in which model complexity is involved is through unusual difficulty in determination of $V(\hat{y})$. A similar position is reached for questions 4 and 5, both of which involve some form of testing, and are resolvable into questions of precision.

Conclusion

The first part of this note deals with problems of propagation of error and the latter part with questions commonly identified as "model validation" questions. It is concluded that all validation questions can be resolved into two parts: (1) strategic aspects, involving logical and inferential structures and (2) tactical aspects, involving some "propagation of error" technique.

It is also concluded that the questions commonly lumped under "sensitivity analysis" can be divided into two distinct categories. One group involves propagation of error, and the other a description of model behavior. Although these are not unrelated, they are sufficiently different operationally that the distinction seems desirable.

The problem of <u>too many variables</u> is discussed and it is recognized that a strategy to resolve this problem must be worked out. It is anticipated that this problem will be even more critical in the study of model behavior and that the note on that subject must deal with the problem.