Statistical Science
1990, Vol. 5, No. 2, 169–174

Role of Models in Statistical Analysis

D. R. Cox

Abstract. A number of distinct roles are identified for probability models used in the analysis of data. Examples are outlined. Some general issues arising in the formulation of such models are discussed.

Key words and phrases: Substantive model, mechanistic model, empirical model, AIDS, rain, calibration, model choice.

1. INTRODUCTION

Many accounts of the theory of statistics start from three premises: (i) that observations on response variables correspond to random variables; (ii) that there is given a family of possible probability distributions for these random variables, the "true" distribution being an unknown member of that family; and (iii) that the objective of the analysis is connected with some aspect of the unknown "true" distribution.

These are decidedly nontrivial abstractions, very fruitful indeed for some purposes of exposition and development, yet in a sense fairly remote from the reality of some applications. In particular, choice of an appropriate family of distributions may be the most challenging phase of analysis.

Possibly partly as a reaction to the overformalization involved in many accounts of statistical analysis, we have seen recently increased emphasis on methods of analysis in which probability considerations play no explicit role. An implied theme of the present paper is that, important and interesting though informal methods are, the confinement of probability models to so-called confirmatory analysis and the separation of "exploratory data analysis" from "statistics" are counter-productive. Indeed the exploration of ground common to the two viewpoints is particularly fruitful. What graphical methods are appropriate for association with a particular formal technique? What probabilistic considerations go reasonably with a particular technique suggested initially on nonprobabilistic grounds?

The primary object of the paper is to review a number of rather different roles for probability models. Any thinking about a complex phenomenon is likely to involve elements of simplification and idealization and hence to rely on a model of some sort, often a qualitative one. To discuss the formation and use of such models would take us too far afield, and in this paper the word model is restricted to probability models.

Sir David R. Cox is Warden of Nuffield College, Oxford OX1 1NF, England.

Probability is used in two distinct, although interrelated, ways in statistics, phenomenologically to describe haphazard variability arising in the real world and epistemologically to represent uncertainty of knowledge. Discussions of statistical inference focus on the latter, whereas here the emphasis is firmly on the former.

We shall distinguish three broad roles for models, each with some variants. These roles can respectively be described concisely as substantive, empirical and indirect. While they may occur in combination in a specific application, it may clarify thinking to bear in mind the rather different emphases involved.

2. SUBSTANTIVE MODELS

In many ways, the most appealing models are those that connect directly with subject-matter considerations. There are several subdivisions of such models.

2.1 Directly Substantive

These models aim to explain what is observed in terms of processes (mechanisms), usually via quantities that are not directly observed and some theoretical notions as to how the system under study "works." A rather simple example concerns the one-hit or two-hit theories of binary response in which at "dose" x the probability of a positive response (success) is either $1 - e^{-\rho x}$ or $1 - (\rho x + 1)e^{-\rho x}$. The background theory is that there is a Poisson process of unknown rate ρ and a period, x, of exposure such that a success is observed if and only if there is at least one point realized in the Poisson process in the first form, and at least two points in the second version. Such models have been suggested in connection with studies of perception and of cancer incidence.

A less simple example concerns stochastic models for, say, hourly rainfall in which (Rodriguez-Iturbe, Cox and Isham, 1988) there is a Poisson cluster process of rain cells, each cell being a random burst of rain of constant but random depth and random duration, the total rainfall intensity at one time being the sum of contributions from cells active at that time. This leads in the simplest formulation to a five-parameter scheme. The parameters represent storm arrival rate,

mean number of cells per storm, mean interval between cells within a storm, mean cell duration and mean cell depth. Such models, while highly idealized, offer a way of summarizing a long time series of complex structure via a relatively small number of parameters of physical significance. Note that the notion of cells is widely used in physical descriptive studies of precipitation.

A third example is that empirical fitting of numbers of cases in an epidemic (e.g., AIDS) may well most fruitfully be done via a representation derived from the transmission models of epidemic theory, even though the process of infection, incubation, etc. is not directly observable.

An example of a substantive model rather closer to some of the more empirical models to be discussed in the next section is provided by the diagonal models used particularly in some sociological contexts (Sobel, 1981, 1985). Consider a two-way square arrangement in which the labelings of rows and columns are essentially the same, so that the diagonal cells have particular meaning. If μ_{ij} is the expected response in cell (i, j), it may be reasonable on subject matter grounds to suppose that

$$\mu_{ij} = \rho\mu_{ii} + (1 - \rho)\mu_{jj};$$

the notion is that "individuals" in cells (i, i) and (j, j) represent "pure" i and j individuals, respectively, whereas individuals in cells (i, j) have moved from i to j and represent some intermediate category. In more complicated versions, ρ depends on further features such as additional explanatory variables.

2.2 Substantive Hypothesis of Dependence and Independence

A second rather weaker probability is that, while there is no detailed knowledge of processes or mechanisms underlying the generation of the data, there are hypotheses, arising from a subject-matter base, about dependencies. Typically these might take the form that, given certain intermediate responses and explanatory variables, some other variables are conditionally independent. These are the research hypotheses of Wermuth and Lauritzen (1989). A fairly typical example is as follows: for a particular population of individuals, given age, weight and gender, blood pressure is conditionally independent of certain measures of personality characteristics. Another weaker possibility is that the hypothesis might specify the direction of some conditional dependency. Probability models are a valuable if not absolutely essential tool for formulating and testing such hypotheses. Note that to some extent the formulation in terms of hypotheses is just a convention. One could just say that subject-matter considerations lead to interest in the nature and direction of any conditional dependency of the kind mentioned.

2.3 Retrospective Discovery of Substantive Issues

In the previous discussion, it is assumed that the subject-matter considerations on which the model is based are available independently of the data under analysis. Sometimes, however, the argument may go in the other direction. If unexplained regularities are detected by more empirical methods, especially if such recur in independent sets of data, it may be worth aiming for a tentative explanation via an underlying process. For instance, suppose that a number of binary response relations are linearized by the complementary log log transformation. That is, if θ denotes the probability of success $\log(-\log\theta)$ is a linear function of an explanatory variable, x. This relation can be generated via a distribution of latent tolerances having an extreme-value form. This suggests at least the possibility that a phenomenon of the general extreme value type underlies the systems under study.

Again, suppose that time series of a particular kind repeatedly show irreversibility, for example slow rises and rapid falls. Such phenomena are not generated by stationary Gaussian processes or point-wise transformations thereof and are evidence either of nonlinearity or of a linear system forced by non-Gaussian innovations. Then it is worth considering whether one of the stochastic models that generate such data is useful either as a basis for data reduction or as an elucidation of an underlying mechanism or process. Similar remarks apply to long-range dependence (asymptotic self-similarity) (see, for example, Cox, 1984). In these, the lag h correlation between the means of long sections h "steps" apart takes a fixed nonzero form regardless of the length of the section. These are essentially fractal processes in which the emphasis is on very long-term (low-frequency) variations rather than very short-term (high-frequency) variations.

3. EMPIRICAL MODELS

The more common type of model in many fields of application is not based on any very specific subject-matter considerations but rather aims to represent in idealized form dependencies, often "smooth" dependencies, thought likely to be present. The parameters in the model capture aspects of what would be observed "in the long run" in hypothetical repetitions and hence to represent features of the system under study free of the accidents of the particular data under analysis.

We shall call such models empirical models: many of the standard models of analysis of variance and of multiple regression and its generalizations are of this

type. It is at the same time the power and the limitation of such models that they are very widely applicable and are not based on highly specific subject-matter considerations.

There are at least two somewhat different roles for such models.

3.1 Estimation of Effects and Their Precision

This is probably by far the most common use of models. The models lead to methods for estimating unknown parameters of interest by confidence intervals or some roughly equivalent inferential procedure. The more "direct" the method of analysis the better, in the sense that dependence on the precise details of the model is thereby lessened. For example, if error of a contrast is estimated by comparison of estimates from independent replicates of some investigation, i.e., in analysis of variance jargon, via a contrast × replicate mean square, this has a more direct appeal than an estimate of precision derived say by explicit appeal to some parametric distributional form. There is, of course, a large literature on these themes. It is important to distinguish the parts of the model that define the aspects of subject-matter interest, the primary aspects, and the secondary aspects that indicate efficient methods of estimation and assessment of precision. The critical secondary assumptions for the study of precision are frequently ones of independence.

Quite often the full specification of a substantive model will call for empirical formulation of some aspects.

3.2 Correction of Deficiencies in Data

While there is a close connection between this and the previous use, it is worth distinguishing situations in which probability models are used particularly to correct for deficiencies in the data; the assumptions made are often quite critical. Examples are correction for attenuation in regression, when explanatory variables are measured with error, the imputation of missing values in data of relatively complex structure and the need to adjust for unusual methods of sampling. Not only do adjustments have to be made but, unless these are minor, the resulting degradation in precision needs to be measured.

There is, of course, an extensive literature on imputation in regression and survey analysis (Little and Rubin, 1988; Rubin, 1988). A rather special example concerns allowance for reporting lags in collecting data on a series of point events. Thus if such events occur in a Poisson process of rate $\rho_D(t)$ and each point is subject to a reporting lag with cumulative distribution function $F_X(x)$, the recorded process of diagnoses based on data available at that time t_0 shows a dip in rate near t_0 because of missing entries. For discussion of statistical problems associated with this system, see Cox and Medley (1989). For corrections involving unusual sampling methods in industrial contexts, in particular length-biased sampling and recurrence-time sampling, see Cox (1968). In a much deeper and broader sense, the whole field of stereology is concerned with this issue.

4. QUASI-DETERMINISTIC MODELS OF RANDOMIZATION THEORY

Somewhat intermediate between the models of Sections 2 and 3 lie the models of unit-treatment additivity used in the randomization theory of experimental design. In these, the assumption is made that the observation obtained on a particular experimental unit depends only on the treatment applied to that unit and is the sum of a constant characteristic of the unit and a constant characteristic of the treatment. The randomization full null hypothesis is that on any unit the observation is the same regardless of the treatment. These are deterministic statements. Probability enters only in the randomization of treatments to experimental units and allows probability assessment of the uncertainty in causal statements of the effect of treatments. The word causal is used because, under the assumptions indicated, the inference compares the response on a unit receiving a particular treatment with what would have been observed had a different treatment been used, all other aspects being unchanged. Of course whether this conforms to causal in a direct physical sense is another matter.

Such hypotheses of unit-treatment additivity are typically empirical in the sense that they are not based on a specific subject-matter "theory" of an underlying process, except perhaps insofar as the choice of a particular function of the response variable is concerned. On the other hand, they are intended as a basis for conclusions rather deeper than the purely descriptive, and considerably stronger than in comparable observational studies.

5. INDIRECT MODELS

Finally we give two situations in which probability models are used rather indirectly either to suggest methods of analysis which can then be assessed via success in some specific application, as judged by a direct practical verification of, for example, predictability or are used to study the properties of particular techniques of analysis.

5.1 Calibration of Methods of Analysis

Some fairly widely used methods of analysis have been suggested from largely or entirely nonprobabilistic viewpoints. Examples are nonmetrical scaling, various forms of cluster analysis, automatic interaction

detection and classification and regression tree analysis. It can throw useful light on such methods to examine their behavior on data generated by known probability models and used without there being any suggestion that such models are appropriate for any particular set of data. A simple example is the finding by Day (1969) that a simple clustering method applied to homogeneous samples from a ten-dimensional multivariate normal distribution usually identified a number of distinct clusters.

In this situation, data of "known" structure are used to study "unknown" methods of analysis rather than "known" methods being used to study data of "unknown" structure. A similar distinction applies to the use of much-analyzed sets of historical data to illustrate new techniques.

5.2 Development of Automatic Data Reduction

Here the emphasis is on the use of a model to suggest a method of data reduction that can then in some sense be tested directly. Image analysis provides a number of examples. The hidden Markov models used in speech technology (Juang, 1985; Jack and Laver, 1988) suggest methods of analysis for, say, recognition purposes where success can be assessed directly. In fact, dynamic time warping, i.e., nonlinear data-based stretchings of the time scale, may be based on optimizing an empirical criterion or on a model in which the time-varying spectrum shifts between a number of forms on the basis of an unobserved discrete-state Markov process. While in a sense such a model is substantive, its success is to be judged via some explicit practical criterion.

A rather different aspect of the calibrative view arises if, say, the difference between the means of two groups of observations is calculated together with a corresponding standard error and normal-theory p-value. Then, without any very specific proposal of a probability model, one may regard the p-value as ranking possible differences in the light of the distribution of differences that would be generated if the data were randomly sampled from a common normal distribution. This is a very weak but not vacuous justification of standard error and p-values.

6. DISCUSSION

We have thus distinguished three broad roles for probabilistic models, the substantive, the empirical and the indirect, with no suggestion that the categories are rigidly defined. In particular, quite often parts of the model, e.g., those representing systematic variation, are based on substantive considerations with other parts much more empirical. The distinctions between the different types are of most importance when formulating models and in checking and modifying such models.

The iterative nature of the passage between data and model will not be discussed explicitly, although its importance is beyond question. In the following, we address briefly, almost in note form, just some of the more detailed issues of model formulation and parameterization.

Meaningful models. For empirical and indirect purposes, it may be enough that a model defines the joint distribution of the random variables concerned, but for substantive purposes it is usually desirable that the model can be used fairly directly to simulate data. The essential idea is that if the investigator cannot use the model directly to simulate artificial data, how can "Nature" have used anything like that method to generate real data? Thus simplified versions of simultaneous models sometimes considered in econometrics, such as

$$\beta_{11} Y_1 + \beta_{12} Y_2 + \alpha_1 = U_1,$$

$$\beta_{21} Y_1 + \beta_{22} Y_2 + \alpha_2 = U_2,$$

where (U_1, U_2) are independently standard normal random variables, certainly define the joint bivariate normal distribution of (Y_1, Y_2) but cannot be used for direct simulation if $\beta_{12}, \beta_{21} \neq 0$. Similar remarks apply to $\{Y_{ij}\}$ defined on a lattice via conditional statements such as that given $\{Y_{i,j\pm1}, Y_{i\pm1,j}\}$, Y_{ij} is normal with mean $\alpha + \beta(Y_{i,j-1} + Y_{i,j+1} + Y_{i-1,j} + Y_{i+1,j})$ and variance σ^2. Of course both models can be recast in a different form, the latter by regarding it as a cross-section of a spatial-temporal process to be simulated, for example by the Gibbs sampler.

Testability of models. Other aspects concern the extent to which models are defined in a narrowly operational sense, i.e., whether the data gathering operation can be repeated many times under virtually the same conditions in reality, in principle hypothetically, or not at all, and the extent to which models contain latent features that are not directly testable; the randomization theory model of Section 4 is a rather extreme form of the latter. Substantive models are very likely to contain latent features, but models for which absolutely no check is available from the data are to be viewed with particular caution. Note, for example, that the assumptions of unit-treatment additivity in randomization theory can be tested as soon as the experimental units are distinguished in some rational fashion, e.g., via the value of a concomitant variable.

Because all models are idealizations, it makes sense to test only features that have a direct or indirect bearing on the conclusions to be drawn.

Empirical models that do not fit will normally be replaced by models that do fit, but for substantive

models in whose basic soundness there is considerable confidence, it may be more fruitful to specify qualitatively the nature of any failure and the broad interpretation to be given to the departures. Indeed one main use of simple models may be to discover the point at which they break down.

Choice of models in light of data. Particularly with empirical models, it is often wise to amend the model in the light of the data under analysis. A distinction should be drawn between doing this so as to change (a) only secondary features of the model in a way not affecting the definition of the parameters of primary interest; (b) the quantitative but not the qualitative aspects of primary importance; and (c) the whole focus of primary concern.

Thus under (b) one might decide that it is better to regress log y on log x than y on x, whereas under (c) the whole focus of interest may change. While (c) can in some ways prove the most rewarding possibility, obviously the dangers of overinterpretation are considerable and conventional measures of uncertainty have to be adjusted if feasible and in any case regarded as giving lower bounds to the uncertainty involved.

Minimal modeling. Especially in empirical modeling, it may be wise to model explicitly only those aspects of the data of direct concern. Thus in so-called repeated measures designs it is often unnecessary to model explicitly the variation "within" individuals, but for the assessment of precision to rely on the variation "between" individuals in suitable summary statistics (see, for instance, Yates, 1982). Of course this applies only when the structure of the "within" individual variation is not of detailed intrinsic interest.

Models in an exploratory context. In both substantive and empirical contexts, the conclusions may be so clearcut that no consideration of random variation, or at most very perfunctory consideration, may be adequate. Indeed, this is the case in many contexts in physics, where a traditional approach to random variation is to improve experimental technique to the point where the random variation is of minor importance. There are, however, an increasing number of parts even of mainstream physics where this is not practicable. A second type of application where probabilistic considerations are sometimes claimed to be unnecessary is in exploratory analysis of data. In many situations, however, the classifications (confirmatory, exploratory) and (probabilistic, descriptive) are quite separate. Most applications are in any case somewhat in between the confirmatory–exploratory extremes and some notion, however approximate, of precision seems highly desirable in the exploratory portions of the analysis, if extremes of overinterpretation are to be avoided. The attachment of standard errors, etc. to the main features of an exploratory analysis, e.g., an exploratory multiple regression, seems often enlightening as indicating a minimum uncertainty.

Objectives of interpretation. Two counterarguments to an emphasis on parameters are sometimes put forward. One is that parameters are not of intrinsic interest, but serve only to index possible probability distributions. Another viewpoint is that the objective of inference should be the prediction of future observations and that parameters are at most a tool toward that end. On the first point, it does seem essential in attempting to unravel relatively complicated problems that individually meaningful features of the system should be identified and interpreted, and parameters seem a key tool in doing that. Especially in empirical models, it is desirable that parameters (e.g., contrasts, regression coefficients and the like) have an interpretation largely independent of the secondary features of the models used.

In particular, the precise choice of the most relevant partial regression coefficients needs considerable care, especially in a time series context.

As to prediction, it is surely true that this has been underemphasized, but, at least for most of the applications with which I am familiar, analysis and understanding are the qualitative objectives, and these seem best expressed via carefully chosen parameters. These can often be reinterpreted via the properties of a large number of hypothetical future observations, and in this sense given a predictive character.

Randomness, determinism and chaos. There is sometimes the question in formulating a model, especially an empirical model, as to whether particular patterns of variation should be represented by systematic (nonrandom) effects or by random variables. In some contexts it does not matter; for example in normal theory balanced randomized block designs, it is unimportant whether or not block effects are regarded as random. In other contexts, effects of some direct interest should be represented as random variables only as a "last resort"; for example, an interaction between treatment effects and intrinsic factors of interest (e.g., "centres") should be taken as random only if they cannot be "explained" in some way. When there are a large number of parameters of secondary interest representing similar effects in an unbalanced design, it will often be good to consider representing them by random variables. This is partly because the occurrence of a large number of nuisance parameters means that unmodified maximum likelihood methods may be inappropriate and partly because higher precision may be achieved by a representation in terms of random variables with a well-behaved distribution. Recovery of between-block information in unbalanced designs is an example.

Finally, at a deeper level, there is the possibility that superficially random variation can be relatively

simply explained by a nonlinear chaotic process. Some very challenging statistical problems arise in distinguishing empirically between a chaotic and a stochastic system and in estimating the dimensionality and structure of a chaotic system from data (for an introduction, see, for example, Berge, Pomeau and Vidal, 1984, pages 150–160):

7. CONCLUSION

Successful use of statistical methods depends crucially on problem formulation. Where probability models are used, the choice of a family of possible models is thus a key step. Distinctions between different kinds of models have, of course, been discussed in the past. I hope that the rather more detailed classification set out in this paper will be a help in clarifying what to do in particular applications, although large elements of subject-matter judgement and technical statistical expertise are usually essential. Indeed, it is precisely the need for this combination that makes our subject such an interesting and demanding one.

ACKNOWLEDGMENT

I am extremely grateful to Nanny Wermuth for searching comments on earlier versions of this paper. A modified version of the paper formed the R. A. Fisher Lecture, Joint Annual Meetings, Washington, August 1989. Some of the material also formed part of a lecture in honor of C. R. Rao, Neuchâtel, August, 1989.

REFERENCES

BERGE, P., POMEAU, Y. and VIDAL, C. (1984). *Order within Chaos.* Wiley, New York.

COX, D. R. (1968). Some sampling problems in technology. In *New Developments in Survey Sampling* (N. L. Johnson and H. Smith, eds.) 506–527. Wiley, New York.

COX, D. R. (1984). Long-range dependence: A review. In *Statistics, An Appraisal* (H. A. David and H. T. David, eds.) 55–75. Iowa State Univ. Press, Ames, Ia.

COX, D. R. and MEDLEY, G. F. (1989). A process of events with notification delay and the forecasting of AIDS. *Philos. Trans. Roy. Soc. London Ser. B* **325** 135–145.

DAY, N. E. (1969). Estimating the components of a mixture of normal distributions. *Biometrika* **56** 463–474.

JACK, M. A. and LAVER, J. (1988). *Aspects of Speech Technology.* Edinburgh Univ. Press, Edinburgh.

JUANG, B. H. (1985). Maximum likelihood estimation for mixture multivariate stochastic observations of Markov chains. *AT&T Technical J.* **64** 1235–1249.

LITTLE, R. J. A. and RUBIN, D. B. (1988). *Statistical Analysis with Missing Data.* Wiley, New York.

RODRIGUEZ-ITURBE, I., COX, D. R. and ISHAM, V. (1988). Some models for rainfall based on stochastic point processes. *Proc. Roy. Soc. London Ser. A* **410** 269–288.

RUBIN, D. B. (1988). *Multiple Imputation for Nonresponse in Surveys.* Wiley, New York.

SOBEL, M. E. (1981). Diagonal mobility models: A substantively motivated class of designs for the analysis of mobility effects. *Amer. Sociol. Rev.* **46** 893–906.

SOBEL, M. E. (1985). Social mobility and fertility revisited: Some new models for the analysis of the mobility effects hypothesis. *Amer. Sociol. Rev.* **50** 699–712.

WERMUTH, N. and LAURITZEN, S. L. (1989). On substantive research hypotheses, conditional independence graphs and graphical chain models (with discussion). *J. Roy. Statist. Soc. Ser. B.* To appear.

YATES, F. (1982). Regression models for repeated measurements. *Biometrics* **38** 850–853.