MUST MODELS MYSTIFY?

A first look at the epistemological status of statistical models

John Bibby

1. Introduction The basic argument of this paper is that the construction of a statistical model inevitably assumes a particular construction of reality. Therefore in as much as this construction is socially and ideologically determined, statistical models must also contain an ideological component. However since this ideological component cannot be explicitly incorporated into the model, we have the basis of a central mystifying contradiction. Hence by purporting to be what they are not - pictures of the real world rather than portraits of the artist - statistical models inevitably confuse and mystify. This conclusion negates the assumptions of the use/abuse paradigm adopted by some radical statisticians.

In order to validate the above proposition we first prepare the gound by asking "What is a model"? Then we describe in more detail our understanding of the term 'statistical model'. Finally we return to describe the meaning of mystification, and consider the question posed in the title of the paper.

2. What is a model? Despite an enormous literature seeking to unravel the distinction between models and such close relations as theories, maps, metaphors and analogies (and paradigms and fairy-tales?) no general consensus on terminology has yet been found. Of course the word 'model' means various things according to contexts. A separate paper, parallel to this, will attempt to clarify these differences ("The great modelling muddle", in preparation.) That paper will distinguish amongst others, between physical models, conceptual models, mathematical models Marks I and II and statistical models. While different in many respects, these models do have one thing in common - they are all designed to extract for the investigator what seems to him to be an essence of reality, in a manner which adds to knowledge, understanding or insight. This might facetiously be presented as the 'model' of a model. Note the problematic insertion above of the idea of "essence of reality". This is to be understood as the result of a value-laden constructive activity and reflects the viewpoint and presuppositions of the investigator. Hence 'essence' means essential for the observer, and not essential for reality.

3. Statistical models In dangerously Althusserian fashion we shall view statistical models as a process which transforms a raw material (input) into a product (output). The input to the modelling process has three parts. Input I is a set of observations or data. Input II is a specification which usually takes the form of one or more mathematical equations. The specification states assumptions made concerning the relationship between the various theoretical concepts whose empirical counterparts are observed in the data. Thus Input II is a statement of assumptions concerning Input I. Input III on the otherhand represents the relationship or isomorphism between the theoretical concepts and their empirical counterparts. Thus Input III links Inputs I and II. The use here of the words like 'input' and 'data' is not meant to imply that these raw materials are unproblematic. On the contrary, neither data nor specification are value-free. They both reflect the ideological pre-suppositions of the analyst and through him of his social environment - non data sed capta (not given but captured).

Having considered the input we turn now to consider the output of the transformation process. This may take many different forms. It could be a set of point or interval estimates of unknown parameters. It could be the result of a hypothesis test - one star, two stars, or three stars, depending on the level of significance. (Not for nothing have the pages of statistical journals been likened to those of the Good Food Guide!) However, in order to paint statistical model-building in as flattering a fashion as possible, we shall use what seems to us to be the most thorough and least mystifying form of statistical model. (Nevertheless, as we shall see, it mystifies.) This is one whose output decomposes each element of observed data into a "fitted value", being that which is predicted by the model, and a "residual" which is simply "observed minus fitted". In other words observed value = fitted value + residual. There will be one equation similar to this for each element of the data. Thus if $\underset{\sim}{d}$ is a vector representing the data, then $\underset{\sim}{d} = \hat{\underset{\sim}{d}} + \hat{\underset{\sim}{u}}$, where $\hat{\underset{\sim}{d}}$ is the vector of fitted values and $\hat{\underset{\sim}{u}}$ is the vector of residuals.

The left hand side of this equation is Input I, while the right hand side is included in the output of the production process. (The output may also include such things as parameter estimates and hypothesis tests, but these seem to be less stringent requirements than those demanded by the above equation.)

So far we have discussed the input and output of the transformation process, but have said little of the process itself. This we shall now do, before passing onto consider its epistemological implications. The transformation process known as statistical modelling may be viewed as having the following stages.

1. One assumes that the elements of Input I (data $\underset{\sim}{d}$) are observed realizations of random variables. For instance in the simple linear regression model we may have data $\underset{\sim}{d} = (x_1,\ldots, x_n, y_1,\ldots,y_n)'$. Stage I of the transformation process then assumes that this $\underset{\sim}{d}$ is an observed realization of a random vector $\underset{\sim}{D} = (X_1,\ldots,X_n, Y_1,\ldots,Y_n)'$.

2. One assumes that the relationship between the random variables in $\underset{\sim}{D}$ accords with Input II (the specification equation). Thus in the simple linear regression model we have

$$E[Y_i] = bE[X_i] \quad \text{or} \quad E[Y_i] = a + bE[X_i].$$

In each of these examples the specification includes n equations, together perhaps with further assertions concerning variances, covariances, normality, and a statement concerning whether the design variables $X_1,\ldots,X_n$ are random or fixed (the latter being viewed as a degenerate special case of the former). The specification may take various forms. In the above examples the specification takes the form $F(\underset{\sim}{D},\underset{\sim}{\theta}) = \underset{\sim}{0}$ where F is a mathematical function, and $\underset{\sim}{\theta}$ is a vector of unknown parameters. However it could be completely nonparametric, and it need not involve an equation. Consider for example the specification $E[Y_i] > E[X_i]$.

3. Any unknown parameters are estimated. This stage will be omitted if the specification is completely nonparametric. It may appear to be omitted e.g. in hypothesis testing, but is usually there beneath the surface (e.g. likelihood ratio hypothesis testing implies the use of maximum likelihood parameter estimates). This stage is the first one where questions of statistical expertise could possibly become relevant, although even here they should not dominate. Whatever estimation procedure is used, let us call $\hat{\underset{\sim}{\theta}}$ the estimated values of the unknown parameters.

4. The 'fitted values' of the input are calculated. If the specification takes the form $F(\underset{\sim}{D},\underset{\sim}{0}) = \underset{\sim}{0}$ then the fitted values $\hat{\underset{\sim}{d}}$ will usually satisfy $F(\hat{\underset{\sim}{d}},\hat{\underset{\sim}{\theta}}) = \underset{\sim}{0}$, although this need not necessarily be the case e.g. the method of monotone regression.

5. The residuals are calculated. The residuals are defined by $\hat{\underset{\sim}{u}} = \underset{\sim}{d} - \hat{\underset{\sim}{d}}$, and the residuals together with the fitted values are part of the product of the transformation process.

4. What is meant by mystification? It is commonly argued that a major weakness of mathematical models is their tendency to oversimplify the complexity of natural events (e.g. J. Gani, Model-building in Probability and Statistics, in T. Shanin, ed., The Rules of the Game). This would perhaps not be a weakness if simplification led to greater ease of understanding. However this is by no means the case. Firstly of course, the mathematical model is expressed in a tersely recondite language, inaccessible to the vast bulk of humanity. More importantly however, the mathematical model cannot abstract the historical conflict implicit in any situation. Just as poetry is lost in translation, dialectical reality tends to be mislaid and obscured by the process of mathematical formalisation. At the same time, the veils of obscurity tend to reflect the hegemonic ideology.

This process of obscurantisation is what we mean by "mystification". Of course this is not peculiar to mathematical models. It pervades the whole of culture, art as well as science.

John Berger has shown how mystifying conventional artistic criteria can be if applied to Hals' painting of The Regents of the Old Men's Alms House in Haarlem (Ways of Seeing, pp.11-16). A conventional art critic concentrates on topics such as the "human condition", "harmonious fusion", "personal vision", and "life's vital forces". Yet, Berger argues, all this merely evades the central historical fact i.e. Hals' masterpiece was painted by a destitute old painter who was forced to live off public charity, and the painter's subjects personified the affluence which necessitated that charity. Hence the mystification was achieved by evading conflict, rendering a-historical, and "explaining away what might otherwise be evident".

A similar situation exists in model-building. Instead of aesthetic criteria we have the antiseptic conventionalities of mathematical formalism. These are ideological in the sense that they tend to obscure the real condition of society, and thereby stabilise it.

Critical path analysis is another pertinent example - it has aptly been called the science which tells you to put on your socks before your shoes, rather than vice versa. This description capures the essence of mathematical mystification, which uses complicated language to state (and confuse) the obvious, thus making things appear much more difficult than they really are.

As Berger points out, the important question to ask is "who benefits from this mystification". The answer is not difficult to see. For model-building

(a) necessitates obscure language, which is a luxury available only to those who are able to obtain initiation into the knowledge elite; (b) it thereby validates the privileged position of the expert, and (c) disenfranchies the lay man. Finally (d) it abstracts from the socio-historical setting. This and other mystificatory functions have been discussed in the context of the general linear model in Bibby (1977) ("The general linear model: a cautionary tale", to appear in C. Payne and C. O'Muircheataigh, The Analysis of Survey Data).