J. R. Statist. Soc. A (1995)
158, *Part* 3, *pp.* 419–466

Model Uncertainty, Data Mining and Statistical Inference

By CHRIS CHATFIELD†

University of Bath, UK

[*Read before* The Royal Statistical Society *on Wednesday, January 18th, 1995, the President,* Professor D. J. Bartholomew, *in the Chair*]

SUMMARY

This paper takes a broad, pragmatic view of statistical inference to include all aspects of *model formulation*. The estimation of model parameters traditionally assumes that a model has a *prespecified known form* and takes no account of possible uncertainty regarding the model structure. This implicitly assumes the existence of a 'true' model, which many would regard as a fiction. In practice *model uncertainty* is a fact of life and likely to be more serious than other sources of uncertainty which have received far more attention from statisticians. This is true whether the model is specified on subject-matter grounds or, as is increasingly the case, when a model is formulated, fitted and checked on the *same* data set in an iterative, interactive way. Modern computing power allows a large number of models to be considered and data-dependent specification searches have become the norm in many areas of statistics. The term *data mining* may be used in this context when the analyst goes to great lengths to obtain a good fit. This paper reviews the effects of model uncertainty, such as too narrow prediction intervals, and the non-trivial biases in parameter estimates which can follow data-based modelling. Ways of assessing and overcoming the effects of model uncertainty are discussed, including the use of simulation and resampling methods, a Bayesian model averaging approach and collecting additional data wherever possible. Perhaps the main aim of the paper is to ensure that statisticians are aware of the problems and start addressing the issues even if there is no simple, general theoretical fix.

Keywords: AUTOREGRESSIVE MODEL; BAYESIAN MODEL AVERAGING; DATA MINING; FORECASTING; MODEL BUILDING; RESAMPLING; STATISTICAL INFERENCE; SUBSET SELECTION

1. INTRODUCTION

It is hard to set universally acceptable limits on the scope of statistical inference. Much traditional theory (e.g. Silvey (1970) and Cox and Hinkley (1974)) is concerned with the following interesting, but narrow, problem. A family of parameter-indexed probability models, P, is postulated. The analyst then examines whether a given single sample of data is consistent with P, and, if so, estimates and/or tests hypotheses about the parameter(s) of P. The members of P usually differ only in the parameter values, and the *structure* of P is assumed known. Silvey admits that 'the setting up of an appropriate probability model . . . calls for considerable experience and judgement' but makes 'no attempt to discuss this aspect of the subject'.

Most statisticians would agree that their work covers a wider ambit than the above, and modern inference is concerned with *model selection* and *model criticism* as well as estimation and hypothesis testing. Some statisticians would widen inference further to include *prediction*, but for the purposes of this paper there is no need to set

†*Address for correspondence*: School of Mathematical Sciences, University of Bath, Bath, BA2 7AY, UK. E-mail: cc@maths.bath.ac.uk

0035–9238/95/158419

exact limits in this regard. However, I do wish to widen statistical inference to include the *whole model building process* which has four main components, namely

(a) model formulation (or model specification),
(b) model fitting (or model estimation),
(c) model checking (or model validation) and
(d) the combination of data from multiple sources (e.g. meta-analysis).

The broad view of statistical inference taken above is consistent with what Chambers (1993) called 'Greater statistics', and with what Wild (1994) called a 'wide view of statistics'. The statistical *scientist* (as opposed to the statistician?) should be concerned with the investigative process as a whole and realize that model building is itself just part of *statistical problem solving* (e.g. Chatfield (1995)). Problem solving, like model building (see Section 3), is generally an *iterative* process (see for example Box (1994) on the continuing search for quality improvement) and involves wider expertise such as

(i) problem formulation, including clarification of objectives,
(ii) consulting skills—the ability to advise and collaborate with investigators from other disciplines and
(iii) the interpretation and communication of the results.

I cannot overstress the importance of thinking carefully about such issues as what problem needs to be solved and what data need to be collected, but say no more about these wider issues here except to note the need for a better *balance* between the three layers of a study, namely

(i) the problem,
(ii) the theory or model and
(iii) the data,

as Leamer (1992) has argued in an econometric context. It is my experience that students typically know the technical details of regression for example, but not necessarily when and how to apply it. This argues the need for a better balance in the literature and in statistical teaching between *techniques* and problem solving *strategies*.

A discussion of component (d) of model building is deferred until Section 6. The model fitting component (b) usually appears straightforward nowadays, thanks to packages which can estimate the parameters of most types of model (though there is a danger that the analyst will choose a model to fit the software rather than vice versa). Packages also typically carry out a range of routine model checks. In contrast, model formulation is often much harder. The more recent references give guidance on model selection methods for choosing a 'best' model from two or more prespecified models having different structures, but rather little help on model formulation in its widest sense—how do you choose the models to be considered? This is arguably the most important and most difficult aspect of model building and yet is the one where there is least help (honourable exceptions include Leamer (1978) and Gilchrist (1984)). A model may be specified partly or wholly on external subject-matter grounds or from past data but is increasingly determined partly or wholly from the present data, perhaps by searching over a wide range of models by using modern computing power. Then the analyst will typically select the model which is best acccording to some predetermined criterion. Having done this, the analyst proceeds to estimate the parameters of this best model by using the *same* techniques as in traditional statistical inference where the model is assumed known *a priori*. It is 'well known' to be 'logically unsound and practically misleading' (Zhang, 1992) to make inferences as if a model is known to be true when it has, in fact, been selected from the *same* data to be used for estimation purposes. However, although statisticians may admit this privately (Breiman (1992) calls it a 'quiet scandal'), they (we) continue to ignore the difficulties because it is not clear what else could or should be done. Little theory is available to guide us, and the biases which result when a model is formulated and fitted to the *same* data are not well understood. Such biases will be called *model selection biases*. This term is a slight generalization of the term 'selection bias' introduced by Miller (1990), p. 111, which referred only to biases in estimates of regression coefficients.

Even when a model is supposedly known *a priori*, it is advisable to remember that there will still be model uncertainty in that the model may be 'wrong' or at best an approximation. Today's analyst is unlikely to proceed without conducting some exploratory data analysis and model checks, and so subsequent inferences may be biased by being carried out conditionally on some features of the data having been examined or tested.

There are typically three main sources of uncertainty in any problem (Draper *et al.*, 1987; Hodges, 1987):

(a) uncertainty about the structure of the model;
(b) uncertainty about estimates of the model parameters, assuming that we know the structure of the model;
(c) unexplained random variation in observed variables even when we know the structure of the model and the values of the model parameters.

Uncertainty about model structure can arise in different ways such as

(i) model misspecification (e.g. omitting a variable by mistake),
(ii) specifying a general class of models of which the true model is a special, but unknown, case or
(iii) choosing between two or more models of quite different structures.

Statistical theory has much to say about (b) and (c) and about the mechanics of the choice in (ii) (e.g. *F*-tests in analysis of variance (ANOVA)), but it has little to say about (iii) and even less about (i), and largely ignores the effects of (a) in ensuing inferences. This is very strange given that errors arising from (a) are likely to be far worse than those arising from other sources. For example, multiple-regression theory tells us about the errors resulting from having estimates of regression coefficients rather than their true values, but these errors are usually much smaller than errors resulting from misspecification, such as omitting a variable by mistake, failing to include non-linear terms, or failing to take account of the fact that the explanatory variables have been selected from a larger set.

This paper discusses model uncertainty in general. In particular it demonstrates the non-trivial biases which can result from data-dependent specification searches. Methods for assessing the size of the problem and of overcoming it are discussed but no simple general solution is found. This partially explains why so little is said about model uncertainty in the statistical literature. Valiant exceptions include

Leamer (1978) (especially chapter 1—a book sadly neglected by statisticians), Hodges (1987), the collection of papers in Dijkstra (1988), Miller's (1990) study of subset selection in multiple regression, Faraway's (1992) simulation study of regression model selection, Pötscher (1991a), Draper's (1995) review of the Bayesian model averaging approach and the work of Hjorth (1982, 1987, 1989, 1990, 1994) and Hjorth and Holmqvist (1981). Yet as computers allow us to examine and compare increasingly more models, the problem is becoming increasingly serious. Perhaps the main message of this paper is that, when a model is formulated and fitted to the same data, inferences made from it will be biased and overoptimistic when they ignore the data analytic actions which preceded the inference. Statisticians must stop pretending that model uncertainty does not exist and begin to find ways of coping with it.

2. EXAMPLES

We begin with some simple examples to illustrate the effects of formulating and fitting a model to the same set of data.

2.1. *Example 1: Estimating the Mean of a Normal Distribution*

A basic inference problem is that of estimating the unknown mean of a normal distribution from a simple random sample. In practice the analyst will rarely *assume* normality *a priori*, but rather will start by assessing whether the data really are (at least approximately) normally distributed. This can be done with a formal test of significance or more informally by just looking at a histogram or graph of normal scores. The analyst may also consider transforming the data as well as rejecting or adjusting outlying values to make the data 'more normal'. (Whether and when such actions are justifiable is of course another matter.) The analyst proceeds to estimate the mean only if the data 'pass' this assessment procedure, possibly after some manipulation. The whole data analytic process can be regarded as a form of model building and the resulting normal assumption as the model. Subsequent inferences should then really be carried out conditionally on this preliminary assessment, but in practice the preliminary data analysis is customarily ignored. What effect does this have? I am not aware of any help in the literature on this question. Moreover we should perhaps step back from the specific inference prolem and ask more broadly why the data have been collected and what background information is available. In other words we should also ask whether, and to what extent, problem formulation affects inference.

2.2. *Example 2: Linear Regression*

A bivariate random sample is taken on a response variable Y and a possible explanatory variable x to fit a linear regression equation of the form $E(Y|x) = \alpha + \beta x$. A common procedure (rightly or wrongly) is to find the least squares estimator of β, say $\hat{\beta}$, and then to fit the line provided that $\hat{\beta}$ is significantly different from 0. Having done this, the analyst must realize that $\hat{\beta}$ is no longer unbiased for β, but that its properties will depend on the data analytic actions which preceded the calculation of $\hat{\beta}$. If we restrict attention to those cases where a line *is* fitted, the appropriate (conditional) expectation of $\hat{\beta}$ is

$$E(\hat{\beta}|\hat{\beta} \text{ is significantly different from 0}).$$

It is intuitively obvious that this conditional expectation is *not* equal to β as can readily be demonstrated either analytically or by simulation. The bias will be negligible when β is 'large' (where the meaning of large depends of course on the sample size and the residual variance) but may be substantial (e.g. over 40% in one simulation) and of practical importance when the residual variance is large and/or the sample size is small. Essentially the bias arises because we may choose an underparameterized model. The bias will vanish asymptotically.

If we regard *not* fitting a line as a special case of linear regression with $\beta=0$, then the (unconditional) estimator that is actually being used here may be written in the form

$$\hat{\beta}_{\mathrm{PT}}=\begin{cases}\hat{\beta} & \hat{\beta} \text{ is significant,}\\ 0 & \text{otherwise.}\end{cases}$$

In this form it can be seen that it is a simple example of what econometricians call a *pretest* estimator (e.g. Judge and Bock (1978)). It is immediately apparent that $E(\hat{\beta}_{\mathrm{PT}})$ is not generally equal to the unconditional expectation $E(\hat{\beta})$ which assumes that the least squares line is always fitted. Moreover it can be shown that the sampling distribution of $\hat{\beta}_{\mathrm{PT}}$ has a different variance, and a different shape, from that of $\hat{\beta}$.

The two morals of this example are that

(a) least squares theory does not apply when the same data are used to formulate and fit a model, and

(b) the analyst must always be clear exactly what any inference is conditioned on.

2.3. *Example 3: Multiple Regression*

The bias in example 2 is magnified in multiple regression when subset selection of the explanatory variables is allowed (e.g. Miller (1990), Hurvich and Tsai (1990) and Pötscher (1991b)). A typical example cited by Miller (1990), p. 92, from Rencher and Pun (1980) is the following. Generate n random variables on a normally distributed response variable and on k *independent* additional variables which will be treated as if they were potential explanatory variables. Thus the true model here is the null model, but suppose that we nevertheless select the best subset of p 'explanatory' variables by using Efroymson's algorithm and evaluate the resulting coefficient of determination, R^2. This procedure can be repeated many times to obtain the null distribution of R^2 by simulation. When $n=20$, $k=10$ and $p=4$, the average value of R^2 is found to be 0.42 with upper percentile $R^2_{0.95}=0.66$. The 'usual' test on the observed value of R^2, which depends on n and p only and ignores the subset selection, has $R^2_{0.95}=0.45$. Thus an observed relationship obtained by the above procedure for which $0.45<R^2<0.66$ would look 'interesting' and be judged 'significant' by the usual test, but could be spurious. Notice that four variables can be chosen from 10 variables in 210 ways, so that 210 models are effectively considered. If data analytic actions such as outlier rejection are allowed, the effective number of models is even higher so that inferences which ignore the model selection procedure will be even more biased (e.g. Adams (1991), Kipnis (1991) and Faraway (1992)).

Of course we could argue that this example is being unfair to statisticians in that it could be silly to choose the best four variables when not all the relevant coefficients are significant. However, the point of the example is to demonstrate the nature of model selection bias rather than to attempt to simulate a more realistic, but even more complex, model building strategy.

The moral is that subset selection can be dangerous using traditional inferential methods which do not take account of the model selection process.

2.4. *Example 4: An Autoregressive Model*

Consider the first-order autoregressive (AR(1)) time series model, namely

$$X_t = \alpha X_{t-1} + \epsilon_t$$

where $|\alpha| < 1$ for stationarity and $\{\epsilon_t\}$ are independently and identically distributed (IID) $N(0, \sigma^2)$. Suppose that n observations are generated from this model (together with an appropriate start-up sequence to obtain a suitable value for X_0). It is straightforward to fit an AR(1) model to the data, but suppose that we are not sure whether the model is really appropriate (as would normally be the case for real data). The identification process for autoregressive integrated moving average (ARIMA) models is complex and hard to formalize. So for illustration consider the following simple (perhaps oversimplified) time series version of the procedure in example 2, namely

(a) calculate the first-order autocorrelation coefficient r_1,
(b) test the value of r_1 to see whether it is significantly different from 0 and
(c) if it is, estimate α and fit the AR(1) model, but, if not, assume that the data are white noise.

Taking $n=30$ and $\alpha=0.4$ as an example, 250 time series were independently simulated, the resulting value of r_1 was calculated for each series and then an AR(1) model was fitted if r_1 was significantly different from 0 (using the approximate critical value $2/\sqrt{n}=0.36$). At first we used the ordinary least squares (Yule–Walker) estimator for α based on r_1, forgetting that this is seriously biased for small values of n. The simulated unconditional mean of $\hat{\alpha}_{\mathrm{YW}}$ was 0.319 which is in line with the theoretical result in Kendall *et al.* (1983), p. 552. This is 20% *below* the true value of 0.4 and is also worse than the asymptotic results of Shaman and Stine (1988) would suggest. The simulated conditional mean of $\hat{\alpha}_{\mathrm{YW}}$ when r_1 *is* significant turns out to be 0.484. This is more than 20% *above* the true value and so the model selection bias has cancelled out the bias in the Yule–Walker estimate but introduced as large a bias in the opposite direction.

The bias in the (unconditional) Yule–Walker estimate reminds us that there can be serious biases in ARMA model parameter estimators for small samples (e.g. Ansley and Newbold (1980)) and that different estimation procedures (which depend primarily on how the start-up observations are treated) can give substantially different results for small samples (e.g. de Gooijer (1985)). When the above simulation was repeated using the non-linear least squares estimation procedure in the MINITAB package, the unconditional mean of $\hat{\alpha}$ was found to be 0.39 whereas the conditional mean exceeds 0.5. Thus a nearly unbiased estimator is turned into an estimator with a serious bias.

The above model selection procedure is much simpler than would normally be the case in time series analysis. It is more usual to inspect the autocorrelations and the partial autocorrelations, to allow differencing, to allow the removal or adjustment of outliers and to entertain all ARIMA models up to say third order. Choosing a best model from such a wide set of possibilities seems likely to make the model selection biases even larger.

Hjorth's (1994) example 2.2 discusses the related case of distinguishing between an AR(1) and an AR(2) model; two other interesting time series examples from Hjorth (1987) are discussed in Section 4.1.

The moral of this example is that estimation biases are likely to be widespread in time series analysis where it is standard practice to formulate and fit a 'best fitting' model to the (one and only) data set.

2.5. *Example 5: What is the Problem?*

Problem formulation is crucial in the possible presence of model uncertainty. An example from time series analysis will make the point. Much effort (e.g. Ahn (1993)) has been devoted to developing methods for *testing for the presence of a unit root*, which would mean that the given series is non-stationary, but that its first differences *are* stationary. Although the presence of a unit root can be of particular interest (e.g. in the search for co-integration), it is hard to see why the presence of a unit root should be chosen as the *null* hypothesis (and Leybourne and McCabe (1994) provide a different approach where it is the *alternative* hypothesis). The desire to carry out many tests stems from the ingrained idea that there is a true model, and from the implicit notion that a unit root either exists or does not exist. In practice we shall never know whether a unit root really exists, or whether such a structure is appropriate for *part* of the series, or whether the degree of differencing changes over time or whether there is some other explanation for apparent non-stationary behaviour. Rather than carrying out such a test (which may in any case give inaccurate levels of significance or power), it could be better to admit the possibility of model uncertainty and to allow for this by making deductions based on averaging over several plausible alternative models, or by choosing a flexible procedure which does not force a particular form of model on the data. For example in forecasting it is generally preferable to model changes in level with a local linear trend, which can vary stochastically, rather than to adopt a deterministic linear trend. The point is that a test for a·unit root is unlikely to be the main objective of the analysis, and could be positively unhelpful in diverting attention from the need to find a flexible approach to solve the given problem.

3. MODEL BUILDING

The overall model building process involves formulating, fitting and checking a model in an *iterative, interactive* way (e.g. Box (1976, 1980)). Model estimation is generally carried out on the assumption that the model is known *a priori* and is true (Box (1994), p. 221). This means that it should have been prespecified on subject-matter considerations such as accepted theory, expert background knowledge and prior information including that obtained from previous similar data sets (though not necessarily in a Bayesian way). Expert background knowledge could include knowing which variables to include, and making sure that the model allows for known constraints (on both the variables and the model parameters) and for known limiting behaviour. However, the external specification of a model does not mean that model uncertainty is eliminated, since the 'expert' may for example erroneously omit an important variable. Our knowledge about the world is always incomplete (Box (1993), p. 3). Thus the unexplained random variation will depend not only on unknown variations in sampling units and nuisance variables but also on all the *ignored* variables and factors. Model uncertainty seems likely to be more serious in what W. E. Deming has called an *analytic study* (e.g. Hahn and Meeker (1993)) and in scientific areas (e.g. economics) where careful enumeration and control of variables, as in laboratory-based experiments, is not possible. Proxy, or surrogate, variables are sometimes used to try to account for missing variables but it is not obvious in general how to deal with model misspecification.

One possible way to circumvent some types of model uncertainty is to use *nonparametric procedures* which make far fewer model assumptions. Although such methods have their place, particularly in hypothesis testing, they are outside the scope of this paper. Likewise we say nothing about *robust* procedures which can avoid problems due to misspecification of secondary assumptions (e.g. Cox and Snell (1981), p. 18) but do nothing about the primary assumptions judged central to the problem.

This paper is concerned mainly with models that are not fully specified *a priori*, but rather are formulated, at least partially, by looking at the *same* data as those later used to fit the model. This practice is increasingly common. It arises in submodel selection in such areas as time series analysis, regression, generalized linear modelling, ANOVA and the analysis of discrete data, as well as in the situation where the analyst looks at a new set of data with virtually no preconceived ideas at all. The rather derogatory terms *data mining* (e.g. Lovell (1983)) and *data dredging* are sometimes used in this context to describe procedures of the last type, particularly when the analyst eschews careful thought based on external knowledge in favour of deriving the best possible fit from a large number of entertained models. The extent of data mining is unclear, though my, admittedly subjective, impression is that certain forms of it are widespread, particularly in subset selection procedures and in time series analysis. The analyst who is willing to entertain any subset of 10 possible explanatory variables with only 20 observations is displaying not so much a caricature but more a somewhat extreme version of behaviour which can be all too familiar. The effect of data mining is not well understood in general. Some limited results are known—see Section 4—but, in most areas of statistics, inference seems to be generally carried out as if the analyst is sure that the true model is known. It is indeed strange that we often admit model uncertainty by searching for a best model but then ignore this uncertainty by making inferences and predictions as if certain that the best fitting model is actually true.

40 years ago it may have been true that a *single* model was typically fitted to a given set of data. Nowadays the increase in computing power has completely changed the way in which statistical analyses are typically carried out (not necessarily for the better!). For example Leamer (1978) distinguished six different *approaches* to model building, called *specification searches*, namely the data-dependent process by which a researcher is led to select a particular model specification. A model is often selected from a wide class of models by optimizing a statistic such as the adjusted R^2 or Akaike's information criterion (AIC), and there are many references on model selection, especially in time series analysis—see for example the reviews by de Gooijer *et al.* (1985) and Choi (1992). The data analysis procedure may also involve strategies such as

(a) excluding, downweighting or otherwise adjusting outliers and influential observations and
(b) transforming one or more variables, for example to achieve normality, additivity and/or constant residual variance.

As a result the analyst may in effect consider tens, hundreds or even thousands of models, and there is a clear risk that the search for a good fit will turn into data mining. The use of transformations and the deletion of outliers are particularly dangerous actions except where they can be justified on subject-matter grounds. Outliers for example should be discarded only if they are thought to be non-exchangeable with other observations on good substantive grounds. Otherwise predictive uncertainty will be underestimated. If the position is unclear, it may be advisable to carry out two analyses, both with and without outliers. If the findings differ, both should be reported.

Unfortunately statistical theory has not kept pace with this computer-led revolution, and still typically assumes that the model is known. Yet, as illustrated in example 2, standard least squares theory, which we (nearly) all teach and use, does not apply when the same data are used to formulate and fit a model. Unfortunately there has been very little published work on inference *after* model selection, as reviewed in Section 4. The analyst needs to assess the model selection *process* and not just the best fitting model (Hjorth, 1989; Kipnis, 1991), but this is difficult in practice when complicated screening procedures are used where the rules of search may be informal and may involve subjective judgment. As such, they are hard to put in a satisfactory mathematical framework and may not be amenable to theoretical analysis. Even when a model *is* data driven in a clearly defined way, the frequentist approach still cannot readily handle model uncertainty. This is no doubt why we 'too often concentrate on the deductive bit (statistical inference) and pretend the rest does not exist' (Box, 1990). It is also relevant to read Tukey's (1991), p. 128, remarks comparing the classical text-book paradigm with an alternative real life paradigm which does allow for the possibility that the model is unknown, that informal judgments must be made (based on simulation and experience as well as mathematics) and that no formal structure may be possible. What is clear is that most references on parameter estimation disregard the model selection process and are therefore fundamentally incomplete.

The literature on *model checking* seems equally suspect. It is known to be theoretically desirable for a hypothesis to be validated on a second confirmatory sample (see Section 6), but this seems to be rather rare in practice (except perhaps in clinical trials). Rather, diagnostic checks are typically carried out on the *same* data as those used to fit the model. If necessary the model is then modified and a revised model fitted. This iterative process can continue indefinitely, but still *using the same data*. Now diagnostic tests typically assume that the model is specified *a priori* and calculate a *P*-value as Probability(more extreme result than the one obtained|model is true). But, if the model is formulated, fitted and checked using the same data, then we should really calculate Probability(more extreme result than the one obtained|model has been selected as 'best' by the model formulation procedure). It is not clear in general how this can be calculated. What is clear is that the good fit of a best fitting model should not be surprising!

3.1. *Is there a True Model?*

A crucial question in model building is the attitude that one takes to the existence of a true model. By assuming exact knowledge of the model structure, estimation theory implicitly assumes that an exact true model does exist. In practice no-one really believes this. For example Tukey (1994) suggested that we need more honest foundations for data analysis which do not rely on 'assuming that we always know what in fact we never know', whereas Fildes and Howell (1979) say that 'It is a truism of forecasting that the model chosen is misspecified'. The growing disenchantment with classical inference based on a true model is exemplified in a rather extreme way by Tsay (1993) who says that 'Since all statistical models are wrong, the maximum likelihood principle does not apply'. Instead Tiao and Tsay (1994), p. 129, say that

> 'if one accepts the premise that any model is, at best, an approximation, then parameter estimation should be treated more in the context of the use for which the model is to be put rather than as an end in itself'.

This suggests that model builders should adopt a more pragmatic approach in which they search, not for a true model, but rather for a *parsimonious* model giving an adequate approximation to the data at hand—see Box (1976) and Leamer (1978) (especially chapter 6)—and then concentrate on determining the model's *accuracy* and *usefulness*, rather than with testing it (Leamer, 1992). The idea that some models are useful whereas others are not (e.g. Box (1976) and de Leeuw (1988), p. 120) is expressed in the well-known saying that 'All models are wrong, but some are useful'. Clearly the *context* and the *objectives* are key factors in deciding whether a model is 'good' and useful. As well as giving more attention to how a model will be used (and less to optimizing the goodness of fit), intelligent model building should also consider the question of *costs*. For example, when considering whether a possible additional explanatory variable is worth having in multiple regression, the question should not be 'Does it lead to a significant improvement in fit?' but 'Does it provide value for money in improving predictions?'.

The notion that there is no such thing as a true model, but rather that model building is a continual iterative search for a better model, is arguably in line with the general philosophy of science. Whereas statistics is often regarded as an *inductive* science (data → model) and probability theory as a *deductive* science (model → behaviour), Popper (1959) asserted that scientific theories are not generally derived inductively from observations. Rather they are invented as hypotheses, speculations and guesses and then subjected to experimental tests. A theory is scientific only if it is in principle capable of being tested and hence is open to the risk of refutation. Popper (1959) also says (p. 251) that 'theories are not verifiable, but they can be "corroborated"'. In other words a theory, like a statistical model, is never 'proved', even when there is extensive empirical justification for it, but it may be disproved or discredited. My view is that the iterative model building process involves a mixture of inductive and deductive reasoning, whereby we search, not for a true model, but rather for a better, and more general, approximate model for data of a similar type collected under possibly different conditions (see Section 6).

An alternative possibility is that there may be *more than one* model which may be regarded as 'useful' (i.e. as a sufficiently close approximation to the data for the purpose at hand). For example Poskitt and Tremayne (1987) discussed how to obtain a *portfolio* of plausible models. The notion of having more than one model is a key element of the Bayesian model averaging approach (see Section 5) which avoids having to select a single best model but rather averages over more than one model. The notion is also implicit in the *combination* of forecasts (e.g. Clemen (1989)) wherein time series forecasts are produced by taking a weighted linear combination of the forecasts obtained from a range of different methods and/or models. A completely different possibility is to use different models to describe different parts of the data, rather than to pretend that a single model can describe all the data. This applies particularly to time series analysis where the properties of the most recent data may differ markedly from those of earlier data and a global model fitted to all the data may give poor predictions.

If we do nevertheless select a single model based on some best fit criterion, then some sort of *sensitivity analysis* (e.g. Leamer (1985)) seems desirable to see how sensitive any conclusions are to the model assumptions and to guard partially against the dangers of data mining. Unfortunately this seems to be rarely attempted.

The more complicated the model that is chosen, the more likely it is that there will be departures from one or more of the model assumptions. The dangers of *overfitting* are 'well known', particularly in multiple regression and when fitting lagged variables in time series models, but these dangers are not always heeded. Although a more complicated model may appear to give a better fit, the predictions from it may be worse. Moreover, the inclusion of unnecessary explanatory variables has cost implications in that superfluous data will have to be collected and processed. *Neural networks* form another class of models which may lead to overfitting. They have been used successfully in some applications, such as pattern recognition, but have recently been suggested for use in time series forecasting. The large number of parameters (and architectures) which may be tried means that they can usually be made to give a good within-sample fit. However, their forecasting ability is still unproven (Chatfield, 1993a), and arguably unpromising, given that past empirical studies suggest that simple time series models often give as good forecasts as more complicated models. Fildes and Makridakis (1994) complained that these empirical findings are ignored by theoreticians who continue to derive results on inference and forecasting which assume the existence of a true model. Likewise Newbold *et al.* (1993) pointed out the difficulty of deciding on the correct form of differencing when fitting ARIMA models. Having the 'wrong' form of differencing may make little difference for short-term forecasts where

> 'the fiction that the analyst has discovered the "true" model is innocuous. Such fiction, however, is far from innocuous when attempting to base inference about long-run behavior on these fitted models.'

Mention of time series forecasts brings to mind the distinction between estimating *unobservable* quantities, such as population parameters, and predicting *observable* quantities, such as future values of a time series. A problem with the former is that the analyst will never know whether the inferences are good since the estimates cannot be compared directly with the truth. We arguably need more emphasis on predicting observables (e.g. Geisser (1993)) because such quantities can be assessed or *calibrated,* are less dependent on the existence of a true model and are vital in assessing whether a model really is useful. A related point is that models which are mathematically very different may be virtually indistinguishable in terms of their fit to a set of data but give very different predictions outside the range of the data, and this is another reason for not necessarily trying to pick a single best fitting model.

4. MODEL SELECTION BIASES

This section takes a general look at model selection biases and considers

(a) how to assess the size of the problem and
(b) how to overcome or circumvent the problem.

Cohen and Sackrowitz (1987) say that 'inference following model selection based on data is widespread among statistical practitioners' and that 'statistical research on such procedures is fairly extensive'. This may be true in regard to questions such as assessing whether a model of the correct order is chosen asymptotically and controlling the overall probability of an error of type I when a series of data-dependent hypotheses is tested. However, as Pötscher (1991a) pointed out, there has been very little research on inference after model selection. Bhansali (1981) and Shibata (1976) appear to be addressing the problem when they evaluate the effect of not knowing the order of an AR process on the mean-squared error of prediction, but in fact they assume that the model selection and prediction are performed on *independent* processes, albeit with the same probabilistic structure. This is not a situation which I have come across and is not the situation considered in this paper.

Pötscher (1991a) derived two loosely connected results, namely

(a) model parameter estimates are asymptotically consistent (which means that the bias problem vanishes asymptotically) when model selection criteria are used which are consistent (e.g. the Bayesian information criteria—Choi (1992)) but also for some other criteria (e.g. the AIC), and
(b) the asymptotic distribution of parameter estimators is unaffected by model selection if the selection procedure is consistent but in some other cases (e.g. AIC and Mallows's C_p) the asymptotic distribution *will* be different from the 'usual' distribution and can be calculated.

Generally speaking the variance will increase as might be expected from the additional uncertainty due to the model selection process. The shape of the distribution may also change. Zhang (1992) also looked at asymptotic results for inference on linear regression models when the final prediction error criterion (e.g. de Gooijer *et al.* (1985)) is used to select a model and showed that the asymptotic estimate of error variance is satisfactory but that asymptotic confidence regions for unknown parameters are generally too small in that coverage probabilities are less than nominal probabilities. The question then is whether these asymptotic results help us for finite samples. Certainly they emphasize that, even asymptotically, results may be different from the 'usual' results which ignore the model selection procedure. Thus model selection biases are not just a 'small sample' problem, although they do tend to be worse for small samples (though a potential danger is that more data mining will be attempted for larger samples, thereby negating the effects of increased sample size). Clearly more work is needed to see whether asymptotic results are relevant in the finite sample case.

Some useful non-asymptotic results are given by Hjorth (1989, 1994). They rely on the fact that the use of a model selection statistic essentially partitions the sample space into disjoint subsets. This approach enables the derivation of various inequalities regarding the expectation of the optimized statistic and also gives further understanding about estimates of model parameters after model selection. For simplicity this paper presents a simplified account which restricts attention to distinguishing between just two models, say M_1 and M_2 (neither of which need necessarily be true), and uses a sensible statistic, say the AIC, to make the choice. This means that we select M_1 for a data set whenever the AIC for M_1, denoted AIC_1, is less than that for M_2, denoted AIC_2. This effectively partitions the sample space Ω into two disjoint subsets (assumed non-empty), say A_1 where M_1 is selected and A_2 where M_2 is selected. Hjorth (1989) distinguished between *global* parameters which are defined for all models (such as the mean or median) and *local* parameters which are not defined for all models (such as AR coefficients in competing AR models of different order). Suppose that we are interested in estimating a (scalar) local parameter of M_1, say θ, and we have an estimator $\hat{\theta}$, which might for example be the maximum likelihood estimator. The properties of $\hat{\theta}$ are normally found by taking expectations over the whole sample space, conditional on the model being true. However, when estimation follows model selection, as in the above case, the properties of $\hat{\theta}$ should arguably be found by taking expectations over A_1. There is no reason why $E(\hat{\theta})$ evaluated over A_1 should equal the expectation over Ω and in general the two quantities will indeed be unequal (as demonstrated by simulation in example 2 for the local parameter β and in example 4 for the local parameter α). It follows in particular that, if the estimator $\hat{\theta}$ is unbiased when used without selection, it will generally be biased when used *after* selection. However, note that the properties of $\hat{\theta}$ thus derived are conditional on the assumptions that

(a) M_1 is true and
(b) M_1 is selected when the choice is M_1 or M_2.

It is not clear whether such restrictive conditional results have any real general value other than to alert us to the implications of inference after model selection.

Suppose instead that θ is a global parameter. Then the properties of a *global estimator* (defined in different ways for M_1 and M_2) can be found by taking expectations over Ω, but the contribution from A_1 (which assumes M_1 true) will be of a form different from that from A_2 (which assumes M_2 true). Hjorth's (1989) example 2 is an example where the global estimator is biased even though the estimators for each of the individual models are both unbiased. As Hjorth (1989), p. 107, says, when studying the properties of such a global estimator from a frequentist point of view, we must convince the user to consider, not only the selected model, but also all rejected models and estimators. This is difficult, but we must get over the key message that *the properties of an estimator may depend, not only on the selected model, but also on the selection process* (Hjorth, 1990, 1994).

We can also say something about the properties of the statistic used to make the model selection. It is well known that the fitting of a model typically gives optimistic results in that performance on new data is on average worse than on the original data—Picard and Cook (1984) called this 'The Optimism Principle'. Hjorth (1989) gave a rather neglected bias theorem which appears intuitively obvious (and can readily be proved) when looked at from the partitioned sample space point of view. Essentially it says that $E(\mathrm{AIC}_{\min})=E\{\min(\mathrm{AIC}_1, \mathrm{AIC}_2)\}<E(\mathrm{AIC}_i)$ for both $i=1$ and $i=2$. Thus if M_1, for example, happens to be the true model, the expectation of

$\mathrm{AIC}_{\min}$ *after* the model selection process (where we sometimes choose M_2 by mistake because it happens to give a lower AIC) is lower than the (unconditional) expectation of AIC_1. As Hjorth (1989) says

> 'it is perhaps not surprising that selection minimizing a criterion will cause underestimation of this criterion'.

A similar result applies to any sensible loss function, including estimates of residual variance which are unbiased for a particular model over the whole sample space.

Turning now to hypothesis testing, most statisticians realize that, if a hypothesis is generated and then tested using the same set of data, the usual *P*-value is potentially misleading especially if attention is focused on some 'unusual' or 'unexpected' feature of the data. However, it is often unclear ***how*** to adjust the *P*-value or even whether it has any value at all. (The Bonferroni correction to the *P*-value for the most extreme of a set of statistics (e.g. Chatfield (1995)) is a rather unsatisfactory approximation.) It is disturbing that many research papers report tests only if they yield 'significant' results. This practice is rightly deplored (e.g. by Dawid and Dickey (1977)) since it will conceal the selection process which led to these particular hypotheses being considered and reported. When (many) non-significant results are *not* reported, there is a clear danger of giving too much credence to the significant results (and sometimes a lack of significance is what is really wanted anyway). *In any data-instigated procedure, the analyst must be clear what the analysis is conditioned on.* More generally it is difficult to assess the effect of carrying out, not one test, but a whole series of tests, as for example in multiple-comparison problems, in multiple-specification tests and in the sort of sequential testing which may arise in ANOVA (e.g. Azzalini and Cox (1984)). The emphasis in published research has been on controlling the overall probability of a type I error (e.g. Phillips and McCabe (1989)) rather than on assessing other consequences of multiple testing. It may be possible to allow explicitly for the fact that a null hypothesis may be (at least partly) determined by the data, as in the Lilliefors variation of the Kolmogorov test for normality (e.g. Sprent (1993), p. 77), but this is the exception rather than the rule.

In multiple regression, the use of subset selection methods is well known to introduce alarming biases (see example 3). Miller (1990), p. 160, suggests that 'there can be biases of the order of one to two standard errors in the estimates' of regression coefficients. Miller (1990) and Kipnis (1991) have shown that 'traditional' results are overoptimistic and biased with regard to assessing the mean-square prediction error (MSPE). Hjorth (1982) showed that prediction errors for time series regression data are much larger when explanatory variables are selected from the data than when a predetermined model of the same order is specified. Unfortunately these results are often (usually?) ignored in practice.

Similar biases arise for other classes of model though it is hard to find any general results on the size of such biases. The bias in estimating the MSPE seems particularly alarming. From the optimism principle (see above) the within-sample fit of a model is typically better than out-of-sample forecasts or the fit to a new sample of data. This is true in regression (the shrinkage effect—see Section 5), in time series analysis (see Section 4.1) and in other problems (e.g. Efron and Tibshirani (1993), p. 239). For example anyone who has tried discriminant analysis will know that the within-sample error rate is typically better (often much better) than the out-of-sample error rate. This explains why measures of model fit such as Mallows's

C_p and the AIC can be highly biased in data-driven model selection situations (and yet the 'naive use of C_p persists' (Breiman, 1992)). These problems can be partially overcome by the use of resampling techniques (see below).

4.1. *Time Series Analysis*

Model selection biases seem likely to be particularly serious in time series analysis, where we cannot normally replicate a data set. Occasionally a time series model may be based on background theory (e.g. econometric theory) or on a model fitted to time series of a similar type. However, this is exceptional and most time series analyses follow an iterative cycle of model formulation, estimation and diagnostic checking, as in the Box–Jenkins model building procedure (Box *et al.* (1994), section 1.3.2). Yet little is known about the biases that such a procedure will generate.

Suppose that we start a time series analysis by entertaining the class of ARIMA(p, d, q) models for say $0 \leqslant (p, d, q) \leqslant 2$, giving a total of 27 possible models. Although fewer than the 210 models entertained in example 3, the number is still sufficiently large to indicate substantial model uncertainty and to make it likely that model selection biases will arise. Furthermore the number of models entertained may increase during the analysis, as for example if seasonality is found (suggesting a seasonal ARIMA model), or non-normality (suggesting a transformation), or outliers, or non-linearities (suggesting a completely different class of models), or discontinuities, or interventions or whatever. Thus it is hard to see how general theoretical progress can be made on evaluating the extent of such biases since any results are conditional on the particular model selection procedure used.

An example of simulation results is Hjorth's (1987) example 5. Data are generated from an ARMA(1, 1) model and the model selection procedure allows the AR and MA orders to be as high as 3 and minimizes the estimated MSPE. The correct type of ARMA model was found in only 28 out of 500 series. The properties of the estimates for the 28 series differed greatly from those for all 500 series. The model selection bias for the MA parameter was particularly bad. For series length 50 and a true MA parameter of -0.4, the average estimated value for all 495 series giving estimates satisfying the invertibility and stationarity conditions was -0.413 but was -0.528 for the 28 series where an ARMA(1, 1) model was correctly selected. Hjorth also found alarming results concerning estimates of the MSPE. For each series the best model was found and the estimated MSPE was calculated. The latter could be compared with the true MSPE for the true model as well as with the true MSPE for the fitted model, both of which are known or can be calculated as the series are simulated. The average estimated MSPE was less than the true MSPE for the true model and *less than a third* of the true MSPE for the model which was actually fitted. Once again the best fitting model from a range of entertained models will make us think that we have a better fit than we really do, whereas our predictions will generally be *much* worse than expected.

Hjorth's (1987) example 7 illustrates the effect of model selection bias on estimates of the MSPE for multivariate time series models. Forecasts were required for one particular series in a real data set consisting of 28 series. The number of models which could be entertained was enormous. Using external knowledge, experts selected just five series to base forecasts on. The resulting model was compared with the best fitting model using a subset of all 28 series. The latter naturally had a lower mean-squared error as regards fit but gave worse predictions as judged by forward validation (a time series version of cross-validation—see Hjorth (1994), chapter 4).

An immediate consequence of underestimating the MSPE is that *prediction intervals will generally be too narrow*. Empirical studies have shown that nominal 95% prediction intervals will typically contain less than 95% of actual future observations. This happens for a variety of reasons (see Chatfield (1993b), section 6) of which model uncertainty is perhaps most important. The model may be incorrectly identified or may change through time. The one-step-ahead prediction error variance is often taken as $\sigma^2(1+p/n)$ where σ^2 denotes the residual variance and the factor $1+p/n$ reflects the effect of *parameter* uncertainty when estimating a p-parameter model using a sample size n (e.g. Hjorth and Holmqvist (1981), section 1). This factor takes no account of model uncertainty and is in any case often omitted. Moreover the estimate s^2 of σ^2 is typically too small when a best fitting model has been selected. (In contrast prediction intervals for general linear models customarily *do* take proper account of parameter uncertainty and so simpler models can give estimates with shorter confidence intervals (e.g. Regal and Hook (1991)). Failure to reject a null model as an alternative to a more complex model is not the same as establishing that the simpler model is closer to the truth. There is an alarming tendency for analysts to think that narrow intervals are good when wider intervals may reflect model uncertainty better.) Steerneman and Rorijs (1988) illustrated the consequences of overfitting and data mining in an econometric forecasting context and recommended parsimonious, economically meaningful, models. Draper (1995) considered an instructive example concerning forecasts of the price of oil. 10 models were entertained which gave a wide range of point forecasts that were nevertheless all well away from the actual values which resulted. There were also large differences in the prediction error variances. A model uncertainty audit suggested that only about 20% of the overall predictive variance could be attributed to uncertainty about the future conditional on the selected model and on the assumptions (the scenario) made about the future. Yet the latter portion is all that would normally be taken into consideration.

4.2. *Computational Methods*

Given that analytical methods are generally not available to study the effects of data-dependent procedures, a variety of computational methods have been tried (e.g. Faraway (1992) and Hjorth (1994)). *Simulation* methods are one obvious avenue when the model selection procedure is simple and clearly defined as in example 4. But with any model selection procedure it can be very difficult to formalize, and hence to simulate, the data analytic steps taken by an experienced investigator faced with real data. There is typically a wide choice of possible actions and models, usually involving subjective judgment. However, it would impose too much inflexibility to insist that all procedures be capable of objective description and hence be capable of automation. Faraway (1992) has written a program to simulate the actions taken by a human in a regression analysis, including the handling of outliers and transformations. Though it cannot fully simulate real human behaviour, it does give a reasonable representation. Faraway's program also enabled him to investigate various other computational ways of dealing with model selection bias, including *resampling* or *bootstrapping*, *jackknifing* and *data splitting*. The last technique involves splitting the data into two parts, fitting the model to one part (sometimes called the *construction* sample) and using the second part (sometimes called the *hold-out*, *test* or *validation* sample) to check inferences and predictions.

Several other researchers have tried computationally intensive methods. The results in Dijkstra (1988) are generally disappointing. We must avoid resampling which is conditional on the fitted model, such as resampling the residuals, as this will not reflect model uncertainty (Freedman *et al.*, 1988). Resampling the data is difficult with (ordered) time series data, and, since the model may change over time, may still not reflect the true extent of model uncertainty. More generally if a data set from model A happens to have features which suggest model B, then the resampled data are also likely to indicate model B rather than the true model A. It can also be difficult to compare results when different transformations are used for different bootstrap samples (Faraway, 1992). Nevertheless Faraway's (1992) simulation results from a linear regression model using a variety of error distributions suggest that careful bootstrapping can overcome much of the bias due to model uncertainty. Breiman (1992) suggested a form of resampling called the *little bootstrap* and showed that it can give nearly unbiased estimates of the MSPE in subset selection. Breiman's section 9 is well worth reading, emphasizing again that models selected by using data-driven selection procedures can give extremely optimistic looking results. Hjorth (1994) suggested a form of resampling called the *spectral bootstrap* for stationary time series data which involves resampling in the spectral domain. Another form of resampling which will not be considered here is the use of *cross-validation* (e.g. Efron and Tibshirani (1993), chapter 17, and Hjorth (1994), especially section 3.6).

With *data splitting*, one problem is deciding *how* to split the sample (for example see Picard and Cook (1984)). Fitting a model to just part of the data will result in a loss of efficiency. Faraway (1992) showed that this procedure may greatly increase the variability in estimates without the reward of eliminating bias. Thus hold-out samples, although perhaps unavoidable in time series forecasting, do not provide a genuine substitute for a true replicate sample, which will, in any case, inevitably be collected under somewhat different conditions from those applying to the original sample (for example see Hirsch (1991) and Section 6).

4.3. *Some General Consequences*

Model selection biases are hard to quantify, but the following general points can be made.

(a) *Least squares theory does not apply when the same data are used to formulate and fit a model.* Yet time series text-books, for example, customarily apply least squares methods to time series models even when the model has been selected as the best fitting model from a wide class of models such as ARIMA models.

(b) After model selection, *estimates of model parameters and of the residual variance are likely to be biased.*

(c) The analyst typically *thinks that the fit is better than it really is* (the optimism principle), and diagnostic checks rarely reject the best fitting model *because it is the best fit*!

(d) *Prediction intervals are generally too narrow.*

Despite the limited progress described above, the overall impression is that the frequentist approach to statistical inference does not adapt easily to cope with model uncertainty. The practitioner may be tempted to use 'fudge factors', based partly on theory and partly on empirical experience, to multiply the widths of confidence and prediction intervals to obtain more realistic values. However, this approach will not appeal to many readers. Thus the next two sections describe two completely different types of approach. They do not *solve* the problem of data abuse but they do provide ways round it.

5. BAYESIAN MODEL AVERAGING APPROACH

The promising Bayesian model averaging approach to coping with model uncertainty should appeal, not only to Bayesians, but also to any 'broad-minded' statistician. The key to its success lies in not having to choose a single best model but rather in averaging over a variety of plausible competing models which are entertained with appropriate prior probabilities. Thus priors are attached to the models rather than (just) to model parameters. The data are then used to evaluate posterior probabilities for the various models. Models with 'low' posterior probabilities may be discarded to keep the problem manageable, and then a weighted sum of the remaining competing models is taken. This approach has been recommended explicitly or implicitly by several researchers, and a thorough recent review and methodology discussion is given by Draper (1995). This section therefore can be brief. The broad issues involved may be clarified by looking at example 2 again from a Bayesian point of view.

5.1. *Example 6: Linear Regression Revisited*

Suppose as in example 2 that we have bivariate regression data but are not sure whether to fit a straight line or no relationship at all (as would be the dilemma of a frequentist who found the *P*-value for the estimated slope to be around 5%). Then two models are entertained, namely

$$Y=\alpha_1+\beta x+\epsilon_1 \qquad \text{(model I)},$$

$$Y=\alpha_2+\epsilon_2 \qquad \text{(model II)},$$

where α_1, α_2 and β are constants, and $\{\epsilon_{1i}\}$ and $\{\epsilon_{2i}\}$ are IID $N(0, \sigma_1^2)$ and $N(0, \sigma_2^2)$ respectively. Further suppose that the posterior probabilities have been evaluated from the data as p_1 and $p_2=1-p_1$. (For simplicity we ignore uncertainty about model parameters.) There are now three possible actions that we could take.

(a) Choose the single model with the highest posterior probability and use this to make predictions. However, if predictions are made conditional on the selected model, then the prediction intervals will not reflect the model uncertainty.

(b) Make two predictions. For example at $x=x_0$ the predictions are

$$\hat{y}=\alpha_1+\beta x_0 \qquad \text{with probability } p_1$$

and

$$\hat{y}=\alpha_2 \qquad \text{with probability } p_2.$$

This is not much help if we require a single prediction. Nor is it clear how prediction intervals should be calculated.

(c) Combine the two predictions in (b) to obtain the single weighted prediction

$$\begin{aligned}\hat{y}_c &= p_1(\alpha_1+\beta x_0)+p_2\alpha_2\\ &= p_1\alpha_1+p_2\alpha_2+p_1\beta x_0. \qquad (1)\end{aligned}$$

This combined forecast is effectively what will be given by the Bayesian model averaging approach, and it will have a lower MSPE in the long run than either of the individual forecasts. The approach also allows an assessment of the distribution of $\hat{y}_c$ which takes account of the model uncertainty. The mixed prediction implicitly suggests that there is a combined model for which

$$E(Y|x)=p_1\alpha_1+p_2\alpha_2+p_1\beta x \qquad \text{(model III)}.$$

Does this combined model make sense? *A priori*, we assume that either model I or model II is true, but we are not sure which. After seeing the data, we use model III even though it cannot be true if either of models I or II is true. Whether this makes sense seems to depend on whether or not you really believe that there is a single true model and on whether you want a single prediction.

The slope of the combined model III in example 6, namely $p_1\beta$, is smaller than that of model I. This is reminiscent of the *shrinkage* effect in regression (e.g. Copas (1983)) and in logistic regression (e.g. Copas (1993)) whereby regression equations tend to give a poorer fit to new data than might be expected from the fit to the original data. This applies even when a single model is entertained, but Copas (1983) also noted that

> 'shrinkage is particularly marked when stepwise fitting is used. The shrinkage is then closer to that expected of the full regression rather than of the subset regression actually fitted'.

When the number of variables is 'high' compared with the number of observations, the shrinkage can be so severe that a fitted model is worse than useless (e.g. Copas (1983), example 3, and Copas (1993), example 2). However, note that the shrunken predictor is not uniformly better for time series AR models for finite samples (Copas and Jones, 1987). In contrast, Hill *et al.* (1991) showed empirically that shrinkage estimators can give substantially improved out-of-sample forecasts for a price promotion model used in marketing research, while the related idea of *damping the trend* in Holt's exponential smoothing can also improve forecasts (Gardner and McKenzie, 1985).

Although most time series forecasts are produced by finding a best fitting model and extrapolating it into the future, there are two other commonly used forecasting strategies which are relevant to our discussion. In long-range forecasting, *scenario analysis* (e.g. Schoemaker (1991)) is often used. Here a variety of different assumptions are made about the future giving a range of forecasts, rather than just one. Each forecast is linked clearly to the assumptions that it depends on, and their spread should clarify the extent of model uncertainty. This will allow organizations to make contingency plans for the various possible futures. This type of forecasting corresponds loosely to action (b) above.

A completely different type of strategy arises from *combining forecasts* (in a non-Bayesian way). Suppose that you have produced forecasts by several different methods (e.g. exponential smoothing, ARIMA modelling, state space modelling, an econometric model, . . .). Then it has been established empirically that a weighted linear combination of these forecasts will often be more accurate on average than any of the individual forecasts (e.g. Clemen (1989)). A simple average is often as good as anything. One drawback is that the client does not receive a simple model to describe the data. The stochastic properties of the combined forecast may also be unclear. This type of forecasting corresponds loosely to action (c) above.

To decide how to proceed, it is clearly necessary to clarify the objectives of a forecasting exercise and to find out exactly how a forecast will be used. In particular the analyst needs to know whether a single prediction is required, whether a prediction interval is required and whether a model is required for description and interpretation.

Successful time series applications of Bayesian model averaging are reported by Draper (1995) in predicting oil prices from 10 econometric models, by Le *et al.* (1993) in robust prediction of AR processes when the AR order is unknown and by Schervish and Tsay (1988) also for AR processes. Recently the method has also been applied (Madigan and York, 1995) to graphical models for discrete data where it is possible to specify a large class of conditional independence models. The approach obviates the need for model selection criteria to select a single best model from within the class of models being entertained. The general idea of mixing several models, rather than having to use a single best model, is attractive and is the idea behind the use of multiprocess or mixture models in Bayesian forecasting (West and Harrison (1989), chapter 12).

Despite its promise, there are difficulties in applying Bayesian model averaging. First the calculation of posterior probabilities from the prior probabilities requires the computation of Bayes factors. Kass and Raftery (1994) discussed this problem in general. Closed form Bayes factors exist in some interesting cases and good approximations are available in others (e.g. generalized linear models). Sometimes extensive computation is required, which has become feasible in recent years, perhaps with the aid of Markov chain Monte Carlo techniques. A second problem is that the number of possible models can be very large. One approach here is to reduce the number of models by discarding those with low posterior odds. This requires an arbitrary cut-off point to be chosen. Alternatively the Markov chain Monte Carlo model composition method of Madigan and York (1995) allows complete model averaging to be approximated arbitrarily accurately.

A third problem is that prior probabilities for the various models must be specified and this will not be easy, especially when data-dependent actions are allowed. If some models are entertained only *after* looking at the data (as can happen especially in time series analysis), the priors cannot be applied beforehand but rather some sort of preposterior analysis will have to be attempted. This can be avoided only by taking extra care beforehand to elicit a sufficiently rich family of models to incorporate all models that you would be willing to consider after looking at the data. Generally speaking, more attention needs to be given to the elicitation of priors. (Frequentists who object to the 'guess-work' involved in obtaining priors should perhaps reflect that they also must 'hazard a guess' at a model. Nothing is entirely objective. Thus statisticians should be willing to go back to adjust initial judgments if that seems sensible in the light of subsequent analysis.)

Finally, as noted above, Bayesian model averaging does not lead to a simple model. This may not matter for forecasting purposes but does matter for description and interpretation. In this regard the model expansion approach advocated by

Draper (1995)—find a good model and expand around it—and Madigan and Raftery's (1994) Occam's window—find a set of parsimonious models which are well supported by the data and average over them—may be preferable to Madigan and York's (1995) Markov chain Monte Carlo model composition approach which averages over all models.

6. COLLECTING MORE DATA

Somewhat belatedly, we turn to component (d) of the model building process as outlined in Section 1. The editor's introduction to Dijkstra (1988) concludes provocatively by saying 'model uncertainty cannot be ignored but is impossible to take into account without new data'. This is rather defeatist and an overstatement but does point us in a possible new direction.

The idea of taking one or more confirmatory samples is a basic feature of the hard sciences, whereas statisticians seem to be primarily concerned (some might say obsessed) with 'squeezing a single data set dry'. Of course it is not always possible to collect more data. For example, in time series analysis, one can rarely obtain more data (except by waiting for several time periods). And some scientific experiments are so costly that it is right to derive as much information out of the data as possible. However, in many other situations it *is* possible to collect more data and this generally seems wise.

The one area of statistics where confirmatory samples are the accepted norm is in clinical trials, though even here it is not always clear how to combine information from different studies. The term *meta-analysis* (e.g. Mosteller and Chalmers (1992) and Draper *et al.* (1992)) has been coined to describe the use of statistical techniques to sum up a body of separate (but similar) experiments in a quantitative way. This was originally seen primarily as a way of searching for a combined *P*-value to see whether there is a significant treatment effect but is now seen more as a way of summarizing all the evidence in both a quantitative *and* a qualitative way. For example the summary might say that study A was not conducted properly and that study B gave atypical results for specified reasons, whereas studies C, D and E all point towards a similar form of relationship.

Statisticians sometimes think that they can overcome the need for new data by splitting a sample into two parts—see Section 4.2. However, as noted earlier, this is a poor substitute for true replication and the same sentiment also applies to techniques like cross-validation. 'The only real validation of a statistical analysis, or of any statistical enquiry, is confirmation by independent observations' (Anscombe (1967), p. 6) and so model validation needs to be carried out on a *completely new* set of data. Unfortunately most references tell you only how to *test* a model, and not how to *tune* or *extend* a model. 'The monitoring of working models is a large and relatively unexplored topic' (Gilchrist (1984), p. 457). The literature also says rather little about the *design* of replicated studies (but see Lindsay and Ehrenberg (1993)).

The emphasis in statistical inference on analysing single sets of data and on testing models contrasts with *scientific inference* which typically involves collecting *many* sets of data and establishing a relationship which generalizes to different conditions. In other words scientists look for what Nelder (1986) has called *significant sameness* rather than for significant differences. In a similar vein Ehrenberg and Bound (1993) have promoted the idea of searching for *law-like relationships* which describe, not a single set of data, but many sets of data collected under similar or perhaps even dissimilar conditions. In general a law or relationship is much more useful if it 'works' under different conditions rather than merely under as near identical conditions as possible. The latter are usually possible only in the physical sciences. It is unfortunate that the words 'reproducibility', 'replication' and 'repetition' seem to have no generally accepted definition although replication often refers to repeats made at the same place and time. We are talking here about repeats at different points in space and time. What is clear is that more than one data set is needed before we can have any confidence in a model. For this reason, Feynman (1986) (especially p. 344) is right to lament the attitude that repeating an experiment (under similar and/or carefully varied conditions) is a waste of time and not to be counted as research.

The replication of studies can also be sadly neglected in the social sciences. Hubbard and Armstrong (1994) demonstrated that it is very rare in marketing by examining over 1000 published papers, of which none were straight replications and less than 2% were replications with extensions (and over half of these contradicted the original findings!). There is a similar story in psychology. A replication which confirms earlier results promotes confidence in them, whereas conflicting results may help to avert erroneous recommendations or to suggest the need for further research. Put bluntly, if a result is not worth replicating, then it is not worth knowing! Hubbard and Armstrong (1994) speculated on the reasons why replications are so scarce (e.g. the original paper may not report enough background information to permit accurate replication or conducting replications is not career enhancing) and also on ways of encouraging them (e.g. modify journal policy to ensure that authors *are* required to give sufficient background information, ensure that data are made available for subsequent evaluation and appoint a replications editor).

The (over?) emphasis on analysing single sets of data permeates the statistical literature and is a serious disease of statistical teaching. Of course research on model uncertainty for the analysis of a single data set, as in Miller (1990), is clearly valuable, both to cover situations where it is not possible to collect further data and also to understand techniques, like subset selection, which are widely used in practice. However, Miller (1990), p. 13, follows accepted dogma in devoting just a single sentence to the possibility of taking an independent sample to test the adequacy of a prediction equation. In contrast the message of this section is to emphasize that obtaining more than one set of data, whenever possible, is a potentially more convincing way of overcoming model uncertainty and is needed anyway to determine the range of conditions under which a model is valid. Thus statisticians need to achieve better balance between

(a) statistical inference for a single set of data (with or without a prespecified model) and
(b) understanding how to build, check, tune and extend models when it is possible (and therefore desirable) to collect more than one set of data.

7. SUMMARIZING REMARKS AND DISCUSSION

The theory of inference regarding parameter estimation generally assumes that the true model for a given set of data is known and prespecified. In practice a model may be formulated from the data, and it is increasingly common for tens or even hundreds of possible models to be entertained (data mining). A single model is usually selected as the 'winner' even when other models give nearly as good a fit. Even when a model is prespecified on subject-matter grounds, it may be formulated incorrectly, or a true model may not exist, or the analyst may carry out some preliminary checks anyway. Thus *model uncertainty* is present in most real problems. Yet statisticians have given the topic little attention.

Least squares theory is known not to apply when the same data are used to formulate *and* fit a model so that *estimation follows model selection*. Substantial model selection biases can arise, particularly with subset selection methods in multiple regression and in time series analysis. Unfortunately ways of overcoming the problem are not so clear. Statistical inference needs to be broadened to include model formulation, but it is not clear to what extent we can formalize the steps taken by an experienced analyst during data analysis and model building, and whether a suitable mathematical framework can be constructed. Some valiant simulation and resampling experiments have been carried out to try to assess the size of model selection biases and to find ways of overcoming them. However, the frequentist approach does not adapt naturally to cope with model uncertainty. The Bayesian model averaging approach offers more promise, though even here there are difficulties. A safer way to proceed is to replicate the study and to check the fit of the model on new data. However, this is not always possible, especially in time series analysis. Thus the task of finding ways to overcome model uncertainty has only just begun.

Perhaps the main message of this paper is that it is time for statisticians to stop pretending that model uncertainty does not exist, and to give due regard to the computer-based revolution in model formulation which has taken place. This applies, not only to statistical practice, but also to what we teach.

Leamer (1978) set out to bridge the gap between econometric theorists and model builders but ended up less optimistic that a complete reconciliation could be achieved. He predicted (p. vi) that 'real inference will remain a highly complicated, poorly understood phenomenon', and I would still agree with that today.

ACKNOWLEDGEMENTS

Special thank go to David Draper and Urban Hjorth for their wise counsel and valuable comments on earlier drafts. I also thank Andrew Ehrenberg, David Madigan, B. M. Pötscher and a referee, among others, for helpful constructive advice. The computations for example 4 were carried out by Michael Dymond as part of a Master of Science dissertation at Bath University.

REFERENCES

Adams, J. L. (1991) A computer experiment to evaluate regression strategies. *Proc. Am. Statist. Ass. Sect. Statist. Comput.*, 55–62.

Ahn, S. K. (1993) Some tests for unit roots in autoregressive-integrated-moving average models with deterministic trends. *Biometrika*, **80**, 855–868.

Anscombe, F. J. (1967) Topics in the investigation of linear relations fitted by the method of least squares (with discussion). *J. R. Statist. Soc.* B, **29**, 1–52.

Ansley, C. F. and Newbold, P. (1980) Finite sample properties of estimators for autoregressive moving average models. *J. Econometr.*, **13**, 159–183.

Azzalini, A. and Cox, D. R. (1984) Two new tests associated with analysis of variance. *J. R. Statist. Soc.* B, **46**, 335–343.

Bhansali, R. J. (1981) Effects of not knowing the order of an autoregressive process on the mean squared error of prediction—1. *J. Am. Statist. Ass.*, **76**, 588–597.

Box, G. E. P. (1976) Science and statistics. *J. Am. Statist. Ass.*, **71**, 791–799.

———(1980) Sampling and Bayes' inference in scientific modelling and robustness (with discussion). *J. R. Statist. Soc.* A, **143**, 383–430.

———(1990) Commentary on a paper by Hoadley and Kettenring. *Technometrics*, **32**, 251–252.

———(1993) Quality improvement—the new industrial revolution. *Int. Statist. Rev.*, **61**, 3–19.

———(1994) Statistics and quality improvement. *J. R. Statist. Soc.* A, **157**, 209–229.

Box, G. E. P., Jenkins, G. M. and Reinsel, G. C. (1994) *Time Series Analysis, Forecasting and Control*, 3rd edn. Englewood Cliffs: Prentice Hall.

Breiman, L. (1992) The little bootstrap and other methods for dimensionality selection in regression: X-fixed prediction error. *J. Am. Statist. Ass.*, **87**, 738–754.

Chambers, J. M. (1993) Greater or lesser statistics: a choice for future research. *Statist. Comput.*, **3**, 182–184.

Chatfield, C. (1993a) Neural networks: forecasting breakthrough or passing fad? *Int. J. Forecast.*, **9**, 1–3.

———(1993b) Calculating interval forecasts (with discussion). *J. Bus. Econ. Statist.*, **11**, 121–144.

———(1995) *Problem Solving: a Statistician's Guide*, 2nd edn. London: Chapman and Hall.

Choi, B. (1992) *ARMA Model Identification*. New York: Springer.

Clemen, R. T. (1989) Combining forecasts: a review and annotated bibliography. *Int. J. Forecast.*, **5**, 559–583.

Cohen, A. and Sackrowitz, H. B. (1987) An approach to inference following model selection with applications to transformation-based and adaptive inference. *J. Am. Statist. Ass.*, **82**, 1123–1130.

Copas, J. B. (1983) Regression, prediction and shrinkage (with discussion). *J. R. Statist. Soc.* B, **45**, 311–354.

———(1993) The shrinkage of point scoring methods. *Appl. Statist.*, **42**, 315–331.

Copas, J. B. and Jones, M. C. (1987) Regression shrinkage methods and autoregressive time series prediction. *Aust. J. Statist.*, **29**, 264–277.

Cox, D. R. and Hinkley, D. V. (1974) *Theoretical Statistics*. London: Chapman and Hall.

Cox, D. R. and Snell, E. J. (1981) *Applied Statistics*. London: Chapman and Hall.

Dawid, A. P. and Dickey, J. M. (1977) Likelihood and Bayesian inference from selectively reported data. *J. Am. Statist. Ass.*, **72**, 845–853.

Dijkstra, T. K. (ed.) (1988) *On Model Uncertainty and Its Statistical Implications*. Berlin: Springer.

Draper, D. (1995) Assessment and propagation of model uncertainty (with discussion). *J. R. Statist. Soc.* B, **57**, 45–97.

Draper, D., Hodges, J. S., Leamer, E. E., Morris, C. N. and Rubin, D. B. (1987) A research agenda for assessment and propagation of model uncertainty. *Report N-2683-RC*. Rand Corporation, Santa Monica.

Draper, D. *et al.* (1992) *Combining Information*. Washington DC: National Academy Press.

Efron, B. and Tibshirani, R. J. (1993) *An Introduction to the Bootstrap*. New York: Chapman and Hall.

Ehrenberg, A. S. C. and Bound, J. A. (1993) Predictability and prediction (with discussion). *J. R. Statist. Soc.* A, **156**, 167–206.

Faraway, J. J. (1992) On the cost of data analysis. *J. Comput. Graph. Statist.*, **1**, 213–229.

Feynman, R. P. (1986) *Surely You're Joking Mr. Feynman*. London: Unwin.

Fildes, R. and Howell, S. (1979) On selecting a forecasting model. *TIMS Stud. Mangmnt Sci.*, **12**, 297–312.

Fildes, R. and Makridakis, S. (1994) The impact of empirical accuracy studies on time series analysis and forecasting. *Discussion Paper*. Management School, Lancaster University, Lancaster.

Freedman, D. A., Navidi, W. and Peters, S. C. (1988) On the impact of variable selection in fitting regression equations. In *On Model Uncertainty and Its Statistical Implications* (ed. T. K. Dijkstra), pp. 1–16. Berlin: Springer.

Gardner, Jr, E. S. and McKenzie, E. (1985) Forecasting trends in time series. *Mangmnt Sci.*, **31**, 1237–1246.

Geisser, S. (1993) *Predictive Inference: an Introduction*. New York: Chapman and Hall.

Gilchrist, W. (1984) *Statistical Modelling*. Chichester: Wiley.

de Gooijer, J. G. (1985) A Monte Carlo study of the small-sample properties of some estimators for ARMA models. *Comput. Statist. Q.*, **3**, 245–266.

de Gooijer, J. G., Abraham, B., Gould, A. and Robinson, L. (1985) Methods for determining the order of an autoregressive-moving average process: a survey. *Int. Statist. Rev.*, **53**, 301–329.

Hahn, G. J. and Meeker, W. Q. (1993) Assumptions for statistical inference. *Am. Statistn*, **47**, 1–11.

Hill, R. C., Cartwright, P. A. and Arbaugh, J. F. (1991) The use of biased predictors in marketing research. *Int. J. Forecast.*, **7**, 271–282.

Hirsch, R. P. (1991) Letter to the editor. *Biometrics*, **47**, 1193–1194.

Hjorth, U. (1982) Model selection and forward validation. *Scand. J. Statist.*, **9**, 95–105.

——(1987) On model selection in the computer age. *Technical Report LiTH-MAT-R-87-08*. Linköping University, Linköping.

——(1989) On model selection in the computer age. *J. Statist. Planng Inf.*, **23**, 101–115.

——(1990) Model selection needs resampling methods. *Technical Report LiTH-MAT-R-1990-12*. Linköping University, Linköping.

——(1994) *Computer Intensive Statistical Methods—Validation Model Selection and Bootstrap*. London: Chapman and Hall.

Hjorth, U. and Holmqvist, L. (1981) On model selection based on validation with applications to pressure and temperature prognosis. *Appl. Statist.*, **30**, 264–274.

Hodges, J. S. (1987) Uncertainty, policy analysis and statistics. *Statist. Sci.*, **2**, 259–291.

Hubbard, R. and Armstrong, J. S. (1994) Replications and extensions in marketing: rarely published but quite contrary. *Int. J. Res. Marktng*, **11**, 233–248.

Hurvich, C. M. and Tsai, C.-L. (1990) The impact of model selection on inference in linear regression. *Am. Statistn*, **44**, 214–217.

Judge, G. G. and Bock, M. E. (1978) *The Statistical Implications of Pre-test and Stein-rule Estimators in Econometrics*. Amsterdam: North-Holland.

Kass, R. E. and Raftery, A. E. (1995) Bayes factors. *J. Am. Statist. Ass.*, **90**, in the press.

Kendall, M., Stuart, A. and Ord, J. K. (1983) *The Advanced Theory of Statistics*, vol. 3, 4th edn. London: Griffin.

Kipnis, V. (1991) Evaluating the impact of exploratory procedures in regression prediction: a pseudosample approach. *Comput. Statist. Data Anal.*, **12**, 39–55.

Le, N. D., Raftery, A. E. and Martin, R. D. (1993) Robust model comparison for autoregressive processes with robust Bayes factors. *Technical Report 123*. Department of Statistics, University of British Columbia, Vancouver.

Leamer, E. E. (1978) *Specification Searches: ad hoc Inference with Experimental Data*. New York: Wiley.

——(1985) Sensitivity analyses would help. *Am. Econ. Rev.*, **75**, 308–313.

——(1992) Testing trade theory. *Working Paper 3957*. Cambridge: National Bureau of Economic Research.

de Leeuw, J. (1988) Model selection in multinomial experiments. In *On Model Uncertainty and Its Statistical Implications* (ed. T. K. Dijkstra), pp. 118–138. Berlin: Springer.

Leybourne, S. J. and McCabe, B. P. M. (1994) A consistent test for a unit root. *J. Bus. Econ. Statist.*, **12**, 157–166.

Lindsay, R. M. and Ehrenberg, A. S. C. (1993) The design of replicated studies. *Am. Statistn*, **47**, 217–228.

Lovell, M. C. (1983) Data mining. *Rev. Econ. Statist.*, **65**, 1–12.

Madigan, D. and Raftery, A. E. (1994) Model selection and accounting for model uncertainty in graphical models using Occam's window. *J. Am. Statist. Ass.*, **89**, 1535–1546.

Madigan, D. and York, J. (1993) Bayesian graphical models for discrete data. *Int. Statist. Rev.*, **63**, in the press.

Miller, A. J. (1990) *Subset Selection in Regression*. London: Chapman and Hall.

Mosteller, F. and Chalmers, T. C. (1992) Some progress and problems in meta-analysis of clinical trials. *Statist. Sci.*, **7**, 227–236.

Nelder, J. A. (1986) Statistics, science and technology. *J. R. Statist. Soc.* A, **149**, 109–121.

Newbold, P., Agiakloglou, C. and Miller, J. (1993) Long-term inference based on short-term forecasting models. In *Developments in Time-series Analysis* (ed. T. Subba Rao), pp. 9–25. Lodon: Chapman and Hall.

Phillips, G. D. A. and McCabe, B. P. M. (1989) A sequential approach to testing for structural change in econometric models. *Emp. Econometr.*, **14**, 151–165.

Picard, R. R. and Cook, R. D. (1984) Cross-validation of regression models. *J. Am. Statist. Ass.*, **79**, 575–583.

Popper, K. R. (1959) *The Logic of Scientific Discovery* (Engl. transl.). London: Hutchinson.

Poskitt, D. S. and Tremayne, A. R. (1987) Determining a portfolio of linear time series models. *Biometrika*, **74**, 125–137.

Pötscher, B. M. (1991a) Effects of model selection on inference. *Econometr. Theory*, **7**, 163–185.

——(1991b) Correspondence. *Am. Statistn*, **45**, 171–172.

Regal, R. R. and Hook, E. B. (1991) The effects of model selection on confidence intervals for the size of a closed population. *Statist. Med.*, **10**, 717–721.

Rencher, A. C. and Pun, F. C. (1980) Inflation of R^2 in best subset regression. *Technometrics*, **22**, 49–53.

Schervish, M. J. and Tsay, R. S. (1988) Bayesian modeling and forecasting in autoregressive models. In *Bayesian Analysis of Time Series and Dynamic Models* (ed. J. C. Spall), pp. 23-52. New York: Dekker.

Schoemaker, P. J. H. (1991) When and how to use scenario planning: a heuristic approach with illustrations. *J. Forecast.*, **10**, 549–564.

Shaman, P. and Stine, R. A. (1988) The bias of autoregressive coefficient estimators. *J. Am. Statist. Ass.*, **83**, 842–848.

Shibata, R. (1976) Selection of the order of an autoregressive model by Akaike's information criterion. *Biometrika*, **63**, 117–126.

Silvey, S. D. (1970) *Statistical Inference*. Harmondsworth: Penguin.

Sprent, P. (1993) *Applied Nonparametric Methods*, 2nd edn. London: Chapman and Hall.

Steerneman, T. and Rorijs, G. (1988) Pitfalls for forecasters. In *On Model Uncertainty and Its Statistical Implications* (ed. T. K. Dijksta), pp. 102–117. Berlin: Springer.

Tiao, G. C. and Tsay, R. S. (1994) Some advances in non-linear and adaptive modelling in time-series. *J. Forecast.*, **13**, 109–131.

Tsay, R. S. (1993) Comment on a paper by Chatfield. *J. Bus. Econ. Statist.*, **11**, 140–142.

Tukey, J. W. (1991) Use of many covariates in clinical trials. *Int. Statist. Rev.*, **59**, 123–137.

——(1994) More honest foundations for data analysis. *American Statistical Association Meet., Toronto.*

West, M. and Harrison, P. J. (1989) *Bayesian Forecasting and Dynamic Linear Models*. New York: Springer.

Wild, C. J. (1994) Embracing the "wider view" of statistics. *Am. Statistn*, **48**, 163–171.

Zhang, P. (1992) Inference after variable selection in linear regression models. *Biometrika*, **79**, 741–746.

DISCUSSION OF THE PAPER BY CHATFIELD

J. B. Copas (University of Warwick, Coventry): This paper raises a very important issue in the practice of statistics. As we have come to expect of Dr Chatfield, his paper is full of sound common sense and is persuasively argued with his customary clarity and style. He would be the first to admit that there is little that is new in the paper, but he does us a service by calling us all to task over what has been called a 'scandal'. The message of the paper is summed up in the last sentence of Section 1: 'Statisticians must stop pretending that model uncertainty does not exist and begin to find ways of coping with it'.

The paper raises the question of whether a model exists. Surely we have to make the crucial distinction between experimental data and observational data. In properly designed experiments a null model *does* exist and is simply a description of the randomization used in the design. We should remember that many of the traditional statistical techniques were originally developed for experimental data. Questions about the modelling of experimental data and the validity of the usual analyses were extensively discussed in literature predating all the references in this paper. In his book *The Design of Experiments*, for example, Fisher (1966) (but first edition 1935) discussed the simple matched pairs experiment in which for each pair the treatment order (A, B) or (B, A) is decided by the toss of a fair coin. If the data for a typical pair are (x, y) then the test statistic is

$$\sum Z(x-y)$$

where Z is $+1$ for (A, B) and -1 for (B, A). If the treatment has no effect then the xs and ys would be the same whichever treatment orders were chosen and so are known constants. Hence the test statistic has a known null distribution and, as Fisher showed, gives almost exactly the same P-values as the usual