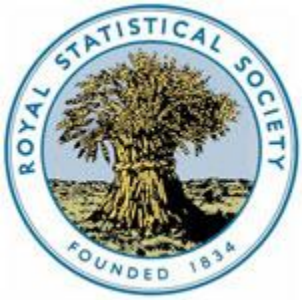

WILEY

Model Uncertainty, Data Mining and Statistical Inference
Author(s): Chris Chatfield
Source: *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, Vol. 158, No. 3 (1995), pp. 419-466
Published by: Wiley for the Royal Statistical Society
Stable URL: http://www.jstor.org/stable/2983440
Accessed: 02-04-2018 20:12 UTC

JSTOR is a not-for-profit service that helps scholars, researchers, and students discover, use, and build upon a wide range of content in a trusted digital archive. We use information technology and tools to increase productivity and facilitate new forms of scholarship. For more information about JSTOR, please contact support@jstor.org.

Your use of the JSTOR archive indicates your acceptance of the Terms & Conditions of Use, available at http://about.jstor.org/terms

Royal Statistical Society, *Wiley* are collaborating with JSTOR to digitize, preserve and extend access to *Journal of the Royal Statistical Society. Series A (Statistics in Society)*

This content downloaded from 98.161.230.204 on Mon, 02 Apr 2018 20:12:13 UTC
All use subject to http://about.jstor.org/terms

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

©1995 Royal Statistical Society 0035-9238/95/158419

This content downloaded from 98.161.230.204 on Mon, 02 Apr 2018 20:12:13 UTC
All use subject to http://about.jstor.org/terms

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

This content downloaded from 98.161.230.204 on Mon, 02 Apr 2018 20:12:13 UTC
All use subject to http://about.jstor.org/terms

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

This content downloaded from 98.161.230.204 on Mon, 02 Apr 2018 20:12:13 UTC
All use subject to http://about.jstor.org/terms

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

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

This content downloaded from 98.161.230.204 on Mon, 02 Apr 2018 20:12:13 UTC
All use subject to http://about.jstor.org/terms

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

This content downloaded from 98.161.230.204 on Mon, 02 Apr 2018 20:12:13 UTC
All use subject to http://about.jstor.org/terms

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

This content downloaded from 98.161.230.204 on Mon, 02 Apr 2018 20:12:13 UTC
All use subject to http://about.jstor.org/terms

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

This content downloaded from 98.161.230.204 on Mon, 02 Apr 2018 20:12:13 UTC
All use subject to http://about.jstor.org/terms

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

This content downloaded from 98.161.230.204 on Mon, 02 Apr 2018 20:12:13 UTC
All use subject to http://about.jstor.org/terms

(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!

This content downloaded from 98.161.230.204 on Mon, 02 Apr 2018 20:12:13 UTC
All use subject to http://about.jstor.org/terms

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

This content downloaded from 98.161.230.204 on Mon, 02 Apr 2018 20:12:13 UTC
All use subject to http://about.jstor.org/terms

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

This content downloaded from 98.161.230.204 on Mon, 02 Apr 2018 20:12:13 UTC
All use subject to http://about.jstor.org/terms

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

This content downloaded from 98.161.230.204 on Mon, 02 Apr 2018 20:12:13 UTC
All use subject to http://about.jstor.org/terms

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

This content downloaded from 98.161.230.204 on Mon, 02 Apr 2018 20:12:13 UTC
All use subject to http://about.jstor.org/terms

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

This content downloaded from 98.161.230.204 on Mon, 02 Apr 2018 20:12:13 UTC
All use subject to http://about.jstor.org/terms

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

This content downloaded from 98.161.230.204 on Mon, 02 Apr 2018 20:12:13 UTC
All use subject to http://about.jstor.org/terms

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

This content downloaded from 98.161.230.204 on Mon, 02 Apr 2018 20:12:13 UTC
All use subject to http://about.jstor.org/terms

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.*

This content downloaded from 98.161.230.204 on Mon, 02 Apr 2018 20:12:13 UTC
All use subject to http://about.jstor.org/terms

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.

This content downloaded from 98.161.230.204 on Mon, 02 Apr 2018 20:12:13 UTC
All use subject to http://about.jstor.org/terms

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

$$\hat{y}_c = p_1(\alpha_1 + \beta x_0) + p_2\alpha_2$$
$$= p_1\alpha_1 + p_2\alpha_2 + p_1\beta x_0. \quad (1)$$

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

This content downloaded from 98.161.230.204 on Mon, 02 Apr 2018 20:12:13 UTC
All use subject to http://about.jstor.org/terms

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

This content downloaded from 98.161.230.204 on Mon, 02 Apr 2018 20:12:13 UTC
All use subject to http://about.jstor.org/terms

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,

This content downloaded from 98.161.230.204 on Mon, 02 Apr 2018 20:12:13 UTC
All use subject to http://about.jstor.org/terms

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

This content downloaded from 98.161.230.204 on Mon, 02 Apr 2018 20:12:13 UTC
All use subject to http://about.jstor.org/terms

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.

This content downloaded from 98.161.230.204 on Mon, 02 Apr 2018 20:12:13 UTC
All use subject to http://about.jstor.org/terms

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.

This content downloaded from 98.161.230.204 on Mon, 02 Apr 2018 20:12:13 UTC
All use subject to http://about.jstor.org/terms

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 pseudo-sample 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.

This content downloaded from 98.161.230.204 on Mon, 02 Apr 2018 20:12:13 UTC
All use subject to http://about.jstor.org/terms

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

This content downloaded from 98.161.230.204 on Mon, 02 Apr 2018 20:12:13 UTC
All use subject to http://about.jstor.org/terms

t-test. Fisher would, I suspect, argue that all the later discussion of testing distributional assumptions, checking for outliers, and so on, is largely irrelevant—the null model is self-evident from the design and is well approximated by the usual model based on normality. Note that the coin tossing is the crux of the model. Even a small amount of bias in the design (perhaps the probability of (A, B) depends on some unmeasured characteristic of each pair) leads to a substantial bias in the t-test. If nothing is known about the way that the treatments were allocated, then nothing can be deduced from the data.

For observational data the situation is quite different. The paper seems to suggest that if we were sufficiently clever to calculate the right conditional probabilities or make the right corrections for shrinkage or subset selection then we could overcome the problem of model uncertainty. For calculating P-values, for example, Section 3 suggests that we need

$$P(\text{data more extreme} | \text{data select this model}).$$

But surely the most important thing in this expression is not the conditioning event but the capital P at the beginning: we must say that the data are *as if* they had resulted from some game of chance in which probabilities are defined. For observational data this must be an assumption, surely unverifiable on the basis of the data alone. Can we observe the world as it is and know for certain whether everything was or was not the result of some gigantic sequence of coin tosses? So we cannot eliminate the effect of model uncertainty, only evaluate one layer of uncertainty by embedding it within the assumption of another. As Dr Chatfield points out, standard errors escalate rapidly as this embedding model becomes more general, and if our inferences are to be useful then we must have further information, either from more data or from knowledge of the context. If we really know nothing about the problem beyond the data themselves then perhaps we should emphasize the descriptive rather than the inferential nature of our analysis and refuse to quote any standard errors or other measures of uncertainty.

When the Social Science Research Council was set up about 25 years ago, it appointed a statistics committee to advise on the statistical aspects of research grant applications. Frequently a project in psychology or sociology would be declared methodologically unsound and so would not be funded. Astute psychologists and sociologists soon learnt to disguise the statistics within their research so that their proposals would not be sent to the statistics committee. The word 'Science' has now been dropped from the title of the now Economic and Social Research Council, and the statistics committee no longer exists. Dr Chatfield tells us that much of *our* statistics is methodologically unsound. If we want to survive as honest analysers of observational data then we should take his paper very seriously.

Section 3 declares that we should not seek a 'significantly better fit', but 'better value for money'. Is this a 'back to basics' campaign? If so the proponents must beware lest their own personal activities are found wanting. Most of us here tonight will have committed many of the sins which Dr Chatfield lists at length in his paper. If Dr Chatfield wishes to make his personal confession in his reply to the discussion, we will listen with interest. In the meantime I have pleasure in proposing the vote of thanks for this challenging and provocative paper.

Neville Davies (Nottingham Trent University): There is a tradition in the Society that proposers and seconders of votes of thanks of read papers are both polite and rude, and often unequivocal in their comments about those papers. Frequently they use it as an excuse to announce the latest aspect of their own research in a totally unrelated area, and tenuous links are made to justify this. I see no reason to depart too far from this tradition: I will not be rude at all about this paper!

Those of us who remember, and perhaps have even read, other papers published in the Society's journals by the author might not only recall the content of those papers but also the lively and provocative discussions that followed. It is fair to say that, even though you might not agree with him, he has always had the ability to generate debate.

I am in agreement with many of the points raised by this paper, but I have no doubt that many colleagues will disagree with Chatfield's approach: I predict that some may even believe the paper trivial and some will say that they knew it all in the first place. I believe that the greatest unsolved problem in statistics is communicating the subject to others. This paper is one which does an excellent job of communicating problems that we all should think about: we all fit models, but most of us do exactly what Chatfield chides us about. Unfortunately, we are not presented with many solutions in this paper.

I shall apply a simple-minded approach and discuss what we would need to do to 'model' this paper. I do this to demonstrate that model uncertainty may change, depending on the use to which a model is put. As Dr Chatfield points out (Section 1) we need to decide what problem needs to be solved, and

This content downloaded from 98.161.230.204 on Mon, 02 Apr 2018 20:12:13 UTC
All use subject to http://about.jstor.org/terms

what data need to be collected to solve that problem. In many cases data have already been collected for us, and this is certainly true if we regard this paper as our data. Let us suppose that we are interested in solving one of the following:

(a) identifying who wrote this paper;
(b) capturing this paper's content and style;
(c) using this paper to predict features of future papers.

The first of these problems is simple to solve, since we can just read the name at the top of the preprint! That is the only information (data) that we have used from the paper. The model is thus very simple but also general: this is a 'Chris Chatfield paper'. We appear to have absolute certainty about the model. However, if the first page were missing how could we then be certain that what we have is a Chatfield paper? This must be connected with obtaining a solution to problem (b), since we might expect at least the style of this paper to be similar to previous, and perhaps even future, papers written by Dr Chatfield. We have now created more problems: we shall have to decide *how* to characterize the content and/or style and to *check* these features by collecting more data, as advocated in Section 6. We need to read more papers authored by Chatfield and validate the chosen feature(s) that we observe in the present paper.

I have selected three papers previously published in journals of the Society (Chatfield, 1977, 1982, 1985). What feature(s) can we choose that will enable problems (a) and (b) to be solved? In Table 1 I present the number of times on each page of the journal that the author used single quotes either to emphasize a word or words, or when a direct quote from other people is presented. I have excluded pages that contain references. The last column of the table is the number of occasions that quotes are used in the present paper. In three of the four papers, it is interesting that there are some pages with a very high number of quotes.

TABLE 1
Page count of quotes

Page	*No. of counts for the following years of paper:*			
	1977	*1982*	*1985*	*1995*
1	0	4	2	2
2	2	1	5	3
3	6	3	2	6
4	7	5	5	2
5	3	4	7	5
6	3	3	8	0
7	5	0	6	0
8	2	4	1	1
9	10	2	4	2
10	1	1	3	10
11	4	3	6	3
12	3	0	0	5
13	0	0	3	1
14	4	0	6	6
15	1	1	10	1
16	1	–	3	0
17	–	–	3	0
18	–	–	–	3
19	–	–	–	2
20	–	–	–	2
21	–	–	–	4
22	–	–	–	4
23	–	–	–	1
Mean	3.3	2.1	4.4	2.7
Variance	7.4	3.2	6.9	6.0

This content downloaded from 98.161.230.204 on Mon, 02 Apr 2018 20:12:13 UTC
All use subject to http://about.jstor.org/terms

One question is, 'To what extent does a simple count of occurrences of quotes characterize Dr Chatfield's style of writing?'. To do this we might need to model within- and between-sample variability for these papers, and this might involve specifying a proper statistical model. Naturally, a more thorough approach would be to investigate papers written by other authors and to investigate whether my simple characterization is adequate to identify a style uniquely. More effort is needed, even though our objective is the same. To solve problem (c) is much more difficult. We would need to decide exactly what features were to be forecasted before specifying a stochastic model. For example, a simple Poisson model may be appropriate to predict page frequency of quotes. However, model uncertainty (misspecification) in terms of overdispersion could be an issue. My point is that model uncertainty will be different, even with the same model, depending how that model is to be utilized. Other information would be needed to build a 'better' model to be able to characterize and predict other features.

I agree whole-heartedly with the lessons to be learned from examples 2 and 3 in Section 2 and I am attracted by the Bayesian model averaging approach given in example 6 in Section 5. The problem that I see in example 6 is the premise that either model I or II is appropriate. The combination is convenient and plausible, but model I with a coefficient, β, that is state dependent could capture the uncertainty instead of having two models that may approximate the data. This would be similar to locally changing parameters in time series. As Chatfield has argued elsewhere (Chatfield, 1993) it is very important to attach prediction intervals to point forecasts. The Bayesian model averaging approach that yields the weighted prediction (1) has a lower mean-square prediction error than either of the individual forecasts. But what can be said about the corresponding prediction invervals for the combined forecast in this case? The author comments about null distributions of R^2. It may be of interest to him that the exact mean of this statistic is given by Smith (1993).

Several times the author refers to time series models and particular difficulties associated with them. All the models that are used for illustration assume time constant parameters. He does, however, allude to the possibility of a changing structure in discussing testing for unit roots (Section 2.5). An evolving system, which statisticians may attempt (mistakenly) to model by using constant parameters, seems to me to be far more natural. I believe that parameter flexibility can compensate for model uncertainty. Although the author makes passing reference to West and Harrison (1989) in the context of multiprocess models, I was surprised that the whole philosophy behind the dynamic linear model (DLM) was not commented on in more detail. The DLM approach to modelling time series data is flexible and many examples of its implementation can be found in Pole *et al.* (1994). The DLM learns as it evolves through the data, but the learning is not just restricted to information in the data. Model uncertainty in the context of this paper may equate to lack of knowledge: presuming that model parameters are the same at the beginning as at the end of a time series should not be part of the null model.

Perhaps the most general formulation for time series models is provided by the state-dependent models (SDMs) of Priestley (1980). Almost all time series models used by practitioners are special cases of this class of models. Why do more time series analysts not start with this class, and would this not help to lessen model uncertainty? Draper (1995), section 5.1, argues in favour of the 'big' model approach, but it has practical difficulties. One problem is, of course, that the glorious generality afforded by SDMs is impractical with most real data sets. Unfortunately, there seem to be very few reported applications of these models.

Now for a not-so-tenuous link with my own research! Chatfield (Section 2.5) discusses the problem of determining whether differencing is needed to reduce a series to stationarity. He states '. . . we shall never know whether a unit root really exists, or whether such a structure is appropriate for part of the series . . .'. 'Never' is a very long time! Considerable controversy surrounds the issue of testing for unit roots, since many of the original tests for this property have been shown to have limitations. See, for example, Agiakloglou and Newbold (1992) and Leybourne and McCabe (1994). A simple statistic, the variogram, that has been used extensively in spatial statistics, can be employed to good effect to investigate non-stationary time series data. Box and Kramer (1992) and Box (1994) suggested the use of a standardized form of the variogram of a time series Y_t ($t=1, 2, \ldots, n$), defined by

$$G(m) = V(Y_{t+m} - Y_t)/V(Y_{t+1} - Y_t),$$

where $V(\ )$ is the variance operator, as a general measure of non-stationarity. Now consider the non-stationary random coefficient autoregressive (RCAR) model $W_t = \alpha_t W_{t-1} + \epsilon_t$, where $W_t = Y_t - Y_0$, $\epsilon_t \sim \text{IID}(0, \sigma_\epsilon^2)$ and $\alpha_t \sim \text{IID}(\alpha, \omega^2)$. Davies and Tremayne (1994) have shown that for this simple RCAR model

This content downloaded from 98.161.230.204 on Mon, 02 Apr 2018 20:12:13 UTC
All use subject to http://about.jstor.org/terms

$$G(m) = \frac{(\alpha^2+\omega^2)^t\{(\alpha^2+\omega^2)^m + (1-2\alpha^m)\} + 2(\alpha^m - 1)}{(\alpha^2+\omega^2)^t(\alpha^2+\omega^2+1-2\alpha) + 2(\alpha-1)} . \quad (2)$$

The case $\alpha = 1$ is of interest, since this corresponds to a process with a stochastic unit root. In this case $G(m) = \{(\omega^2+1)^m - 1\}/\omega^2$, and Davies and Tremayne show that estimates of $G(m)$ for RCAR processes follow closely the template provided by these formulae. Using plots of the variogram, there is some evidence to suggest that the degree of differencing required for series is not constant throughout those series. Examples include the IBM series of Box and Jenkins (1976) and the Hong Kong Hang Seng index.

A much stronger link with past research of mine is evident from the work of Davies and Newbold (1980). They provided a framework under which the cost of misspecification, in terms of increased expected squared error of prediction, can be assessed. They concentrated on forecasting ARIMA(p, d, q) processes with pure autoregressive models, taking estimation error into account. The measure of forecast loss presented in that paper has been used by Ray (1993) for assessing the percentage increase in mean-squared forecast error when long memory (fractionally differenced) processes are modelled by using pure autoregressive structures. It may well be fruitful to develop further the measure for a broader range of time series model misspecifications.

Finally, I believe that the author has given us some food for thought and it gives me great pleasure to second the vote of thanks.

The vote of thanks was passed by acclamation.

D. J. Hand (The Open University, Milton Keynes): I am very pleased to see this paper. It represents the latest in a series of publications in the *Journal of the Royal Statistical Society* dealing with broader issues of statistical inference. Others have dealt with such things as study replication (Ehrenberg and Bound, 1993), problem formulation (Hand, 1994) and model uncertainty (Draper, 1995).

However, such papers are still very much in the minority. A glance at almost any modern statistical journal will demonstrate the current gross imbalance in favour of papers dealing with relatively small refinements of modelling. This is the middle of the three layers of a study that Dr Chatfield identifies. The first and last layers, respectively 'problem' and 'data', receive very little attention, despite the fact that the intrinsic variation in these often leads to variation in the conclusions which swamps that arising from the variation in model building.

This is, of course, a consequence of the *success* of statistics. It has achieved this success by putting a premium on theoretical development of statistical modelling and estimation procedures. However, it seems to me that now the most significant improvements will arise by shifting attention to the first and third layers: by stepping outside the bounds of work focusing on model fitting and model checking. So I agree with the author on this point.

However, I think that the author's focus on *models* is misdirected.

We must distinguish between two ways in which statisticians use the word 'model'. One usage is a *representation* of what is thought to be going on. This might, for example, be a system of equations based on some theory. The other usage is a *description* of the data. In the former case there will be relatively little model uncertainty. It is in the latter case where problems arise. In this case, there is great scope for selecting models—and consequently for model uncertainty. But, in this case, models are often, perhaps typically, merely a means to a *predictive* end. Statisticians, in such cases, seem to me to have become seduced by the attractions of models. The *problem*—prediction or whatever—needs to be returned to centre stage. The attractive model averaging approach, which has been increasingly widely applied in the last few years, is a step in the right direction: a step towards replacing the problem at centre stage, with the models merely being paths leading to it.

Clifford E. Lunneborg (University of Washington, Seattle, and The Open University, Milton Keynes): Chris Chatfield reminds us of how great the need is to introduce more reality and usefulness into the inferences which we draw from data-dependent models. We might do so by attending to two principles:

(a) take account of the noise created by the modelling process;
(b) consistent with the modelling goal, create as little noise as possible.

Can we do it? I am more optimistic, particularly with respect to frequentist inference based on resampling, than Chatfield appears to be.

This content downloaded from 98.161.230.204 on Mon, 02 Apr 2018 20:12:13 UTC
All use subject to http://about.jstor.org/terms

Chatfield's example 1

Chatfield describes a data-dependent strategy for estimating a mean, points out that any subsequent inferences should be conditional on that strategy but finds no help for doing so in the statistical literature. Although the location estimators used in their example are robust, the general resampling approach described by Léger and Romano (1990) and Léger *et al.* (1992) to assess uncertainty induced by a strategy should be of interest.

Complicated model screening

In Sections 3 and 4.2 Chatfield notes that our need to assess the model selection process is made difficult by the necessity for informal rules of search and subjective judgment. Such subjectivity, I believe, is certain only to add unnecessary noise to the modelling process. Faraway (1992) illustrated how to express a complicated model development strategy as a series of objectively defined decision steps. I think that Chatfield is wrong to conclude that such an approach is too inflexible. Any loss in modelling flexibility will be made up for in more accurate knowledge of modelling uncertainty.

Model averaging

In Section 5 Chatfield notes that more accurate prediction can be obtained by averaging over several models rather than by using any single model. Weighting models by their Bayesian posterior probabilities, as Chatfield suggests, can be difficult as it depends on the specification of prior probabilities (not easily elicited where there is no true model) as well as the computation of Bayes factors (problematic if the population form is uncertain). However, something very like posterior probabilities are readily available where model selection is replicated over bootstrap resamples. Efron and Gong (1983) displayed predictor selection results for a run of bootstrap resamples and commented that although 'no theory exists' for interpreting the results their variability certainly 'discourages confidence in the causal nature' (or correctness) of the model selected in the original sample. We still may lack theory, but there are many practical uses to be made of resample-based model selections.

A. S. C. Ehrenberg (South Bank University, London): Chris Chatfield's welcome paper seems to take the uncertainty out of traditional statistical modelling. He rehearses how assessing model validity for just a single set of data does not work very well.

But Chatfield still has hold of the wrong end of the stick. There is little uncertainty about working models in any reasonably well-tilled area (e.g. Ehrenberg (1994), Ehrenberg and Bound (1993) and references there). Anyone analysing yet further data can start with a reasonably well-based working model such as Boyle's law $PV \simeq C$ or $g \simeq m_1 m_2 / d^2$ for gravity.

Unfortunately, however, Chatfield mentions using such 'substantive knowledge' only in passing. And he touches on the underlying notion of many sets of data only 'Somewhat belatedly' in Section 6. He then confusingly dubs it not only 'wise' but also 'provocative', 'defeatist' and 'new' (which it certainly is not).

He also refers to it as merely 'the idea of having one or more confirmatory samples'. But that is wrong. Firstly it is a matter of determining what, if any, law-like relationship holds under a wide range

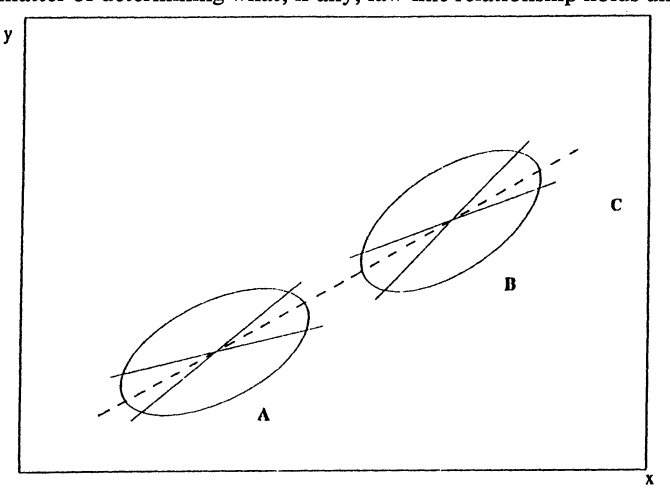

Fig. 1. A generalizable model for many sets of data A, B, etc. and model fit for single sets of data

This content downloaded from 98.161.230.204 on Mon, 02 Apr 2018 20:12:13 UTC
All use subject to http://about.jstor.org/terms

of different circumstances, such as the broken line for data A and B in Fig. 1. (Does it then also hold for C etc.?) In this way analysing many sets of data replaces the best fit criterion.

Chatfield is well aware of the issue of modelling many sets of data but still seems unable to face up to it. He worries for example about unfortunate cases where one cannot, or not *easily*, collect more data. But why does he, and others, not focus instead on cases where there *are* many sets of data?

To reinforce this plea, I briefly mention two central criticisms of the 'best fit for a single set of data' tradition.

(a) Any reasonable model will give virtually as good a fit as any other (Ehrenberg, 1982). In Fig. 1 the different lines for data set A have much the same residual scatter (which is why statistics uses least squares to pick out 'the best'). Typically, for bivariate data with a correlation of 0.6 say, a 100% difference in slopes leads only to a 3% increase in the residual standard deviation (RSD). The RSD is a very flat-bottomed function. Best fit is not the way to discriminate between alternative models.

(b) A best fit model will generally fit less well for any further data, as Chatfield notes. But the procedure would not even give the same model (Ehrenberg, 1963). This is illustrated by the two quite different pairs of (regression) lines for data A and data B in Fig. 1. I am much more worried about different or wrong answers than about the precise residual scatter about such wrong answers.

Steven G. Gilmour (University of Reading): I work with scientists who use response surface methods to study the effects of several factors on one or more response variables. In this context, I think that more emphasis must be placed on *objectives*. There is much discussion in the paper of prediction, which raises the question 'prediction of what?'. In many of my experiments, the objectives are to find the combination of factor levels which optimizes the response and to discover the pattern of response around this optimum, rather than to predict the reponse. Picturing the response surface leads to better understanding and is often more important than prediction. This makes the Bayesian model averaging approach unappealing. Of course the predicted optimal responses will be too optimistic, but the scientists would never believe such predictions without confirming them experimentally.

A typical procedure with these experiments is to fit a full second-order polynomial and then to drop terms which seem to be small. Sometimes a few higher order terms are included if the model displays lack of fit. Hence there will be model selection bias, as described in Section 4. However, two important features of the designs used are that they are nearly orthogonal and that they allow estimation of pure error. This should reduce the consequences of model selection given in Section 4.3, since the parameter estimates and their estimated standard errors hardly change when other terms are added to or deleted from the model.

A very important point is the effect of model uncertainty on the *design* of the experiment. Designs should be chosen which are good for fitting a range of different models. Criteria such as D-optimality are aimed at fitting a single model and can lead to designs which are very inefficient for fitting the model which is eventually chosen.

Parts of Section 6 could be interpreted as encouraging badly designed studies. Meta-analysis of too small clinical trials is acceptable only for political reasons. Scientifically, it would be much better to have one large properly (sequentially) designed multicentre clinical trial. Replicated studies are just large, badly designed, single studies. The reason why scientists cannot reproduce each other's results is that they have ignored important factors and each scientist holds these factors constant at different levels. We must always emphasize to scientists the importance of considering *all* factors which may be important, including noise factors. Holding a factor constant implies that the conclusions will be valid only for that level of the factor.

David Draper (University of Bath): It may sharpen one of Dr Chatfield's points, on the degree of overfitting and predictive miscalibration that routinely result from variable selection in regression, to add a numerical example to his discussion in Sections 2.3 and 4. I have reanalysed Cox and Snell's (1981) example G, on estimating the construction cost of nuclear power-plants, from the point of view of predictive validation. This analysis involves a regression of cost on $p = 10$ predictors, five of which are dichotomous, in a sample of size $n = 32$. Model specification details include the scale on which the outcome and five continuous predictors should enter the regression (e.g. raw or logarithmic), the choice of error distribution and which subset (if any) of the predictors to use. For simplicity I have conditioned on all choices made by Cox and Snell (e.g. cost on the log-scale, back-transforming as needed) except

This content downloaded from 98.161.230.204 on Mon, 02 Apr 2018 20:12:13 UTC
All use subject to http://about.jstor.org/terms

for the predictors selected; if the additional model uncertainty implied by these other data-driven choices were assessed and propagated, the predictive validation performance would be slightly worse than here.

I pretended that the regression model—including variable selection—was known, which yielded the following results: the mean of the predictive z-scores (actual − predicted)/(predictive standard error), on the 32 power-plants used to fit and estimate the model, was roughly 0, as it should be, but the standard deviation (SD) was only 0.76, and '80%' predictive intervals (using the normal theory t-multiplier) covered the actual values 91% of the time. I then jackknifed the modelling process, as far as variable selection is concerned, by successively setting aside each power-plant and using the model that maximized adjusted R^2—an approach which produces results similar to those with Cox and Snell's backward deletion method—to generate an estimated cost and predictive standard error for each plant omitted. The resulting z-values again had a mean near 0 but an SD of 1.32, and nominal 80% predictive intervals only covered out of sample 72% of the time. In this example the predictive signal is sufficiently strong that you would not expect much room for overfitting (the adjusted R^2 of the apparently best model on the full data set exceeded 82%), and yet even here the variable selection process has produced an understatement of the predictive standard errors by about 30%. Draper (1994) gives one possible general solution to this problem, based on hierarchical modelling that updates a somewhat informative prior—obtained from substantive considerations suggesting the signs of the coefficients—to a posterior on the entire β-vector, and only then, following Lindley (1968), uses a utility analysis that drops variables if they do not predict sufficiently well given how much they cost to collect. Dr Chatfield mentions this cost–benefit idea briefly in Section 3.1; it deserves wider consideration.

Peter J. Green (University of Bristol): I welcome both this paper and the previous Bath University contribution on model uncertainty that was read a few months ago (Draper, 1995). My comments, on the Bayesian approach described in Section 5 of the present paper, could really be addressed to either author.

While I agree with Dr Chatfield both that the Bayesian approach to model uncertainty via a hierarchical formulation is appealing and that Markov chain Monte Carlo (MCMC) techniques can be used to perform the computations necessary for model averaging, I am slightly reluctant to embrace *blind* model averaging as a general principle. This seems to me to have two demerits. Firstly, it is generally the case that the choice of models to be entertained and their associated probabilities is partly or completely arbitrary. Parts of model space may be over-represented or under-represented relative to the analyst's real prior beliefs simply for want of imagination. Secondly, if an average over models is *all* that is produced, then we lose the important diagnostic value in Bayes factors, and in the separate interpretation of different explanations of the data. So by all means average over models, but let us also see individual posterior analyses.

One of the delights of sample-based computation such as the MCMC method in Bayesian inference is its sheer flexibility: it is not limited to model averaging but can be used to extract all the information needed for more subtle analyses, simply by conditionally se'lecting from the sample output. I have described (Green, 1994a, b) a general framework for setting up MCMC methods to sample from the joint posterior distribution of both model indicator and parameters; this has been used to develop methodology for changepoint analysis in one and two dimensions, for Bayesian analysis of factorial experiments, and for mixture estimation; many other applications are envisaged.

Some implications in the context of the present paper are as follows:

(a) all model-specific posterior distributions can be generated in a single run;
(b) you can decide how to average over models after seeing posterior model probabilities, following Madigan and Raferty (1994) for example, or otherwise;
(c) if only Bayes factors are required, model probabilities can be chosen for computational convenience;
(d) if prior model probabilities are changed, importance sampling can be used to reweight the existing sample, rather than start again.

A. C. Davison (University of Oxford): Given some data, suppose that we are interested in the variance of a statistical estimator T. Let M indicate which of a range of models is chosen from the data, and suppose that this range contains the true model m'. Then Dr Chatfield's paper reiterates that often

$$E(T|M=m) \neq E(T) \tag{3}$$

and

$$\text{var}(T) = \text{var}_M\{E(T|M)\} + E_M\{\text{var}(T|M)\} \geqslant \text{var}(T|M=m), \tag{4}$$

for the particular model m chosen. This prompts some further thoughts.

This content downloaded from 98.161.230.204 on Mon, 02 Apr 2018 20:12:13 UTC
All use subject to http://about.jstor.org/terms

First, if the data point fairly unambiguously towards a single model m, the inequality in expression (3) is unlikely to be very great. If the choice of m results from a detailed scrutiny of the fit of different possible models to the data, $\mathrm{var}(T|M=m)$ is unlikely to be very different from $\mathrm{var}(T|M=m')$.

Second, although $\mathrm{var}_M\{E(T|M)\}$ is usually sufficiently large to ensure that the inequality in expression (4) holds, it need not. If not, the effect of allowing for model uncertainty is to increase the precision of T. Although this is rare in practice, it might arise if $\mathrm{var}\,(T|M=m)$ was large relative to most other such conditional variances and m was relatively unlikely to be chosen under repeated sampling from m'.

Third, and crucially, T must estimate a quantity defined for all the models under consideration. This might apply if the models are regarded as statistical artefacts rather than having substantive meanings, e.g. if T is a time series prediction, and the statistical procedure selects among models indistinguishable on subject-matter grounds. But, if the interpretation of the estimand of T depends on the model, the quantity of interest is $\mathrm{var}(T|M=m)$, because $\mathrm{var}(T)$ is meaningless—though we would wish to know which other models might have been chosen. This applies to variable selection in linear models with non-orthogonal design matrices, where the meaning of a parameter typically depends on all the explanatory variables. Unless the interpretations of β in the models $y=\alpha+\beta x+\epsilon$ and $y=\alpha+\beta x+\gamma z+\epsilon$ are the same, how can a combined variance be relevant?

Fourth, when T estimates a primary aspect of the problem, to quote results from a single model is analogous to point estimation, which we rightly discourage. For a confidence region approach we might use some criterion of fit to find all models consistent with the data at some confidence level, and then report summary results for all these models. Here the most useful practice seems to be the usual reporting of the estimates and standard errors for each of the most likely models, together with some idea of their relative plausibilities. The catch here—which applies equally to the Bayesian model averaging approach—is to ensure that all possible models have been considered.

To sum up: whether model uncertainty should be taken seriously depends on what the models are for.

Toby Lewis (University of East Anglia, Norwich): I would like to congratulate Dr Chatfield on his usual lively presentation and splendid audibility, and on his courteous and self-denying decision to limit the time that he took to speak to leave ample time for discussion. Following his presentation there was a whole succession of complimentary contributions, so he can happily ignore my unappreciative comments.

He said in his verbal presentation that robust inference (and I think he said the same about non-parametric inference) is a 'completely different alternative approach' which he was 'not saying much about'. In fact, there was no other mention of it, either verbally or in the preprint. Yet it is stated in the summary that

'. . . the main aim of the paper is to ensure that statisticians are aware of the problems'—

of overcoming the effects of model uncertainty—

'and start addressing the issues . . .'.

Ensure? *Start addressing*? How has the profession survived these past 20 or 30 years pending the arrival of this paper?

And now two matters of detail—there are others, but I shall keep to two. With regard to fitting a linear regression equation $E(Y|x)=\alpha+\beta x$ (example 2, Section 2.2), the author says

'A common procedure . . . is to find . . . $\hat{\beta}$, and then to fit the line provided that $\hat{\beta}$ is significantly different from 0',

and he goes on to discuss properties of this 'common' procedure. But it is surely a highly *uncommon* procedure! Whatever the size of $\hat{\beta}$ (even 0), we might want a confidence interval for β to see how large it could reasonably be. If any preliminary significance testing is to be done, it would commonly be to estimate γ, δ, . . ., the coefficients of x^2, x^3, . . . in a polynomial regression model, and then to fit $E(Y|x)=\alpha+\beta x$ if the estimates $\hat{\gamma}$, $\hat{\delta}$, . . . do *not* differ significantly from 0.

Secondly there is Section 6 of the paper entitled 'Collecting more data', Dr Chatfield's question 'What are the pedagogical implications?' in his verbal presentation, and his statement in Section 6:

'The . . . emphasis on analysing single sets of data . . . is a serious disease of statistical teaching'.

It must have been a dream, all those cohorts of students over the years, taught either by me or by many others whose teaching I encountered as an external examiner, who had learned about doing a t-test or a χ^2-test or whatever and who dutifully wrote up their answers to examples. Good marks were awarded for reaching the result 'so there is no significant difference between the treatments', but then they received full marks if they went on to say 'and therefore more data need to be collected'.

This content downloaded from 98.161.230.204 on Mon, 02 Apr 2018 20:12:13 UTC
All use subject to http://about.jstor.org/terms

The following contributions were received in writing after the meeting.

P. V. Allin (Department of National Heritage, London): The phrase 'data mining' is increasingly used in another context but raising issues along the same lines as those in the paper. Data mining, also known as knowledge discovery or unlocking corporate data, is marketed as a technique for uncovering 'information', correlations or patterns in the data held in one or more databases.

There may be clear cost and timing advantages to be gained in applying well-established, data exploratory techniques to data that have already been gathered for some other purpose, rather than collecting new data. But there are also two points to bear in mind. First, as is well known in secondary analysis, the data miner must work with the concepts and codings previously applied to the data set. This may in particular cause difficulties in data mining if data sets are to be joined. Variables that appear to be common between the data sets may be on different bases or use differing coding systems. Secondly, there is a degree of having to trust the 'black box' of some data mining software, so that this kind of data mining is not part of the iterative model building process driven by the investigator described in the paper.

Some feel for the noise in the data, and the completeness of the data set, is necessary. Data mining, especially when powerful software tools are used on good data, can be effective. However, I suggest that it still needs to be used as part of a statistical model building and testing process and not as a stand-alone technique.

Jamal R. M. Ameen (University of Glamorgan, Pontypridd): Dr Chatfield's paper addresses one of the most fundamental problems in the philosophy of science, namely the way that scientific problems are formulated. Fundamental to some of the arguments raised in the paper is what is meant by a *model*. The conflict regarding the existence of a or the true model, model uncertainty, identification, selection, misspecification, etc. can be resolved if a clear definition of a model is stated. Models have different meanings in different fields. Once a young woman was interviewed for a data handling job in Unilever. She was puzzled by one of the interviewers who asked her whether she had done any modelling. It was found later that to her the question meant whether she had posed for photographs! If a model is assumed to be a 'small scale' representation of an object, then by definition all models are wrong, simply because they are not the true objects that they are to represent. However, if a model is seen to be a device that a scientist uses to understand some natural phenomena better, then it is natural for the model to be

(a) satisfactory in performance relative to the stated objective,
(b) logically sound,
(c) representative,
(d) questionable and subject to on-line interrogation,
(e) able to accommodate external or expert information and
(f) able to convey information.

Unless they are misspecified in terms of one or more of these properties, all models are acceptable. Some models are more acceptable than others.

The subject of the paper has deep roots in objectivity and subjectivity arguments. The objective idea of searching for the true model is that of classical mathematics ('You believe in a God that plays dice, and I in complete law and order', said Albert Einstein to Max Born) and had a relatively short life even in mathematics. Probability theory was used to extend the ability of deterministic models to represent more open systems but the existence of the true model has remained with many modellers. The Bayesian philosophy views models from a subjective viewpoint.

Even when the scientist is clear about the irrelevance of including seasonality (say) or a local linear trend instead of a third-degree polynomial, a multiprocess model is still relevant. This is one of the main features of Bayesian modelling not only for model averaging (West and Harrison, 1989) but also when the state of the system is disturbed by unforeseen events (Ameen and Harrison, 1984).

G. A. Barnard (Colchester): Dr Chatfield's emphasis on the fact that we statisticians serve *clients* is only one of many welcome points which brevity demands that I pass over. The one point that I can concentrate on concerns the extent to which computers now allow us to reduce the number and strength of assumptions we need to enable us to give useful answers to our clients. Dubious assumptions of normality, for example, have been logically unnecessary since Fisher's (1934) paper on conditional inference; but computers now allow us in practice to calculate the distribution of location and scale pivotals,

This content downloaded from 98.161.230.204 on Mon, 02 Apr 2018 20:12:13 UTC
All use subject to http://about.jstor.org/terms

conditional on ancillaries, for wide ranges of distributional forms. Given a random sample x_i, $i=1, 2, \ldots, n$, the assumption that the shape ϕ of the population density is unimodal is verified much more often than that it is symmetric, let alone that it is normal. In such a case we can take the population mode μ as a location parameter, with the direct interpretation that it is the most likely value for predicting the next observation, while the scale parameter which we denote by σ can be equally directly interpreted as the length of the shortest interval containing at least half the population. It is a pity that it does not have a generally appropriate name. The pivotals $p_i=(x_i-\mu)/\sigma$ then transform to (s, t, c_j), $j=1, 2, \ldots, n-2$, with $p_i=s(t+c_i)$ and, with the usual sample notation, $s=s_x/\sigma\sqrt{n}$, $t=(\bar{x}-\mu)\sqrt{n}/s_x$, $c_i=(x_i-\bar{x})\sqrt{n}/s_x$. The joint conditional density of (s, t) given the observed values c_{i0} of the c_i can then be written as

$$\psi(s, t \mid \underline{c}_0)=Ks^{n-1}\prod_i \phi\{s(t+c_{i0})\}$$

and it may turn out that ψ varies little when ϕ is changed over a wide range of shapes ϕ. Such 'conditionally robust' behaviour is even more frequent with the marginal density $\xi(t \mid \underline{c}_0)$, since integrating out s has a powerful smoothing effect. For example, we have found that 'normal looking' samples taken from skewed Cauchy densities (pathological because such densities do not have even generalized means) can often be treated without serious error as coming from normal densities.

Before querying the phrase 'without serious error' we may ask for more study, for example, of the extent to which typical clients can distinguish between probabilities of 0.06, 0.05 and 0.04, or between 0.006, 0.005 and 0.004. Fisher helped to cause much unnecessary trouble by excessive use of words such as 'wholly' and 'exact', for instance in Fisher (1970), pages 9–10:

> '. . . the theory of inverse probability . . . must be wholly rejected. Inferences respecting populations . . . cannot . . . be expressed in terms of probability, save in those cases in which there is an observational basis for making exact probability statements in advance about the populations in question.'

David Bartholomew (London School of Economics and Political Science): The issues which Dr Chatfield raises were live ones in the manpower planning world in the early 1970s. In the work of the Civil Service Department's Statistics Division at that time we identified and tried to allow for four kinds of uncertainty. Two were identical with (b) and (c) in the list on the third page, one was uncertainty about the accuracy of the data and the fourth was what we called 'specification error'. This was essentially the same as (i) and (ii) of the author's category (a). We found that specification error was usually by far the most important. Sensitivity analysis therefore played a much more important role in our work than the usual calculation of standard errors and suchlike. A paper making and illustrating this point was originally submitted for reading at an Ordinary Meeting but was eventually published in a non-statistical journal (Bartholomew *et al.*, 1976). In that paper we also noted that similar questions had been raised in demography and Hoem (1973) was cited. Statisticians have been slow to give proper recognition to model uncertainty.

Trevor Bedeman (TSB Consumer Credit, Brighton): A related problem is that of comparative model performance in cross-sectional modelling. Claims are currently being made that computationally intensive methods such as neural networks and genetic algorithms can outperform traditional methods such as logistic regression in credit assessment. In this field even a small increase in predictive power could result in large savings. The dangers of overfitting, however, mean that great care is needed in evaluating the competing claims of differing models. A few per cent difference could either represent real improvement and cost savings or be the result of error made at several stages of comparison.

I recommend the following structure for operational testing. From within the samples taken for model development, a substantial hold-out subsample should be taken. This is becoming standard practice with techniques such as neural networks and genetic algorithms. It is used to provide a basic check against overfitting these models. But there are dangers in reliance purely on the hold-out sample. Once such a sample is being used to influence model building, then the results are no longer truly independent. A model developer will typically look several times at preferred versions of the final model and at the hold-out sample and results obtained from this hold-out sample are likely to have influenced the finished model.

So it is also necessary to assess competing model performance against a completely separate sample, which is effectively unseen and unused until after the final model is complete. This validation sample

This content downloaded from 98.161.230.204 on Mon, 02 Apr 2018 20:12:13 UTC
All use subject to http://about.jstor.org/terms

should ideally consist of fresh performance information taken after the original development sample. It then provides a test of the operational performance of the model. It is this testing against the unseen validation sample that is most revealing of the actual predictive capacity of such competing models. Validation results which are of a different order from those of the development and hold-out samples show the importance of such full operational testing.

R. J. Bhansali (University of Liverpool): In view of the recent developments on model selection, the question of statistical inference after model selection is timely. I am, however, disappointed by this paper in that the author seems only to nibble at the underlying problem and does not appear to introduce new methods for dealing with it. I am also puzzled by the emphasis on the Bayesian approach. The underlying idea is not new (e.g. Akaike (1979)) and it is a direct consequence of Bayes's theorem when a model is treated as a random quantity. It is unclear, however, how to use this approach when the number of candidate models is large and allowed to increase with n. Also, the approach is predicated on the hypothesis of a 'true' model, which is incongruent with the view (Rissanen, 1987) that for observed data there can be no true model. It is possible to develop statistical procedures based on the latter point of view. There is currently much interest in lead-time-dependent model selection and/or parameter estimation for time series forecasting. The latter may be considered when using a possibly under-parameterized method, e.g. exponential smoothing, for multistep forecasting and the former when adopting a nonparametric approach, e.g. autoregressive model fitting: see Findley (1985), Tiao and Xu (1993) and Bhansali (1993, 1994), among others.

Leo Breiman (University of California, Berkeley): Dr Chatfield has admirably undertaken the job of looking at limitations of models and inference. Unfortunately, our profession has tacitly encouraged the use of inference to the point of extreme abuse.

Here is a hypothetical (but realistic) example: company X is suspected of sex discrimination in salary. 18 relevant variables are extracted from the personnel record of each employee. A linear regression fit to salary is computed. Using the model that salary equals a linear function of the predictor variables plus independently and identically distributed normal noise, the coefficient of the binary sex variable is significant at the 5% level. An expert testifies that this constitutes proof of discrimination. Similar regression modelling is used throughout the social sciences to establish causality.

Setting aside the vaporous claims of causality, there is a fundamental problem with inference in this setting. The inference in the above example can be restated as: if company X is independently replicated many times and if, in the replicates, the 'true' coefficient of the sex variable is 0, then in fewer than 5% of the replicates will the estimated coefficient of sex have a significant t-value. But company X is non-replicable, a stochastic model is not possible for non-replicable data, and the inference makes no sense.

To try a fix, we estimate the ordinary least squares (OLS) coefficients by using half the data and compute prediction errors by using the other half. If the predictor using the sex variable gives appreciably more accurate predictions than the predictor not using the sex variable, we have the more modest conclusion that the sex variable is an important predictor of salary.

But another prediction method might be more accurate than OLS linear prediction and have the same accuracy whether or not the sex variable was used. Thus, even the modest assertion that a certain variable is important in predicting future outcomes cannot be decided by using inference on a linear model, unless we establish its approximate validity—an impossible undertaking in data that are not low dimensional.

Some models are sound and useful—i.e. analysis-of-variance models. Yet, inference using models in data with many mutually dependent covariates is generally unsound. Can a borderline be defined? Some of the clearest work in this area is by David Freedman and Freedman (1991, 1995) are excellent introductions to his views (which I share) and to opposing views.

D. R. Cox (Nuffield College, Oxford): Tension between theory and application arises in many fields of activity and an issue with Dr Chatfield's thought-provoking paper concerns whether he has correctly identified major matters, the resolution of which would help to reduce such tensions in our subject. The answer must depend in part on the kinds of application concerned.

His formulations all seem to assume that the objective is fixed, the forecasting of something or the estimation of a given parameter, whereas my own worries in applications centre more on whether the right qualitative goal has been selected, whether the parameter or parameters chosen to encapsulate

This content downloaded from 98.161.230.204 on Mon, 02 Apr 2018 20:12:13 UTC
All use subject to http://about.jstor.org/terms

that goal in idealized form are well judged and whether some major feature of data collection or potential bias has been totally overlooked or some important subject-matter considerations ignored. Within a given broad framework it may be essential to recognize explicitly that different models with different interpretations fit about equally well; procedures which then force a single choice are bound to be potentially misleading. By recognizing that data admit different interpretations many considerations about multiple testing become irrelevant (Cox and Snell, 1974).

Where it is a question of uncertainty assessment with an agreed target, a comparison of different models, formally or informally, is an important possibility, although even then underestimation of the 'real' uncertainty is likely, partly because of some instability in the target and partly because the dangerous uncertainties are those that have been overlooked.

Often the term data mining is used in a rather derogatory sense, although Dr Chatfield himself is not be criticized over this. I understand mining to be a very carefully planned search for valuables hidden out of sight, not a haphazard ramble. Mining is thus a rewarding but, of course, dangerous activity. It is an interesting issue, very specific to each subject-matter field, as to what extent important conclusions from data 'lie on the surface'.

Finally it does not seem helpful just to say that all models are wrong. The very word model implies simplification and idealization. The idea that complex physical, biological or sociological systems can be exactly described by a few formulae is patently absurd. The construction of idealized representations that capture important stable aspects of such systems is, however, a vital part of general scientific analysis and statistical models, especially substantive ones (Cox, 1990), do not seem essentially different from other kinds of model.

Simon Day (Leo Laboratories Ltd, Princes Risborough): Dr Chatfield presents a well-reasoned caution clarifying the fact that uncertainty exists in defining the boundaries between data mining and careful, thorough, data analysis. The dangers are real and the importance of this paper seems, therefore, without question.

I do not question the reality of the problem but I question its seriousness. If I introduce a new medical procedure, manufacturing process, perhaps even a new law and with the passing of time a better procedure, process, law etc. is found then I can and should change. If I make a mistake in analysing a set of data then I can re-do the analysis and re-present the results. It may be that my previous decision about a procedure or law was wrong, or at least not as good as it might have been. That is unfortunate, possibly even disastrous; there may be nothing that I can do to rectify past mistakes but at least I can learn and change for the future. But what if my data analysis that led to that decision was wrong; wrong, not just for one particular problem, but the basic philosophy was wrong? Should we, as statisticians, now retract all our data analyses as being unsound and so possibly leading to wrong conclusions? I doubt that we would be willing to do this. Nor am I convinced that it is necessary. Most of us have probably made these types of 'error' at some time. If the analyses that we have performed are suspect then we have a duty to retract them, or at least to revisit them. The only exception can be if we judge (or better if we can prove) that the consequences are not substantial. Can we take any comfort from the test of time and observe that, despite our errors, statistics is not 'getting it wrong' daily and so the consequences may not be substantial? I am not absolutely convinced, but I hope so.

Robert Fildes and Mike Pidd (Lancaster University): Chatfield's paper is an extremely welcome reminder that statistical analysis is as much about model building as about inference. However, there is an undoubted risk that attempts at statistical model building may make very efficient use of computer software and may even be theoretically elegant, yet they may produce models which are very ineffective in use. Indeed Fildes and Makridakis (1995) reviewed 20 years of the research literature on time series analysis and concluded that most effort has been devoted to minor extensions of the autoregressive integrated moving average model class and to tests of related hypotheses. Discussions of model validation and robustness, especially those which linked problem context, data and model, were extremely rare.

To make sense of this and to make progress, it is important to think through the nature of modelling, particularly of statistical modelling. In this we agree with Chatfield that a model will always be an approximation and, to add to Chatfield, a model should provide a convenient vehicle for experimentation, more convenient than the system being modelled. In discrete simulation, Zeigler (1976) popularized the idea that models are valid only within a fully specified experimental frame. The specification of that frame is not intended as part of a search for truth but is simply a statement of the intended use

This content downloaded from 98.161.230.204 on Mon, 02 Apr 2018 20:12:13 UTC
All use subject to http://about.jstor.org/terms

for the model. Even within a specified experimental frame, several models may be valid, in the sense of fitness for purpose. The idea of true models is, as Chatfield says, not helpful.

Statistical analysis is rarely an end in itself and many other communities make use both of statistics and statisticians. Our own management science community is one such and it has a well-developed literature on model building as a process (see, for example, Mitchell (1994)). In management science we take for granted that fuzziness, uncertainty and problems of interpretation abound—in addition to the technical statistical issues of model choice highlighted by Chatfield.

Hence our response is to applaud Chatfield's discussion of model uncertainty, data mining and the link to inference. But we part company if he regards this as primarily a statistical problem. Instead, we believe that progress is most likely to be made through this mine-field when statistical techniques are combined with a rigorous analysis of the intended use of a model. This will help to define a class of potentially useful models within which analysis can be conducted.

A. D. Gordon (University of St Andrews): Some of the problems of model uncertainty and validation of the results of an analysis that are described in this paper also arise in cluster analysis. Different clustering procedures can provide markedly different analyses of the same set of objects; in effect, each clustering criterion implicitly involves a model for the data. One approach to this problem has been to analyse the data by using several different clustering procedures, which theory or background information suggests might be appropriate, and to synthesize the set of results in a *consensus* classification (Gordon (1981), chapter 6).

There has also been concern to assess the validity of the cluster output; an overview of relevant tests and procedures is given by Gordon (1994). Two approaches of particular interest have involved assessing replicability across subsamples and simulation tests. In the first approach, the data set is randomly divided into two and the objects in each subset are

(a) clustered separately and
(b) assigned to the 'nearest' cluster in the other classification.

A high correspondence between relevant partitions increases confidence in the validity of the results (McIntyre and Blashfield, 1980; Breckenridge, 1989). In the second approach, data sets randomly generated under a null model of the absence of group structure are analysed by using the same clustering procedure as was applied to the original data set, allowing a Monte Carlo test of the strength of support for its clustering (e.g. Arnold (1979) and Milligan and Sokol (1980)). The results depend on the choice of null model, and use has been made of data-dependent null models; for example, simulated data sets can be required to have similar covariance structure to the original data set, or data points can be generated uniformly within the convex hull of the original data.

Although cluster analysis is not concerned with the estimation of parameters, I believe that it could usefully be included in discussions of the effects of model uncertainty in the analysis of data.

Howard Grubb (University of Reading): I would like to discuss some issues which arise from trying to measure uncertainty in a geophysical procedure (marine seismic surveying). In this case, our model M is of some property of the earth estimated at various locations. These have been determined by physical considerations, so there can be considered to be no model structure uncertainty; however, there is uncertainty in the estimated values. This model is then used in a further stage of processing to calculate other properties of the earth (P, Q) and these final coefficients are of interest to geologists. It is therefore uncertainty resulting at this stage and on the scale of the earth parameters which is important, rather than, say, confidence in a particular model M. Notice that this uncertainty varies spatially and is of different orders for each of the coefficients (P, Q) (i.e. some are better determined than others).

Uncertainty in this procedure can be considered to arise from three sources:

(a) variability in the parameter estimates of model M;
(b) precision of the final estimates (P, Q) due to the processing procedure;
(c) quality or validity of physical assumptions for particular M.

(a) and (b) correspond to Chatfield's (b) and (c) in Section 1, although, while mentioning model validation as part of the model building process, he does not consider this as an element of uncertainty which could be measured, rather than just taken as good or bad. In this application, the physical model may be less valid at certain points, depending on the particular earth structures.

This content downloaded from 98.161.230.204 on Mon, 02 Apr 2018 20:12:13 UTC
All use subject to http://about.jstor.org/terms

This analysis raises the issues of measuring the uncertainty (we must have suitable scales on which to do this), and of combining these uncertainties in meaningful ways. An initial solution to this has been implemented using simulation to estimate the effect of variability, although this is a very expensive approach. Analysis of the processing procedure can give us a measure of precision, whereas physical knowledge allows us to devise measures of quality or validity, although the relative scales of these measures and their combination can be difficult to determine in a complex procedure.

Urban Hjorth (Linköping University): Surprisingly, it is not easy to comment on a work with which one agrees as completely as I do with Dr Chatfield's presentation. I congratulate him on his wide coverage and well-articulated description of the difficulties and challenges in inference after model selection.

We now know that classical inference cannot be trusted after data-based model selection and that in particular classical measures of fit and predictability, based on the properties of the fitted model only, will be overoptimistic. Of course there is nothing wrong in searching different models for a good fit. This is just another version of the maximum likelihood principle. The mistake is to pretend afterwards that only a small subset of all the possible models is involved in the analysis. However, we often do not know how to handle the big problem and are sometimes even unable to define the class of models.

Chatfield mentions two possibilities: the use of computer-intensive methods, when the modelling can be at least approximated by an automatic procedure, and the challenging Bayesian analysis where no particular model is selected. Both kinds of method are of great interest. I would like to add that, since the gain of model selection appears to be smaller than the fit indicates, and the complexity much larger, it can be useful to bring some problems back to the classical situation by more careful thinking before the analysis of data and by defining more general parametric models. On a more practical level, we need to incorporate model selection analysis in software for regression and time series modelling. A far reaching vision is to structure a package where any use of explorative search, transformations and other modelling tools could be accounted for in a statistical evaluation of the results. Chris Chatfield's paper will hopefully stimulate much new research in the area.

Raymond Hubbard (Drake University, Des Moines): I applaud Dr Chatfield for advocating a 'broader' view of statistical inference to include the whole model building process. Likewise, I share his concerns about the widespread practice of data mining and its effects on the veracity of published empirical results.

It is ironic, then, that journal editors and reviewers promote data mining. They do so because of their bias in favour of publishing manuscripts with statistically significant ($p<0.05$) results (Denton, 1985), which can fuel data mining among those researchers who failed to obtain them initially. As Mayer (1980), p. 175, observed, 'If you just torture the data long enough, they will confess'.

I particularly agree with Dr Chatfield's argument for validating models by *replicating* them on new data sets. But I would caution that ideally model replication should be undertaken by independent (different) researchers. Researchers replicating their own work are more likely to report a confirmation of previous results. For example, in a study concerning replication research in five business disciplines during 1970–91, it was discovered that 266 of 4270 empirical papers (6.2%) qualified as replications with extensions (Hubbard and Vetter, 1995). 51 of these 266 replications involved authors replicating their own work, and only five (9.8%) of them reported findings that conflicted with earlier results. In contrast, of the 215 replications carried out by independent researchers 116 (54%) conflicted with earlier results.

Unfortunately, editors and reviewers again impede the development of knowledge through their bias against publishing replication research (Neuliep and Crandall, 1993). Thus, instead of accumulating results based on significant sameness, we tend to be left with fragmented and isolated findings whose value is minimal.

Because of editorial–reviewer biases, I am not optimistic that the changes necessary for improving the climate of empirical research will occur. The bias against publishing statistically insignificant results encourages data mining, thereby exacerbating the problems regarding model uncertainty and inference. The bias against publishing replication research discourages empirical generalization. And these biases continue to exist despite numerous pleas to eliminate them. So, even after acknowledging Dr Chatfield's valuable insights, I am afraid that for most researchers it will be back to business as usual.

David Madigan (University of Washington, Seattle): Dr Chatfield's paper adds to the rapidly expanding number of references on model uncertainty. That this activity is taking place more than a decade after the appearance of Leamer (1978) is indeed a 'quiet scandal'.

This content downloaded from 98.161.230.204 on Mon, 02 Apr 2018 20:12:13 UTC
All use subject to http://about.jstor.org/terms

I shall comment on five specific points in Dr Chatfield's paper. First, readers of Section 4 might have the impression that large samples alleviate 'model selection biases'. York *et al.* (1995), Madigan and York (1995) and Kass and Raftery (1993) described applications which demonstrate that there is little room for complacence, no matter how large the sample.

Second, elicitation of uncontroversial informative prior model probabilities for Bayesian model averaging (BMA) is possible. Madigan *et al.* (1995) described one approach and presented a medical application where an informative prior distribution provides improved out-of-sample predictive performance when compared with a vague alternative.

Third, Dr Chatfield states that BMA does not lead to simple models which 'may not matter for forecasting purposes but does matter for description and interpretation'. However, BMA can provide a posterior distribution for *any* quantity of interest, not just forecasts. For example, Raftery *et al.* (1995) present posterior distributions for effect sizes averaged across models.

Fourth, there is mounting empirical evidence that BMA consistently provides improved out-of-sample predictive performance for a range of model classes including linear regression (Raftery *et al.*, 1994), graphical models for categorical data (Madigan and Raftery, 1994) and event history analysis (Raftery *et al.*, 1995). The results from these studies are quite similar: in most cases, BMA improves the predictive performance over the single best model by about the same amount as would be achieved by increasing the sample size by 4% (Raftery, 1995). Madigan and Raftery (1994), equation (4), provide a theoretical underpinning for these results.

Finally, Dr Chatfield makes somewhat disparaging remarks about data mining. Used wisely, however, apparently egregious data mining methods can provide useful results (see, for example, Riddle *et al.* (1994)). Within the emerging 'knowledge discovery in databases' (KDD) community there is considerable disenchantment with traditional statistical methods. Although it seems important to point out the inadequacies of certain KDD approaches, we must do more than merely criticize.

John M. Marriott (Nottingham Trent University): Dr Chatfield discusses what he calls the Bayesian model averaging (BMA) approach. I believe that this approach, as described in the paper, is just a special case arising from a Bayesian decision analysis that is not necessarily applicable in situations where there is no belief in a true model. The approach can be summarized as combining the predictions $\hat{y}_i$ from the individual models to obtain an overall prediction as

$$\hat{y}_C = \sum_i \hat{y}_i P(M_i|\mathbf{x}).$$

In presenting model choice in a decision context Bernardo and Smith (1994) discussed *three* alternative ways in which the possible models might be viewed. If we restrict our consideration to the first of these, the $\mathcal{M}$-closed case, in which the true model is assumed to be one of the set of models being considered, and then *further restrict* the decision problem to pure prediction, the Bayesian procedure for model selection is then to choose an action that will minimize the posterior expected loss

$$\mathrm{EL} = \int_y L(\hat{y}, y) p(y|\mathbf{x}) \, \mathrm{d}y$$

where the loss function $L(\hat{y}, y)$ measures the loss from using $\hat{y}$ when the true, as yet unobserved, value is y.

In the *special case* of a quadratic loss function, the optimal predictor is provided by averaging the individual posterior predictive means over the posterior distribution of the models,

$$\hat{y}_B = \sum_i P(M_i|\mathbf{x}) \int_y y p(y|\mathbf{x}, M_i) \, \mathrm{d}y$$

which coincides with the BMA approach if $\hat{y}_i$ is interpreted as the posterior predictive mean.

In the context of time series models there are many references in which either $\hat{y}_B$ is employed (see for example Monahan (1983) or Marriott and Tremayne (1988)) or different loss functions and their effect on model choice criteria are considered (see for example Chow (1981), Kashyap (1982) and Poskitt (1987)). In all these consideration is restricted to the $\mathcal{M}$-closed case.

None of the results referred to above automatically extend to the case in which the true model is assumed unknown.

In presenting his example of the BMA approach Dr Chatfield does not make this clear, nor does he indicate that his results do not apply regardless of the choice of loss function or prior beliefs about model parameters.

This content downloaded from 98.161.230.204 on Mon, 02 Apr 2018 20:12:13 UTC
All use subject to http://about.jstor.org/terms

Alan J. Miller (CSIRO Division of Mathematics and Statistics, Melbourne): I very much welcome this excellent review of the current state of the art of inference after model building. The comprehensive set of recent references is particularly valuable. Computational methods of model building, such as alternating conditional expectation, classification and regression trees, projection pursuit and generalized additive modelling, have proliferated in recent years yet we have only just started research into how to draw inferences for selecting subsets of populations, subset selection in linear regression and autoregressive integrated moving average modelling. One name which is conspicuous to me though by its absence is that of T. A. Bancroft. See particularly Bancroft and Han (1977).

Let me put my contribution in the form of a question. Given a set of data with no accepted model, which of the following should we do?

(a) Should we split the data into parts which we use
  (i) to build a model,
  (ii) to estimate the parameters in the chosen model and
  (iii) finally 'validate' the model (though invalidation is a preferable attitude)?

If we do this, how should we do the splitting?

(b) Alternatively, should we use all the data for the three phases above but in inference make allowance for the data mining process, and if so, how?

Common practice is either to use the second alternative but with no allowance for overfitting, or to split off a small part of the data for validation. The few investigations which have been carried out (e.g. Roecker (1991)) suggest strongly that the alternative of using all available data for model building is better. The reason appears to be that, the more data we use for the data mining phase, the closer we approach to a realistic model.

Suppose that we then carry out a second data collection or experiment after the first has narrowed our choice of models. Should we combine both data sets for the next phase of model building? I think so.

I would like to question one statement in (d) of Section 4.3 'prediction intervals are generally too narrow'. Chatfield may be right but my limited experience is that the width of the intervals (in regression) is roughly correct; the problem is that they are in the wrong place!

Anthony O'Hagan (University of Nottingham): Dr Chatfield is to be congratulated for a timely and thought-provoking paper. His criticism of so much of common statistical practice, and particularly of classical inference, in the presence of model uncertainty is thorough and very welcome. I was disappointed to see, however, so little space on solutions, particularly when I find the discussion in Section 6 quite unconvincing. How is collecting more data different from arbitrarily splitting the existing data? *After* collecting more data, we again have a single, albeit larger, set of data. To regard the part that we received first as that to be used for model building, and the new part as that for model confirmation, is just as arbitrary as splitting the original data into two parts for these purposes. Furthermore, the process makes little sense in either case. What if the 'confirmation' sample fails to confirm the model? However one describes it, the new data are being used for further learning about the model, and it is proper to use *all* the data explicitly for model inference, as a single data set. This is what the standard Bayesian procedure, now called Bayesian model averaging, would do.

The latter part of Section 6 similarly unconvincingly tries to say that statistical analysis of several data sets is something other than analysis of one large data set. All analysis is of a single data set, but some data sets have more complicated structures than others, which must be acknowledged in the form of model and analysis adopted.

Bayesian model averaging is the most appropriate form of inference, but Dr Chatfield correctly identifies the principal practical difficulty as the choice of the prior probabilities for different models. It is dangerous when used with an unstructured collection of models, and particularly if some of those models have been suggested by the data. (Non-Bayesian model averaging with fixed weights, which Dr Chatfield suggests in the context of 'combining forecasts', is not consistent.) Dr Chatfield does not discuss Bayesian model averaging in detail, so this is not the place for an extensive comment on applications of the technique, but I would like to draw attention to a new computational technique which promises to be extremely powerful in applications. Green (1994b) shows how to use Markov chain Monte Carlo (MCMC) techniques on the union of the parameter spaces of all the models. This enables inference both within and across models to be derived from one MCMC computation. Current work suggests that Green's method combines particularly well with the fractional Bayes factors approach of O'Hagan (1995) when prior information on parameters within models is weak.

This content downloaded from 98.161.230.204 on Mon, 02 Apr 2018 20:12:13 UTC
All use subject to http://about.jstor.org/terms

Benedikt M. Pötscher (Universität Wien): The author is to be commended for exhorting statisticians to face the implications of model uncertainty and to try to take on this old and nagging problem. My comment relates to the asymptotic results discussed at the beginning of Section 4. The perhaps most important result in Pötscher (1991) concerns model selection procedures obtained from a sequence of hypothesis tests and describes the asymptotic distribution of parameter estimators conditional on the event that a particular model has been selected by this model selection procedure. It turns out that conditionally on having selected the minimal true model this asymptotic distribution coincides with the classical normal distribution (seemingly implying that there is no effect from using a model selection procedure in this case), but that conditionally on selecting an overparameterized true model this asymptotic distribution can have a very different form. (Underparameterized models are never selected asymptotically.) These asymptotic results must be used with great care, however, since the convergence of the finite sample distributions to their asymptotic equivalents is not uniform as the true parameter varies over the parameter space. The non-uniformity arises near lower dimensional models and is related to the fact that selection of underparameterized models, which may occur in small samples but not asymptotically, is not taken into account in the asymptotics. See Kabaila (1995), Pötscher (1995) and Shibata (1986) for more discussion. As a consequnce, in finite samples the picture can be quite different from that predicted by asymptotic theory. A study of the finite sample distributions of parameter estimators conditional on having selected a particular model is reported in Pötscher and Novak (1994). It is found that the asymptotic results mentioned above provide good approximations for the finite sample distributions in case we condition on the event that a true but overparameterized model has been selected. However, conditionally on selecting the minimal true model, the finite sample distributions can sometimes differ dramatically from their asymptotic counterparts. Hence, although the asymptotic result seems to tell us that there is no effect from model selection when selecting the minimal true model, this is not what we find in small samples. This is linked to the possibility of selecting underparameterized models in small samples; see Pötscher and Novak (1994) for more discussion. (There is, however, a mark of this phenomenon also in the asymptotics, namely the non-uniformity alluded to above.)

W. D. Ray (Leatherhead): The subject of this paper is ideal for discussion since it focuses directly at the basics of statistical inference. It raises fundamental issues on which every statistician must have definite views and opinions. These issues deserve to be widely debated.

Firstly is 'the single-model issue'. It may be accepted that there is no ideal model, but, and here is the rub, with what do you replace it . . . many models? That certainly gives flexibility but when do you stop counting and who chooses the final set? It seems a little presumptuous then to choose priors, either to soothe one's conscience or to be able to make some mathematical steps out of convenience. How do you evaluate the priors and who does the evaluating?

In Section 5 this theme is pursued and the author seems to approve the Bayesian approach. However, this appears as equally fallacious as the 'single-model' direction. What we are doing is replacing wishful thinking in choosing one model with wishful thinking about choosing several models!

The only sure thing (or should be) is the *data*! This viewpoint puts the nonparametric approach, bootstrap and all, as the safest path to follow, and it is encouraging to see the developments in this direction.

The difficulties, however, remain and as the author remarks there is no comprehensive answer.

Referring to the time series field I have often thought that a clutch of different models might be used to advantage when forecasting, each being a function of the number of steps ahead to be forecasted. It is asking too much of a *single* model to trap the *long*- and *short*-term structure simultaneously.

Evidence for this is often apparent from a look at the spectrum when there is significant power at both low *and* high frequencies.

R. A. Sugden (Goldsmiths College, London): Dr Chatfield brings to our attention some unpleasant home-truths about conditioning.

His warnings are certainly relevant for model-based survey samplers. Suppose that the two models under consideration are as in example 2 and a linear least squares predictor of the finite population mean, essentially a 'global' parameter, is required. Under model I the linear regression estimator

$$\bar{y}+\hat{\beta}(\bar{X}-\bar{x})$$

is used (assuming that the finite population mean $\bar{X}$ is known) and, under model II, the sample mean $\bar{y}$. If a sample-based choice of model is made, then this 'pretest' estimator certainly suffers from model selection bias.

This content downloaded from 98.161.230.204 on Mon, 02 Apr 2018 20:12:13 UTC
All use subject to http://about.jstor.org/terms

It is important to realize that design-based survey samplers, who base their inference solely on the distribution generated by repeated applications of the sample design and not on models, can also suffer from selection bias, although a better term might be 'estimator selection bias'. Consider the same problem as above, where the design is say simple random sampling (applied to a fixed but unknown finite population). The linear regression estimator is chosen for certain samples, which are subsets of size n, and the sample mean for the rest. Conditionally on either choice, the population units now no longer have equal survey weights but unequal and moreover unknown weights. These are the inverse inclusion probabilities given that the sample s now falls in a restricted sample space. Faced with difficulties like these, a design-based survey statistician might retreat to unconditional inferences. Similar problems are encountered in the decision whether or not to post-stratify, even when the stratum sizes are known.

The **author** replied later, in writing, as follows.

I thank all the discussants for their generally encouraging and constructive contributions. The large number of comments means that I cannot reply to all of them, and, in any case, many of them need no specific reply. Thus the absence of a response to a particular point should not necessarily be taken to imply lack of interest on my part.

I am grateful for the additional references, especially Pötscher and Novak's (1994) recent small sample results on inference after model selection, and Bancroft and Han's pre-1977 bibliography of what was then called *conditional specification*. I also welcome the additional insights into model uncertainty gained from a variety of areas of science and statistics such as manpower planning (Bartholomew), cluster analysis (Gordon), geophysics (Grubb) and sample surveys (Sugden).

Wider issues of modelling and problem solving

Several discussants have commented on wider aspects of statistical inference. I agree with Hand that such issues need more attention, in contrast with the detailed refinements of techniques, where there is a rapidly growing literature. We should indeed keep the *problem*, rather than the model, centre stage, and the *clarification of objectives* (Gilmour, Cox) is crucial. Nevertheless, *modelling* is of course important and general questions thereon also receive much attention (e.g. Copas, Hand, Ameen, Davison, and Fildes and Pidd). What constitutes 'a model' and what is the purpose of modelling? Hand helpfully distinguishes between models used as *representations* and as *descriptions*, whereas Ameen describes a model as 'a device that a scientist uses to understand better some natural phenomena'. Of course all models are approximations (Fildes and Pidd) and it is better to describe models as better or worse approximations rather than 'right' or 'wrong' (Cox), but I still think that the old aphorism 'All models are wrong, but some are useful' can be a salutary reminder of the frailties of modelling. My lack of belief in a 'true' model means that I do not accept the premise behind what Marriott calls the $\mathcal{M}$-closed case. We also need to remember that several different models may fit the data about equally well (Ehrenberg, Cox, Fildes and Pidd) and that the choice of model will depend partly on the particular application. For example Ray reminds us that different time series models may be appropriate for short- and long-term forecasts. I also agree with Davies that I should have said more in Section 3.1 about models with *time varying* parameters. In time series analysis in particular, the use of such (local) models, which may readily be updated by the Kalman filter, is increasing and has many benefits compared with constant parameter (global) models.

Copas says that, in properly designed experiments, a null model is 'simply a description of the randomization used in the design'. This is only helpful up to a point. Knowing the null distribution of some test statistic does not help you to describe (model) the effects of the treatments when the null hypothesis is rejected. However, I agree with Copas that, with observational data, 'we should emphasize the descriptive rather than the inferential nature of our analysis'. Breiman's remarks on company X are relevant here, whereas Gilmour's remarks remind us that good (e.g. orthogonal) design is a necessary precursor to well-grounded analysis and can help to reduce model uncertainty.

Data mining

The phrase 'data mining' has caused some comment (Allin, Cox, Hubbard). It appears that the phrase can be used to denote sensible 'digging' into data to try to reveal what they are saying. But, rightly or wrongly, it has more often become synonymous with 'torturing the data till they confess', especially in the econometric literature (Lovell, 1983; Denton, 1985). There is of course a fine line between the careful study or evaluation of a set of data and an overzealous modelling spree (Day, Madigan).

This content downloaded from 98.161.230.204 on Mon, 02 Apr 2018 20:12:13 UTC
All use subject to http://about.jstor.org/terms

Many sets of data or one?

Statisticians must keep a sensible balance between solving problems with one-off data sets and looking for confirmation or generalization with two or more data sets. I have routinely had to tackle both types of problem. I therefore strongly disagree with those comments which dispute this balance (I seem to have become a 'moderate'!). At one extreme, Ehrenberg complains that statisticians worry too much about 'unfortunate cases' where one cannot easily collect more data. I think that describing the whole of time series analysis, for example, as an unfortunate case is preposterous. Although we should always be looking to replicate and generalize wherever possible, we cannot 'do science' all the time. Rather we often must solve problems of a one-off nature, and Ehrenberg must face up to the fact that people *do* have to spend considerable time analysing single data sets.

At the other extreme, I disagree strongly with Gilmour that it is always better to have one large properly designed multicentre clinical trial. We cannot in advance think of *all* the factors which might be important, so that, even if the large design is modified sequentially, it may still not give representative results. Replicated studies are much more than a series of 'badly designed, single studies'. Rather they provide an essential check to see whether results taken by a different experimenter under inevitably somewhat different conditions (though it may be not be clear in advance how they are different) are similar. Or of course we may wish deliberately to design differences into a replicate study (Lindsay and Ehrenberg, 1993).

By the same token, I disagree even more strongly with O'Hagan's suggestion that arbitrarily splitting data is the same as collecting new data. I refer him to the comments of Bedeman, Hubbard and Miller as well as to Hirsch (1991). To put my views in Bayesian jargon, new data will not necessarily be fully exchangeable with existing data because of all sorts of unforeseen factors (a new experimenter, a different country, a different year. . .). Thus new data are vital in challenging our complacency in automatically assuming exchangeability. Hubbard's gloomy remarks on the 'bias against publishing replication research' are therefore sad. Incidentally, I think that I prefer the term 'research synthesis' to 'meta-analysis' and note that the former term is increasingly used. I also note that Bayesian updating can be regarded as another way of combining information.

While on the subject of data splitting, I must voice my support for Bedeman's comments that a 'hold-out sample' which is used to help to *choose* a 'best' model is *not* really a hold-out sample at all. Forecasting comparisons must be made on genuine *ex ante* or out-of-sample forecasts (Fildes and Pidd).

Miscellanea

Some other comments and responses in brief are as follows. In response to Copas, I confess that I also have sinned in ignoring the effects of model uncertainty and am still unclear about how much this will matter in specific situations, and what can be done about it. Hence there is the need for more research. When the data point unequivocally to a single model, the effect of model uncertainty is likely to be small (Davison) but how often does this happen? In my experience, it is rather rarely, especially in time series analysis. Nevertheless Day will be pleased to know that I am not proposing we retract all our previous analyses, even in time series analysis, though I do think that we should be more circumspect in our qualitative inferences. When one sees an example like that presented by Draper (showing that the within-sample standard deviation of standardized residuals is about *half* the corresponding jackknife value) the need for caution is clear!

Bhansali's references seem to relate to model selection rather than inference *after* model selection which is my main concern.

O'Hagan comments that the (non-Bayesian) combination of forecasts 'is not consistent'. I have no idea what he means by this. I do know that combining forecasts often works well in practice. Bayesians have an unfortunate habit of dismissing the results of alternative approaches as being 'incoherent' or 'inconsistent' without stating the assumptions or premises on which this is based.

I was also tempted to agree with Ray that 'the only sure thing is the *data*', but then realized that

(a) data often contain errors and will not then be 'sure' (and Bartholomew is right to add this to the list of sources of uncertainty) and
(b) the quote implies that we should ignore context, and I would not agree with that.

I enjoyed Lewis's forthright remarks but cannot agree that the scenario I examine in example 2 is uncommon. I am delighted to hear that his students were taught to say that more data may need to be collected when a non-significant result is obtained, but I wonder whether they were taught how much more to collect, and what to do with them when collected (e.g. combine new data with old?). Is this

This content downloaded from 98.161.230.204 on Mon, 02 Apr 2018 20:12:13 UTC
All use subject to http://about.jstor.org/terms

really an area where we are doing a thorough job? More generally I am disappointed that no-one took up the pedagogical implications of model uncertainty.

Finally, I mention two additional relevant papers which have been brought to my attention. Easterling and Anderson (1978) give results which are relevant to example 1. Altman *et al.* (1994) show that when a continuous variable is categorized into discrete groups, with groupings determined by the data to give an 'optimal' *P*-value, then the latter will (obviously?) be biased. They recommend that this procedure and the word 'optimum' should be abandoned.

REFERENCES IN THE DISCUSSION

Agiakloglou, C. and Newbold, P. (1992) Empirical evidence on Dickey–Fuller type tests. *J. Time Ser. Anal.*, **13**, 471–483.

Akaike, H. (1979) A Bayesian extension of the minimum AIC procedure of autoregressive model fitting. *Biometrika*, **66**, 237–242.

Altman, D. G., Lausen, B., Sauerbrei, W. and Schumacher, M. (1994) Dangers of using "optimal" cutpoints in the evaluation of prognostic factors. *J. Natn. Cancer Inst.*, **86**, 829–835.

Ameen, J. R. M. and Harrison, P. J. (1984) Discount Bayesian Multiprocess Models with cusum's. In *Time Series Analysis*, vol. 5, *Theory and Practice* (ed. O. Anderson), pp. 117–134. Amsterdam: North-Holland.

Arnold, S. J. (1979) A test for clusters. *J. Marktng Res.*, **16**, 545–551.

Bancroft, T. A. and Han, C.-P. (1977) Inference based on conditional specification: a note and a bibliography. *Int. Statist. Rev.*, **45**, 117–127.

Bartholomew, D. J., Hopes, R. F. A. and Smith, A. R. (1976) Manpower planning in the face of uncertainty. *Persnnl Rev.*, **5**, 1–17.

Bernardo, J. M. and Smith, A. F. M. (1994) *Bayesian Theory*. New York: Wiley.

Bhansali, R. J. (1993) Order selection for linear time series models: a review. In *Developments in Time Series Analysis* (ed. T. Subba Rao), pp. 50–66. London: Chapman and Hall.

———(1994) Asymptotically efficient autoregressive model selection for multistep prediction. To be published.

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

Box, G. E. P. and Jenkins, G. M. (1976) *Time Series Analysis: Forecasting and Control*. San Francisco: Holden-Day.

Box, G. E. P. and Kramer, T. (1992) Statistical process monitoring and feedback adjustment—a discussion. *Technometrics*, **34**, 251–267.

Breckenridge, J. N. (1989) Replicating cluster analysis: method, consistency and validity. *Multiv. Behav. Res.*, **24**, 147–161.

Chatfield, C. (1977) Some recent developments in time-series analysis. *J. R. Statist. Soc.* A, **140**, 492–510.

———(1982) Teaching a course in applied statistics. *Appl. Statist.*, **31**, 272–289.

———(1985) The initial examination of data (with discussion). *J. R. Statist. Soc.* A, **148**, 214–253.

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

Chow, G. C. (1981) A comparison of the information and posterior probability criteria for model selection. *J. Econometr.*, **16**, 21–33.

Cox, D. R. (1990) Role of models in statistical analysis. *Statist. Sci.*, **5**, 169–174.

Cox, D. R. and Snell, E. J. (1974) The choice of variables in observational studies. *Appl. Statist.*, **23**, 51–59.

———(1981) *Applied Statistics: Principles and Examples*. London: Chapman and Hall.

Davies, N. and Newbold, P. (1980) Forecasting with misspecified models. *Appl. Statist.*, **29**, 87–92.

Davies, N. and Tremayne, A. R. (1994) Exploratory techniques for identifying classes of time series models. *Report*. Department of Mathematics, Statistics and Operational Research, Nottingham Trent University, Nottingham.

Denton, F. T. (1985) Data mining as an industry. *Rev. Econ. Statist.*, **67**, 124–127.

Draper, D. (1994) Hierarchical models and variable selection. *Statistics Research Report 94:05*. University of Bath, Bath.

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

Easterling, R. G. and Anderson, H. E. (1978) The effect of preliminary normality goodness-of-fit tests on subsequent inferences. *J. Statist. Computn Simuln*, **8**, 1–11.

Efron, B. and Gong, G. (1983) A leisurely look at the bootstrap, the jackknife, and cross-validation. *Am. Statistn*, **37**, 36–48.

Ehrenberg, A. S. C. (1963) Bivariate regression analysis is useless. *Appl. Statist.*, **12**, 161–179.

———(1982) How good is best? *J. R. Statist. Soc.* A, **145**, 364–366.

———(1994) *A Primer in Data Reduction*. Chichester: Wiley.

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.

Fildes, R. and Makridakis, S. (1995) The impact of empirical accuracy studies on time series analysis and forecasting. *Int. Statist. Rev.*, **63**, in the press.

This content downloaded from 98.161.230.204 on Mon, 02 Apr 2018 20:12:13 UTC
All use subject to http://about.jstor.org/terms

Findley, D. F. (1985) Model selection for multi-step-ahead forecasting. In *Identification and System Parameter Estimation*, pp. 1039–1044. Oxford: Pergamon.
Fisher, R. A. (1934) Two new properties of mathematical likelihood. *Proc. R. Soc.* A, **144**, 285–307.
——(1966) *The Design of Experiments*, 8th edn. London: Oliver and Boyd.
——(1970) *Statistical Methods for Research Workers*, 14th edn. Edinburgh: Oliver and Boyd.
Freedman, D. (1991) Statistical models and shoe leather (with discussion). *Sociol. Methodol.*, 291–358.
——(1995) Some issues in the foundations of statistics. *Found. Sci.*, **1**, 19–83.
Gordon, A. D. (1981) *Classification: Methods for the Exploratory Analysis of Multivariate Data*. London: Chapman and Hall.
——(1994) Clustering algorithms and cluster validation. In *Computational Statistics* (eds P. Dirschedl and R. Ostermann), pp. 497–512. Heidelberg: Physica.
Green, P. J. (1994a) Discussion on Representations of knowledge in complex systems (by U. Grenander and M. I. Miller). *J. R. Statist. Soc.* B, **56**, 589–590.
——(1994b) Reversible jump MCMC computation and Bayesian model determination. Submitted to *Biometrika*.
Hand, D. J. (1994) Deconstructing statistical questions (with discussion). *J. R. Statist. Soc.* A, **157**, 317–356.
Hirsch, R. P. (1991) Letter to the editor. *Biometrics*, **47**, 1193–1194.
Hoem, J. M. (1973) Levels of error in population forecasts. *Artikle fra Statistik Sentralbyra 61*. Statistik Sentralbyra, Oslo.
Hubbard, R. and Vetter, D. E. (1995) An empirical comparison of published replication research in accounting, economics, finance, management, and marketing. *J. Bus. Res.*, to be published.
Kabaila, P. (1995) The effect of model selection on confidence regions and prediction regions. *Econometr. Theory*, **11**, in the press.
Kashyap, R. L. (1982) Optimal choice of *AR* and *MA* parts in autoregressive moving average models. *IEEE Trans. Pattn Anal. Mach. Intell.*, **4**, 99–104.
Kass, R. E. and Raftery, A. E. (1993) Approximate Bayes factors and accounting for model uncertainty in generalized linear models. *Technical Report 255*. Department of Statistics, University of Washington, Seattle.
Leamer, E. E. (1978) *Specification Searches: ad hoc Inference with Experimental Data*. New York: Wiley.
Léger, C., Politis, D. N. and Romano, J. P. (1992) Bootstrap technology and applications. *Technometrics*, **34**, 378–398.
Léger, C. and Romano, J. P. (1990) Bootstrap adaptive estimation: the trimmed-mean example. *Can. J. Statist.*, **18**, 297–314.
Leybourne, S. J. and McCabe, B. P. M. (1994) A consistent test for a unit root. *J. Bus. Econ. Statist.*, **12**, 157–166.
Lindley, D. V. (1968) The choice of variables in multiple regression (with discussion). *J. R. Statist. Soc.* B, **30**, 31–66.
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., Gavrin, J. and Raftery, A. E. (1995) Eliciting prior information to enhance the predictive performance of Bayesian graphical models. *Communs Statist. Theory Meth.*, to be published.
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. (1995) Bayesian graphical models for discrete data. *Int. Statist. Rev.*, to be published.
Marriott, J. M. and Tremayne, A. R. (1988) Alternative statistical approaches to time series modelling for forecasting purposes. *Statistician*, **37**, 187–197.
Mayer, T. (1980) Economics as a hard science: realistic goal or wishful thinking? *Econ. Inq.*, **18**, 165–178.
McIntyre, R. M. and Blashfield, R. K. (1980) A nearest-centroid technique for evaluating the minimum-variance clustering procedure. *Multiv. Behav. Res.*, **15**, 225–238.
Milligan, G. W. and Sokol, L. M. (1980) A two-stage clustering algorithm with robust recovery characteristics. *Educ. Psychol. Measmnt*, **40**, 755–759.
Mitchell, G. H. (1994) *The Practice of Operational Research*. Chichester: Wiley.
Monahan, J. F. (1983) Fully Bayesian analysis of *ARMA* time series models. *J. Econometr.*, **21**, 307–331.
Neuliep, J. W. and Crandall, R. (1993) Reviewer bias against replication research. *J. Soc. Behav. Pers.*, **8**, 22–29.
O'Hagan, A. (1995) Fractional Bayes factors for model comparison (with discussion). *J. R. Statist. Soc.* B, **57**, 99–138.
Pole, A., West, M. and Harrison, P. J. (1994) *Applied Bayesian Forecasting and Time Series Analysis*. New York: Chapman and Hall.
Poskitt, D. S. (1987) Precision, complexity and Bayesian model determination. *J. R. Statist. Soc.* B, **49**, 199–208.
Pötscher, B. M. (1991) Effects of model selection on inference. *Econometr. Theory*, **7**, 163–185.
——(1995) Comment on "The effect of model selection on confidence regions and prediction regions by P. Kabaila". *Econometr. Theory*, **11**, 550–559.
Pötscher, B. M. and Novak, A. J. (1994) The distribution of estimators after model selection: large and small sample results. *Working Paper*. Department of Statistics, University of Vienna, Vienna.
Priestley, M. B. (1980) State dependent models: a general approach to non-linear time series models. *J. Time Ser. Anal.*, **1**, 57–71.
Raftery, A. E. (1995) Bayesian model selection in social research (with discussion). *Sociol. Methodol.*, to be published.
Raftery, A. E., Madigan D. and Volinsky, C. T. (1995) Accounting for model uncertainty in survival analysis improves predictive performance (with discussion). In *Bayesian Statistics V* (eds J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith). Oxford: Oxford University Press. To be published.

This content downloaded from 98.161.230.204 on Mon, 02 Apr 2018 20:12:13 UTC
All use subject to http://about.jstor.org/terms

Ray, B. K. (1993) Modeling long memory processes for optimal long-range prediction. *J. Time Ser. Anal.*, **14**, 511–525.
Riddle, P., Segal, R. and Etzioni, O. (1994) Representation design and brute-force induction in a Boeing manufacturing domain. *Appl. Artif. Intell.*, **8**, 125–147.
Rissanen, J. (1987) Stochastic complexity. *J. R. Statist. Soc.* B, **49**, 223–239.
Roecker, E. B. (1991) Prediction error and its estimation for subset-selected models. *Technometrics*, **33**, 459–468.
Shibata, R. (1986) Consistency of model selection and parameter estimation. In *Essays in Time Series and Allied Processes* (eds J. Gani and M. B. Priestley), pp. 127–141. Sheffield: Applied Probability Trust.
Smith, M. D. (1993) Expectations of ratios of quadratic forms in normal variables: evaluating some top-order invariant polynomials. *Aust. J. Statist.*, **35**, 271–282.
Tiao, G. C. and Xu, D. (1993) Robustness of maximum likelihood estimates for multi-step predictions: the exponential smoothing case. *Biometrika*, **80**, 623–641.
West, M. and Harrison, P. J. (1989) *Bayesian Forecasting and Dynamic Models*. New York: Springer.
York, J., Madigan, D., Heuch, I. and Lie, R. T. (1995) Birth defects registered by double sampling: a Bayesian approach incorporating covariates and model uncertainty. *Appl. Statist.*, **44**, 227–242.
Zeigler, B. P. (1976) *Theory of Modelling and Simulation*. New York: Wiley.

This content downloaded from 98.161.230.204 on Mon, 02 Apr 2018 20:12:13 UTC
All use subject to http://about.jstor.org/terms