Title
Planning, Criticism, and Revision.

Journal
Journal of Applied Econometrics, 4(S)

Author
Leamer, Edward E

Publication Date
2023-02-17

JOURNAL OF APPLIED ECONOMETRICS, VOL. 4, S5–S27 (1989)

PLANNING, CRITICISM, AND REVISION

EDWARD E. LEAMER
Department of Economics, UCLA, Los Angeles, CA 90024, U.S.A.

SUMMARY

This paper presents a more complete theory of data analysis which allows for changes in the state of mind of the observer and also for approximations that limit planning costs. Discussion is included on the form that criticism should take, and the extent to which planned responses to the data can legitimately be revised after the data are reviewed. The proper role of diagnostics is discussed. Some diagnostic statistics are genuinely criticisms, but many are pre-test diagnostics that play a role in a complex multi-step method of estimation. A third category is elicitation diagnostics, which ask data-dependent questions about the prior distribution.

1. INTRODUCTION

One of the major changes that is taking place in practical econometric analysis is the increasing use of 'diagnostic' statistics which apparently serve as 'criticisms' of the model and which can stimulate model 'revisions'. The Durbin–Watson statistic and the adjusted R^2 have been used as 'diagnostic' statistics for several decades; but recently the list of diagnostics has expanded dramatically.

Diagnostics play a particularly great role in the LSE–Hendry style of applied econometrics that has been reviewed by Pagan (1987a, 1987b), and Gilbert (1988). In an application, Baba, Hendry, and Starr (1987) surround an estimated money demand function with three tests for residual autocorrelation, several Chow tests for parameter constancy, White's test of functional form misspecification–heteroscedasticity, an F-test against a more general model, the Jarque–Bera test for normality, and a test for autoregressive conditional heteroscedasticity of order r (unspecified).

The most extensive use of diagnostics is still confined to Hendry and his adherents, but the approach seems sure to spread as computer packages routinely compute these diagnostics. For example, Microfit (previously called Data-fit), a recent computer package designed by M. Hashem Pesaran and Bahram Pesaran, includes: Godfrey's test of residual serial correlation, Ramsey's RESET test of functional form, Jarque–Bera's test of the normality of regression residuals, tests for heteroscedasticity, the Chow test of the stability of the regression coefficients, Sargan's misspecification test, Sargan's test of serial correlation of instrumental variable residuals, and leverage plots.

The LSE–Hendry style of applied econometrics is distinct not just in the number of diagnostic tests that are employed but also in the method of choosing variables, beginning with a highly over-parameterized general model, shrinking it by eliminating a large number of variables, and then expanding it to include variables that were not in the initial 'general model'. Both this and the use of diagnostics seem to raise issues of statistical methodology but Baba,

0893–7252/89/0S00S5–23$11.50

Received April 1989

Hendry, and Starr (1987) claim that 'precisely how one should construct empirical models is primarily a matter of research efficiency; in principle, no method of construction need be invalid since nothing precludes an investigator from thinking of or chancing upon useful and robust relationships prior to or during data analysis'. Leamer (1985, p. 112) quips that this seems like 'a combination of backward and forward stepwise (better known as unwise) regression'.

Bruce Hill's (1980, 1988) position seems not far from Hendry's. Hill (1988), echoing Keynes (1921, Ch. 25), observes that from the Bayesian perspective 'once a model has been formulated, whether pre- or post-data, the likelihood function for the parameters of that model, conditional upon the truth of that model, does not in any way depend upon the circumstances under which that model was discovered' (Hill, 1988, p. 11). The Bayesian problem with data-instigated models, which was recognized by Keynes (1921, Ch. 25), is only how to form a prior distribution that is not 'contaminated' by the data. Hill (1988, p. 13) acknowledges that 'the difficulties are primarily psychological', but apologizes that 'the force of a Bayesian analysis of data must depend upon an agreement among scientists that specific prior distributions and likelihood functions are pertinent to the problem, and can be considered on their own merits, even after the data has [*sic*] been observed'.

In apparent contrast to Hill and Hendry, I believe that the use of diagnostic statistics does present a challenge to statistical theory, classical or Bayesian. Traditional statistical theory deals with the evaluation of *planned* responses to *hypothetical* data sets. Indeed it is impossible to compute sampling properties without a set of plans indicating the response to the data for every conceivable data set. The use of a diagnostic statistic to criticize a model is an advance announcement that the planned responses are not fully committed and may be revised when the actual data are observed.

However, very few if any of the diagnostics that are traditionally employed in the econometrics literature are criticisms in my sense of precipitating an unplanned, unpredictable response to the data. Many are 'pre-test' diagnostics that play a part in a complex multi-stage method of estimation of a very general model. A statistic is a pre-test diagnostic if both the general model and the response to the data can be fully defined and programmed before the data are observed. The proper evaluation of pre-test diagnostics involves either the study of the sampling properties of these complex procedures or the search for a prior distribution that could partially justify them.

But not all responses can or should be planned. The actual response to real data can differ from the planned response to hypothetical data for at least two reasons. The first reason is that the desired response to the data depends on the state of mind of the observer, which can change with changes in mood and expertise. Second, even if there were no variability in the state of mind, a complete set of plans applicable to every conceivable data set is very costly to formulate. Plans accordingly will be formulated only for data sets that are regarded to be probable, and responses to improbable data sets will be formulated only if and when they are observed. Contrast for example the scatter of observations in Figures 1 and 2. Though the t-statistics and the R^2 values are the same, the messages seem very different, and the plan of regressing y on x seems not very wise for the scatter in Figure 2. One would not sensibly have planned for this possiblity since it seems so remote, but once these data are observed the original plan to run a regression seems highly inappropriate, and cries out for revision.

This paper presents a more complete theory of data analysis which allows for changes in the state of mind of the observer, and also for approximations that limit the planning costs. Discussion is included on the form that criticism should take, and the extent to which planned responses can legitimately be revised. The proper role of diagnostics is discussed. I argue that

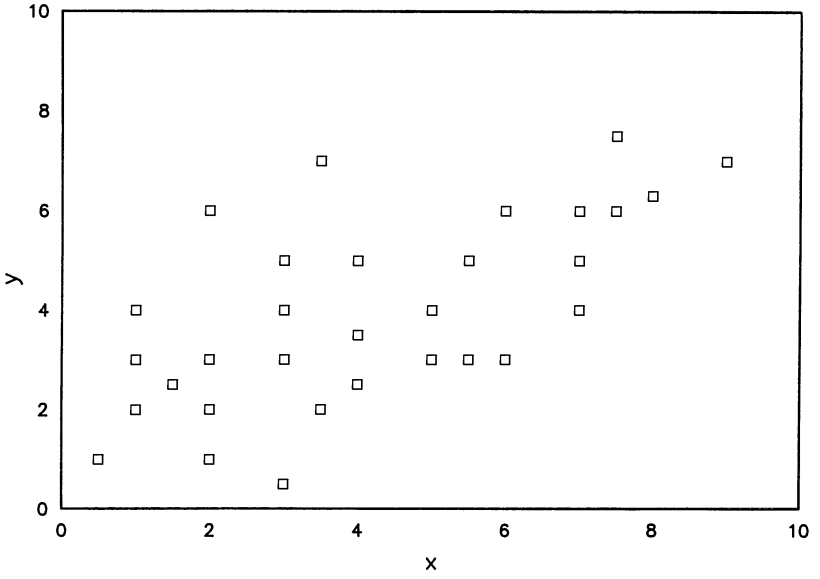

Figure 1. A probable scatter

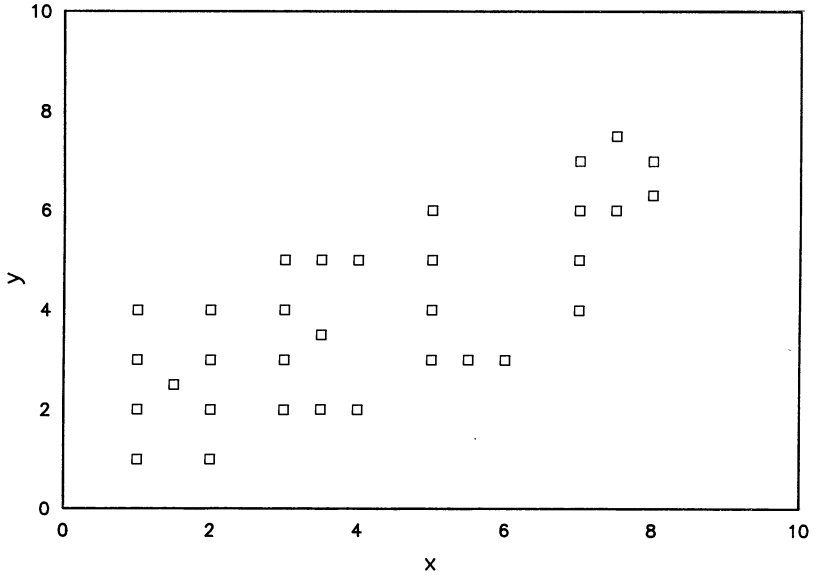

Figure 2. An improbable scatter

there are three different categories of diagnostics:

1. Pre-test diagnostics which select between a pair of alternative estimates.
2. Elicitation diagnostics which indicate if the inferences are sensitive to the choice of prior distribution, and which call for a more accurate measurement of the state of mind.
3. Criticisms which suggest a 'fundamental' change in the model and/or prior distributions.

It is possible to have a theory of pre-set diagnostics and a theory of elicitation diagnostics, but I know of no approach that allows a theory of criticism. Selecting the form that genuine

criticism should take is by its very nature an unsolvable problem, since the solution must be based on the corrective action that a successful criticism should precipitate, yet knowledge of that action means that criticism in my sense has not taken place. Furthermore, in the absence of a completely appropriate theory of criticism, the adjustments to the inferences that are required to correct for both successful and unsuccessful criticisms must remain to some extent *ad hoc*, though I will argue that there are adjustments that are sensible.

2. CRITICISM AND REVISION

Essentially all of statistical theory is concerned with the evaluation of *planned* responses to *hypothetical* data sets. A Bayesian approach allows these idealized plans to depend on the state of mind of the observer, but even the Bayesian theory ignores the possibility that this state of mind responds to influences other than previous data sets. Actual responses, Bayesian or not, may be quite different from these idealized plans, because the state of mind of the observer may change substantially for reasons not adequately captured in the Bayesian model.

The model of data analysis that is presented here allows for changes in the state of mind of the observer. This model of data analysis is broad enough to include 'exploratory' data analysis as well as 'confirmatory' data analysis. Confirmatory data analysis is characterized by a substantial commitment to the original planned responses. Exploratory data analysis has weak plans, if any, and may use displays and diagnostics to suggest the 'model' on which a response might be formulated. The problem with exploratory data analysis is that there is a substantial tendency to overfit, to see patterns in the data set that are not really there. A theory of exploratory data analysis indicates how the overfitting problem can be avoided.

Incidentally, the word 'response' is here used in a general sense. A response to a data set is sometimes a decision (for example: Act as if a hypothesis were true) but more often is a judgement (for example: The data suggests that the value of β is close to zero). The response to the data depends on the state of mind of the observer. You and I may see the same observations but draw very different conclusions from them. For that matter, you may analyse a data set today and draw very different conclusions than you did last week.

The three determinants of the state of mind are former observations, 'mood' and 'expertise'. Bayesians have a well-developed theory of how past observations affect the prior distribution, which can be important for interpreting the current data set. The classical problem of pooling different data sets yields essentially the same results. But neither the traditional Bayesian theory nor the classical pooling theory admits the influence of variable factors other than data.

The two non-data components of the state of mind that are considered here are the 'mood', which is defined as those random effects that are impermanent and stationary, and the 'expertise', which is defined to be those random effects that are permanent and nonstationary. Emotions, among other things, can cause changes in mood. Reasoning and flashes of insight can cause changes in expertise. Social interactions are a very important source of changes in mood and expertise. Fads influence mood; fashions may be either permanent or impermanent.

Reference to the problem of estimating/discovery of the functional form of a relationship has helped to organize my thoughts about these difficult issues. It seems useful to distinguish six different ways in which a functional form is selected:

Planned responses:

1. Stepwise estimation in which a quadratic term is included if its t-value exceeds some critical level.

2. Visual inspection of the scatter of points to decide whether to include a quadratic term or not.
3. Visual inspection of the scatter of points to decide which functional form to estimate (linear, quadratic, log-linear, etc.).

Unplanned Responses:

4. Visual inspection of the scatter of points without the expressed intent of considering other than a linear form, but the discovery of evidence of curvature that leads to the estimation of a nonlinear function.
5. Discovery of curvature in the scatter of points that stimulates a theoretical insight and alters the level of 'expertise' to such an extent that plans in the future allow for this kind of curvature.
6. Use of the *t*-statistic on the quadratic term as a diagnostic statistic to suggest an unspecified alteration of the model such as the inclusion of some new variable.

The line separating (3) from (4) separates settings in which a planned response to the data is carried out from settings in which the original plans are weak and major revisions occur. The former methods will be said to be 'above the line'; the latter are 'below the line'. The former are 'confirmatory'; the latter 'exploratory'.[1] The former methods ought to be subjected to the critical scrutiny of traditional statistical theory; the latter may (or may not) be free from that form of scrutiny.

'Diagnostic' statistics could be said to be used in all six cases. In the first three cases the diagnostics statistics are part of a multi-step method of estimation, and the overall method of estimation should be evaluated in the traditional manner. But when a diagnostic statistic is used to stimulate an unpredictable response, the estimation method falls 'below the line' and resists evaluation of any kind.

More generally, the analysis of data is summarized in the following data:

PLANNING RESPONSES	
Precise responses	*Imprecise responses*
Judgements	Confusion
Actions	Indecision
CRITICISM	
REVISION	

Below the line in the middle of this diagram are the activities of criticism and revision.[2] The

[1] Hendry uses the word 'evaluation' and 'design' which might be interpreted as corresponding to my use of 'above the line' and 'below the line'. But much of that Hendry calls evaluation is or should be 'above the line' in the sense that the response to the data set can be fully planned. His words might correspond better with 'estimation' and 'hypothesis testing'.

[2] The phenomena of planning, criticism, and revision have analogies in computer programming, contracting, and human learning. A computer program is designed to work well for normal inputs, and will signal the user with a diagnostic message when inputs are unusual and difficult to process. Then the user may want to revise the plans by finding or building a more suitable program. Contracts are usually written contingent on expected events. When unexpected events occur, either party can appeal to the court to have the contract rewritten. The economic pathology of unemployment may be linked to conditions that are unusual, but not so unusual that the contract is rewritten. When humans learn new tasks, the discretionary activities of criticism and revision are frequent but, with time, responses become 'programmed' and automatic. Instincts are genetic programs that can be overridden as circumstances require.

planning of responses occurs 'above the line'. A response may be either a judgement or an action; a response may be either precise or imprecise. 'Confusion' is the state of mind that is present when a judgement is imprecise; 'indecision' is the outcome of an imprecise action. 'Confusion' and 'indecision' are important subjects that have received inadequate attention in the theory of inference, exceptions including Leamer (1987). My concern here, however, is with the separation between the activity of planning from the activities of criticism and revision.

It should be clearly understood that both classical and Bayesian inference require *complete commitment* to the initial plans, and disallow criticism and revision. Classical inference, which refers to sampling properties, requires a complete commitment to the initial plans since *sampling properties* can be computed only if the response to every conceivable data set is known. A Bayesian treatment also implicitly requires a complete commitment to the initial plans in the sense that the plans are a consequence of the choices of prior and sampling distributions, which choices are made after the data are observed only with discomfort to the data analyst and suspicion by the reader of results.

By its very nature we cannot know the form that criticism should take, but it is clear that both successful and unsuccessful criticisms have implications for drawing conclusions from the data. The phenomenon of criticism, even when it does not lead to a revision, reveals that there is a lack of complete commitment to the assumptions that underlay the original plans. This lack of complete commitment requires some alteration of the plans; for example, the standard errors of the coefficients should be enlarged to reflect the fact that there surely are omitted variables that cause bias in the estimates. When the criticism is successful there is a double-counting problem, because the data are used once to alter the assumptions, and then again to estimate the parameters, as if these were the assumptions that were used from the beginning.

However, the distinction between planned and unplanned responses is not obvious. Responses that are predictable but not explicitly and consciously planned can be said to be implicit plans. Genuine revisions are unpredictable and quite rare. For example, I don't know in advance exactly how a multidimensional stepwise regression program will work in particular settings, but I know it will always do the same thin and in that sense is predictable. This computer program will be regarded to embody a complete plan, even though many aspects of the plan do not have my conscious review.

Human intervention is necessary, but not sufficient to establish that a revision has occurred. It depends on whether the intervention is predictable. For example, my detection of curvature in a scatter of points is predictable, and simulates the predictable response of including a parameter that allows curvature. In this case the human and the computer combine to carry out the implicit plan. This is essentially the same as a stepwise program that adds a nonlinear term when it is 'significant'.

3. ELEMENTS OF A THEORY OF DATA ANALYSIS

The model of data analysis that is presented here will use the following notation:

Y_t = a sequence of data matrices;
S_t = current state of mind;
M_t = the mood, a stationary random process;
E_t = expertise, a nonstationary random process.

The elements of the model follow:

Data Distribution

The data are assumed to have been drawn from a distribution that depends on β_t, the parameter of interest, and ϕ_t, a vector of 'nuisance' parameters:

$$F_Y(Y_t \mid \beta_t, \phi_t)$$

This notation allows for time-series dependence in the sampling process if the vector ϕ_t includes other data sets such as Y_{t-1}. This assumption of the existence of a data distribution is essentially vacuous if there is sufficient freedom to select the nuisance parameter.

Prior Distribution

The nuisance parameter ϕ_t is assumed to be infinite dimensional to allow for essentially any assumption about the way the data are generated. In order to make inferences about β_t, the values that ϕ_t can take on must somehow be limited. Often this is done in practice by restricting all but a few of the components of ϕ_t to take on preselected values. This can be regarded as a special kind of prior distribution which is dogmatic about some of the components of ϕ_t and diffuse about the others. For purposes of discussion it will be assumed more generally that there exists an implicit or explicit prior distribution that indicates the probable regions for (β_t, ϕ_t):

$$G(\beta_t, \phi_t \mid S) \quad (\beta_t, \phi_t) \in \Phi_t$$

This prior distribution depends importantly on the state of mind, S. I will not assume that this distribution can be elicited without error, and it may be impossible to base a data analysis on the 'true' prior distribution.

The State of Mind

The state of mind depends on past observations, on the mood M and on the expertise E of the observer:

$$S_t = f(Y_{t-1}, T_{t-2}), \ldots, Y_1, M_t, E_t).$$

The Mood

The mood is a stationary stochastic process which for purposes of discussion is assumed to depend only on the current observations and a white noise random variable ε_t:

$$M_t = g(Y_t, \varepsilon_t)$$

The mood may vary with the personal emotional state of the observer and may also be influenced by social interactions (fads and fashions). The mood may be very different for analysis of hypothetical data sets than for analysis of real data sets, since the latter are treated with greater thought and care. The formal elicitation of a prior distribution can alter substantially the mood of the observer. This can increase the amount of care but also cause a high level of commitment to the current model; more on this below.

The Expertise

The level of expertise is a nonstationary stochastic process which for purposes of discussion will

be written as:

$$E_t = E_{t-1} + h(Y_t, e_t)$$

where e_t is a white noise random process. Expertise changes with contemplation, study, enlightenment, and training, among other things. The process is nonstationary, since once a level of expertise is obtained there is no tendency to return to the former level; indeed the process may be irreversible.

The Idealized Response

Given the state of mind of an observer, there is an idealized response to the data. This idealized response can be found using either a Bayesian or a classical approach, although the solutions may differ:

$$\rho(Y_t \mid S)$$

The traditional theory of data analysis is almost completely a theory of ideal, planned, fully committed responses, not a theory of actual responses. Plans by definition are formulated before the real data are observed. A planned response to hypothetical data will differ from the actual response to real data for at least two reasons. The first is that, at the time the plan is formulated, the future state of mind is uncertain and can be forecast only with error. Secondly, even if there were no variability in the state of mind, a complete set of plans applicable to every conceivable data set is very costly to formulate. For example, it may be infinitely costly to elicit the prior distribution fully and without error. Plans accordingly will be formulated only for data sets that are regarded to be probable. Responses to improbable data sets will be formulted only if and when these improbable data are observed. If a plan is applicable for a range of possible data sets, then it will be said to be a 'wide' plan. A plan will be wide in a setting in which there is a great deal of knowledge about the process that generates the data; that is to say when there is little variability in the state of mind.

The planned and actual responses that are made can only approximate the ideal response function. The sense in which the response approximates the ideal is most easily discussed from the Bayesian perspective which can base a data analysis on an approximate prior distribution. This Bayesian perspective will now be used, and a sampling theory treatment will be presented subsequently.

Approximate Prior Distribution

Using the Bayesian approach, the formulation of the idealized response function would require the elicitation of the prior distribution over the infinite dimensional parameter space Φ_t, a task that would require an unlimited amount of time. Instead, an approximate prior distribution is formulated. First the parameter space Φ_t is abbreviated, and then a mathematically convenient approximate prior distribution is formed over the abbreviated parameter space:

$$\tilde{G}(\beta_t, \phi_t \mid S) \quad (\beta_t, \phi_t) \in \tilde{\Phi}_t(S)$$

An equivalent characterization of this approximate prior distribution uses the device of an approximate state of mind:

$$\tilde{G}(\beta_t, \phi_t \mid S) = G(\beta_t, \phi_t \mid \tilde{S}) \quad (\beta_t, \phi_t) \in \Phi_t$$

where $\tilde{S}$ is a state of mind similar to S but one that implies an abbreviated parameter space and a simple data analysis.

Planned Response

The planned response is selected before the current data are observed. A Bayesian plan is formulated based on a prediction of the state of mind:

$$P(Y_t) = \rho(Y_t \mid \hat{S}_t)$$

where $\hat{S}_t$ is a prediction of S_t given the information available in period $t-1$. This prediction is selected to imply a relatively simple data analysis. Note that if mood and expertise are stochastic and affect the state of mind, then the plan is also stochastic in the sense that the same response is not always made to the same data set.

The Actual Response Function

The planned response is carried out if there is little change in the state of mind, but otherwise a revision occurs:

$$R(Y_t) = \begin{cases} \rho(Y_y \mid \hat{S}_t) & \text{if } \mid \hat{S}_t - \tilde{S}_t \mid < w(\hat{S}_t) \\ \rho(Y_t \mid \tilde{S}_t) & \text{otherwise} \end{cases}$$

where $\mid \hat{S}_t - \tilde{S}_t \mid$ is a measure of the difference between the predicted state of mind $\hat{S}_t$ and the approximate state of mind $\tilde{S}_t$ measured after the data are observed.

The function $w(\hat{S}_t)$ is the 'width of the plan'. If the approximate state of mind $\tilde{S}_t$ is a known function of the data Y_t, then the width of the plan can be characterized in terms of 'the' probability that the plan will be carried out. If this probability is evaluated with respect to the approximate prior distribution, it is likely to underestimate the true probability of a revision. More of this below in the discussion of the choice of significance level for diagnostic statistics.

4. SPECIAL CASES

The following are special cases:

Textbook Classical Theory

Most programs for electronic data analysis cannot alter themselves, and when confronted with the same inputs always produce the same outputs. The absence of memory and randomness means that a computer program cannot have a variable state of mind. In that event the width of the plan is infinite in the sense that the actual response and the planned response are necessarily the same and equal to the idealized response for one special state of mind S_0:

$$R(Y_t) = P(Y_t) = \rho(Y_t \mid S_0)$$
$$w(\ ) = \infty$$

A simple example is least squares regression which uses inputs $Y_t = [y_t, X_t]$ to form an estimate:

$$R(Y_t) = (X_t' X_t)^{-1} X_t' y_t$$

Another example is stepwise regression which can be written as

$$R(Y_t) = (X_{1t}' X_{1t})^{-1} X_{1t}' y_t$$

where the included variables X_{1t} are columns of X_t selected depending on the data:

$$X_{1t} = q(X_t, y_t)$$

Classical Model with Inputs Selected Non-stochastically Without Reviewing the Current Data

The textbook classical model of data analysis ignores altogether the fact that a human has to write the computer program and select the inputs. A real data analysis must therefore be viewed as the output of a dual effort by human and electronic computer. No sharp distinction should be made between responses that are carried out completely by electronic computers, and responses that are partly selected by a human computer. For example, I might use stepwise regression to decide if a function is quadratic or not: if the t-value on the quadratic term exceeds some critical value, then the electronic computer will include the quadratic; otherwise, it will be excluded. This is not fundamentally different from deciding to include a quadratic term if something looks 'suspicious' in the scatter of observations (or plot of the residuals).

Given that a human being must be involved in a data analysis, the closest one could come to the ideal classical model is to have the human write the computer program and select the inputs into the electronic computer without reviewing the current data and without influence of the stochastic elements in mood and expertise. An equivalent would be a computer program with memory. In terms of the model, this amounts to selecting a predicted state of mind that does not depend on the mood or the expertise and a plan with an infinite width:

$$\begin{aligned} R(Y_t) = P(Y_t) &= \rho(Y_t \mid \hat{S}_t) \\ \hat{S}_t &= f(Y_{t-1}, Y_{t-2}, \ldots, Y_1). \\ w(\ ) &= \infty \end{aligned}$$

Here again, the planned response and the actual response are identical.

Classical Model with Stochastically Selected Inputs

In practice it is unlikely that a human could approximate an electronic computer and make choices that do not depend at all on mood and expertise. For example, stepwise estimation of a quadratic equation will always produce the same estimated model, linear or quadratic, when excited by the same data set. But a human observer of a scatter of observations will sometimes include the quadratic term, but sometimes will not. This makes the response have a random component. It is as if the stepwise computer program were to use a stochastic critical value to determine if the quadratic term should be included or excluded.

In terms of the elements of the model, this requires only that we allow the past levels of mood and expertise to affect the response function:

$$\begin{aligned} R(Y_t) = P(Y_t) &= \rho(Y_t \mid \hat{S}_t) \\ \hat{S}_t &= f(Y_{t-1}, Y_{t-2}, \ldots, Y_1, M_{t-1}, E_{t-1}). \\ w(\ ) &= \infty \end{aligned}$$

Multi-step, Infinite Width Plans

One aspect of exploratory data analysis is that the width of the plan is zero. Many methods of estimation masquerade as exploratory by pretending to have plans with narrow or zero

widths, when in fact the widths are infinite. For example, stepwise regression is an example of confirmatory data analysis that can be written in a form which makes it appear to have an exploratory component:

$$R(Y_t) = \begin{cases} (X_t' X_t)^{-1} X_t' y_t & \text{if } F(y_t, X_t) < w \\ (X_{1t}' X_{1t})^{-1} X_{1t}' y_t & \text{otherwise} \end{cases}$$

where F is the F-statistic for testing if the variables X_{2t} belong in the equation. One might be tempted to say that the planned response to the data is to include all the variables in the regression, but if the data are 'unusual' in the sense that the F statistic is small, then the plan is revised and only a subset of variables is included.

In this case of stepwise regression, however, the planned response to the data set is complete and fully carried out. Stepwise regression should be viewed as a form of confirmatory data analysis, and subjected to the same kind of critical scrutiny as other confirmatory analyses. Either sampling properties should be determined, or the implicit prior distribution should be unearthed. If the Bayesian approach is taken, it is natural to assume that the implicit prior distribution summarizes the notion that the subset of variables X_{2t} might be neglected because they have coefficients close to zero.

Apparently Exploratory Data Analysis with Implicit *ex-post* Plans

The cost of planning can be completely avoided if the width of the plan is genuinely set to zero. In that event a response need be formulated only for the actual data set once it is observed, not for all hypothetical data sets. The distinction between exploratory and confirmatory data analysis might be based on the width of the plan. A fully confirmatory data analysis occurs when the width is infinite. An exploratory data analysis might be said to occur when the width of the plan is zero.[3] But the proper distinction between exploratory and confirmatory data analysis cannot be made on the basis of the apparent width of the plan only, since plans may be implicit yet still be said to exist and subject, in principle, to the same kind of scrutiny as explicit plans.

Consider again the example in which one looks at a scatter of points to decide what functional form should be estimated, not necessarily committed to choosing between the linear or quadratic forms. Is this genuine exploratory data analysis? Not if the subject is merely carrying out an implicit plan. In principle we could find out what the plan is by confronting the subject with a sequence of hypothetical scatters of observations, and asking if the quadratic term should be included. This is analogous in inputting a sequence of data sets into a stepwise regression program to see if the program selects or omits the quadratic term. One difference is that there is variability in response of the human that is not normally present in the computer. This makes the plan stochastic. Another big difference is that the response of a human to hypothetical data sets may be very unlike the response to real data sets.

The formal model of apparently exploratory data analysis is:

$$\begin{aligned} R(Y_t) &= \rho(Y_t \mid \tilde{S}_t) \\ \tilde{S}_t &= m(Y_{t-1}, Y_{t-2}, \ldots, Y_1, \varepsilon_t, e_t). \\ w(\,) &= 0 \end{aligned}$$

[3] These words, 'exploratory' and 'confirmatory', are traditional. The word 'exploratory' evokes the image of an explorer entering uncharted territory with little or no preconceived idea about what is to be found there. The word 'confirmatory' evokes no similarly strong image. What is one called who uses a map for navigation? A traveller? How about 'navigatory' data analysis?

where $m(\ )$ is a measurement function that depends on the random components of mood and expertise. The difference between this and the classical model with a stochastic plan is only the existence of a complete set of explicit plans in the former case, and the complete absence of same in the latter case. But the plan does exist implicitly, even though it is not articulated or programmed. And it can in principle be uncovered by an experiment in which the observer is confronted with a sequence of observations Y. Once uncovered, it can be subjected to the traditional kinds of scrutiny.

Exploratory Data Analysis

What makes a data analysis genuinely exploratory? It cannot be merely the width of the plan. As I see it, *we should reserve the word 'exploratory' to those response functions which are not 'idealized' responses for some state of mind*. Responses that are unplanned, but are nonetheless idealized response functions for some state of mind, are apparently exploratory, not genuinely exploratory. Apparently exploratory analyses could and probably should be scrutinized in the traditional way; genuine exploratory analyses, however, may require some major amendments to our theories of inference.

Genuinely exploratory data analysis occurs when the approximate state of mind on which the data analysis depends is a function of the current data. This would occur either because the mood and expertise depend on the current data, or because the approximation to the current state of mind depends on the current data.

The model for genuinely exploratory data analysis is:

$$\begin{aligned} R(Y_t) &= \rho(Y_t \mid \tilde{S}_t) \\ \tilde{S}_t &= f(Y_t, Y_{t-1}, Y_{t-2}, \ldots, Y_1, \varepsilon_t, e_t). \\ w(\ ) &= 0 \end{aligned}$$

In this case the response function $R(Y_t)$ is not equal to an idealized response $\rho(Y_t \mid S)$ for any state of mind S because the data Y affect the state of mind S on which the response is based. *The basic problem with exploratory data analysis is that the data play two roles in the analysis, one to determine the state of mind (instigate a hypothesis) and the other to select a response given this state of mind*. The solution to this problem of double counting is proper policing of the inferences to make the response function conform as closely as possible to an idealized response for some state of mind. More on this below.

From the standpoint of an outside observer, who can see the response but not the logic for it, it is impossible to distinguish exploratory from confirmatory data analysis with implicit plans. A clear distinction could be made if the plan were required to be fully articulated before the data were observed. It is also possible to test if the response to hypothetical data sets is the same as the response to real data sets. Complications would arise, however, when there are changes in expertise.

Ideal Bayesian Model

Bayesian programs have prior information as inputs. These define the state of mind, which is a function of past observations, and time, in the sense that the prior for analysing Y_t depends on past observations and on the process that is observed at time t. Thus:

$$\begin{aligned} R(Y_t) &= P(Y_t) = \rho Y_t \mid \hat{S}_t) \\ \hat{S}_t &= S_t = f(Y_{t-1}, Y_{t-2}, \ldots, Y_1, t) \\ w(\ ) &= \infty \end{aligned}$$

meaning that the true state of mind depends on the sequence of past observations and on the process that is currently being analysed, and this state of mind can be perfectly measured. This Bayesian model is like the nonstochastic classical model but with a well-developed theory of the state of mind.

Here is an example: suppose that a sequence of vectors is observed that are generated by regression functions:

$$y_i = X_i\beta_i + u_i$$
$$y_i \sim N(0, \sigma_i^2 I)$$
$$\beta_i \sim N(\beta, V_{it})$$

where the last assumption indicates the relationship between the coefficients of interest and past coefficients. Then we can substitute to obtain

$$y_i = X_i\beta + e_i$$
$$e_i \sim N(0, \sigma_i^2 I + X_i V_{it} X_i')$$

The posterior mean of β_i can then be written in the notation of the state of mind as

$$R(Y_t) = (X_t' X_t + S_{1t})^{-1}(X_t' y_t + s_{2t})$$

where $S_{1t}^{-1} s_{2t}$ is the prior mean vector and S_{1t} is the prior precision matrix. These components of the state of mind are dependent on past observations according to:

$$S_{1t} = \sum_{i=1}^{t} X_i' [\sigma_i^2 I + X_i V_{it} X_i']^{-1} X_i$$

$$s_{2t} = \sum_{i=1}^{t} X_i' [\sigma_i^2 I + X_i V_{it} X_i']^{-1} y_i$$

Practical Bayesian Model: Precommitted Prior Distribution

In practice, however, there is a substantial amount of whimsy in defining the prior distribution, which anyway is only an approximation to the state of mind. A model of practical Bayesian analysis in which the prior distribution is selected before reviewing the data is thus:

$$R(Y_t) = P(Y_t) = \rho(Y_t \mid \hat{S}_t)$$
$$S_t = f(Y_{t-2}, Y_{t-1}, \ldots, Y_1, M_t, E_t, t)$$
$$\hat{S}_t = m(E(S_t), M_{t-1}, E_{t-1})$$
$$w = \infty$$

where m is a measurement function. Here the response depends on a *measurement* of the *expected* state of mind. The accuracy of the measurement of this expected state of mind depends on the mood and expertise of the observer.

Bayesian Analysis with an Implicit, *ex post* Plan

The initial prior distribution that was formed before the data were observed may seem undesirable once the data are observed. The selection of a prior distribution after observation of the data could be based on any of three assumptions regarding the effect of the current data on the state of mind and its measurement. These three cases are:

1. Neither the state of mind not its measurement depends on the current data.

2. The state of mind does not depend, but its measurement does depend, on the current data.
3. The state of mind, and consequently its measurement, depend on the current data.

The second two cases lead to what I would call 'exploratory data analysis'; the first is not exploratory. An example of the first case is offered in the subsequent section on Bayesian diagnostics, in which I suggest eliciting the prior distribution carefully only if an affirmative answer is given to the question: 'Is the prior variance greater than c?' where the number c is selected after the data are observed in such a way that an affirmative answer justifies the approximation that the prior variance is infinite. I argue that the answer to this question is not likely to be (greatly) affected by the fact that c depends on the data, and consequently this does not give rise to the double-counting problem of exploratory data analysis.

This case in which the prior distribution is elicited after the data are observed but not (substantially) dependent on the data takes the same form as apparently exploratory data analysis from the classical perspective:

$$\begin{aligned} R(Y_t) &= \rho(Y_t \mid \tilde{S}_t) \\ \tilde{S}_t &= S(Y_{t-1}, Y_{t-2}, \ldots, Y_1, \varepsilon_t, e_t). \\ w(\ ) &= 0 \end{aligned}$$

Note that here the response function is an ideal response for a state of mind that is measured after the data are observed. This is not what I call exploratory analysis, even though the width of the plan is zero.

Exploratory Bayesian Analysis

The other two possibilities in which the measurement of the state of mind depends on the current data fall under the heading of exploratory data analysis because the response function is not an idealized response for any state of mind. The first model of exploratory data analysis has the true state of mind independent of the current observations, but has the measurement of some influence of the current observations:

$$\begin{aligned} R(Y_t) &= \rho(Y_t \mid \tilde{S}_t) \\ \tilde{S}_t &= m(S_t, M_t, E_t, Y_t) \\ S_t &= f(Y_{t-1}, Y_{t-2}, \ldots, Y_1, \varepsilon_t, e_t). \\ w(\ ) &= 0 \end{aligned}$$

The other model has the state of mind as well as its measurement dependent on the current data set

$$\begin{aligned} R(Y_t) &= \rho(Y_t \mid \tilde{S}_t) \\ \tilde{S}_t &= m(S_t, M_t, E_t, Y_t) \\ S_t &= f(Y_t, Y_{t-1}, Y_{t-2}, \ldots, Y_1, \varepsilon_t, e_t). \\ w(\ ) &= 0 \end{aligned}$$

These two cases allow the data to play two roles in the analysis; once to affect the measured prior distribution and again to affect the inferences given the measured prior distribution. Some kind of adjustment to the inferences is required to correct for the possibility of double counting, and to limit the chance of overfitting (seeing patterns in a random data set).

5. SAMPLING PROPERTIES OF RESPONSE FUNCTIONS: CLASSICAL DESIGN

Mood and expertise cause great problems for classical inference which attempts to base the choice of response function on its sampling properties. If the response function is simple, like ordinary regression, then its sampling properties such as bias and variance can be determined mathematically. If the response function is more complex, like stepwise regression, but programmable on an electronic computer, the function can be plotted and its sampling properties can be established by Monte Carlo methods. If a human, subject to mood swings and accumulation of expertise, shares the choice of response with an electronic computer, it may also be possible to determine the joint response function and to estimate its sampling properties using Monte Carlo experiments in which the human-cum-electronic computer is excited by a sequence of randomly chosen inputs and the corresponding responses are tabulated. Randomness in the human response is clearly allowable with this experimental approach if there is no intersample dependence in the response. If there is intertemporal dependence, the experimental approach may uncover it, but sampling properties will not be clearly defined.

Consider, for example, the problem of estimating the mean μ_i of a sequence of normal populations each with variance one. The following are three estimators suggested by the Bayesian tradition. Each is a weighted average of the sample means m_i and some other value which can be thought to be the location of the prior distribution.

$$\begin{aligned}
\mu_1 &= wm_i & 0 \leqslant w \leqslant 1 & \\
\mu_2 &= wm_i + (1-w)\varepsilon_i & 0 \leqslant w \leqslant 1 & \varepsilon_i \sim N(0,1) \\
\mu_3 &= wm_i + (1-w)\sum_{j \leqslant i} e_j & 0 \leqslant w \leqslant 1 & e_i \sim N(0,1)
\end{aligned}$$

The location of the prior for the first estimator is always zero. The location of the prior for the second estimator is a stationary random variable, suggestive of changes in mood. The location of the prior for the third estimation is a nonstationary random variable, suggestive of changes in expertise. What are the sampling properties of these estimators? We can all agree that the first estimator has mean $w\mu_i$ and variance w^2/n where n is the sample size. The second estimator might be said to have mean $w\mu_i$ and variance $w^2/n + (1-w)^2$. Or it could be said that, conditional on ε_i, the mean is $w\mu_i + (1-w)\varepsilon_i$ and the variance is w^2/n. The question that must be confronted is whether changes in mood should be embedded in the sampling error or not. This question is made more pointed by reference to the third estimator which might be said to have mean $w\mu_i$ and variance $w^2/n + (1-w)^2 n_i$ where n_i is the number of means that have been observed, or alternatively could be said to have mean $wm_i + (1-w)\sum_{j \leqslant i} e_j$ and variance w^2/n. Yet a third alternative is to condition on everything that is given up to the ith mean. Then this third estimator could be said to have mean $wm_i + (1-w)\sum_{j \leqslant i-1} e_j$ and variance $w^2/n + (1-w)^2$.

The point that this example makes is that the conceptual experiment of repeated sampling that underlies classical inference can be ambiguous. Then the ranking of alternative estimators can also be ambiguous. My own instinct here would be to embed changes in mood in the sampling distribution, but not changes in expertise. If the state of mind on which a data analysis rests does not depend on mood or expertise, then a fully nonrandom response function can be selected before the latest data are observed, and sampling properties of this response function can be straightforwardly established. These sampling properties remain relevant after the data are observed because the data do not affect the sampling properties that the observer considers

relevant. When the state of mind is variable, so too are the relevant sampling properties. Which sampling properties should dictate the choice of procedures?.

In addition, there is a substantial problem in determining what the response function really is when it includes a random component chosen by a human. If the response were fully selected by an electronic computer with a random component it would be possible to input repeatedly the same data set to see how the computer would respond. For example, a stepwise regression program could have a random critical value for the t-statistic that selects the variables to include in the equation. The sampling properties of this response function can be established by Monte Carlo methods.

But a computer does not distinguish real from hypothetical situations. Humans do. What would you do if an attractive stranger proposed a rendezvous? Your answer to the hypothetical question may be very different from your response to a real proposal. Or I might ask you, if you observed a particular scatter, would you think the model to be linear or quadratic. Your answer to this hypothetical can be very different from your response to real data for a variety of reasons, one of which is that you treat the real situation with greater care and thought. To use my language, the mood that you approach a hypothetical data analysis may be very different from the mood that you approach a real data analysis.

Anyway, is it really sensible to try to find the sampling properties of an estimator that is partly selected by a human? Who is going to sit still for this?

My conclusion: sampling theory is not very useful for selecting responses to real data sets except in those cases in which the state of mind is perfectly predictable and a fully committed set of plans can be formulated before the data are observed. These cases may be more prevalent than you might imagine. Many diagnostic statistics precipitate a predictable response, and cannot be said to affect the state of mind of the observer in the sense that I have defined. These diagnostics form part of a complex multi-step method of estimation which ought to be scrutinized in the traditional way: either sampling properties should be determined, or the implicit Bayesian prior distribution should be unearthed.

6. DIAGNOSTIC STATISTICS

The theory outlined in Section 3 allows diagnostic statistics to play three different roles:

1. A 'pre-test diagnostic' may be used to select between a pair of alternative estimates.
2. An 'elicitation diagnostic' indicates if the inferences are sensitive to the choice of prior distribution, and may call for a more accurate measurement of the state of mind.
3. A 'criticism' may suggest a change in the original model/state of mind.

Each of these is now discussed.

Diagnostics as Part of a Multi-step Planned Response

Suppose that the response to a 'bad' Durbin–Watson statistic is only to correct for first-order serial correlation. Technically, this is the same as stepwise regression in which a variable is added to the model if it is sufficiently correlated with the estimated residuals. A t-statistic on this potential variable could serve as the 'diagnostic', indicating the need to add this other variable. This type of diagnostic is just part of a complex method of estimation. It really should not be called a diagnostic at all. The complex method of estimation should be subjected to the traditional scrutiny: either sampling properties should be established, or the prior distribution that underlies the estimate should be disclosed.

These pretest diagnostics are, I believe, the most prevalent in practice. Usually peculiarity in a diagnostic precipitates a predictable response. A test for non-normality selects a correction for non-normality; a test for serial correlation can lead to a correction for serial correlation; a test for heteroscedasticity selects a heteroscedasticity correction. The particular form of the correction may vary with the mood of the observer, but in principle this variability can be determined by an outside observer. In that sense the response is predictable, though possibly random. If so, this is a complex planned response, but the problems are entirely 'above the line' and do not raise the difficult double-counting issues associated with criticism and revision.

Diagnostics that Suggest More Accurate Measurement of the Prior

A Bayesian approach requires the elicitation of a prior distribution which can be done most efficiently after the data are observed, since there are many prior distributions that are practically equivalent to the diffuse prior, and there are many others that are practically equivalent to the dogmatic prior. A companion paper (Leamer, 1989) presents some elicitation diagnostics for the normal linear regression model. These diagnostics indicate when it is a good approximation to use either a diffuse prior distribution or to use the sharp prior that calls for a subset of variables to be altogether omitted. If the sample size is small, one might as well omit the variables; for large samples one might as well include them and estimate with maximum-likelihood. For intermediate sample sizes the prior distribution matters, and needs to be more accurately elicited.

These elicitation diagnostics do depend on the chi-squared statistic that tests the traditional hypothesis that the coefficients of the doubtful variables are collectively zero, and also on the t-statistic that tests if the omission of the doubtful variables causes bias in the estimates of the issue of interest, but other aspects of the data are also relevant.

These elicitation diagnostics do raise a double-counting problem because they reveal features of the data which may affect the prior distribution that is elicited. My guess is that the measured prior distribution would not be greatly affected by knowledge of these diagnostics, but this is an hypothesis that could be experimentally tested.

Diagnostics as Criticisms

Diagnostics may also serve as criticisms of either the model or the prior distribution. The form that criticism should take is not clear-cut whether one takes a Bayesian or a classical perspective. What feature of the data might suggest that the search for a new model would be successful? If one knew the answer to this question in advance, then the response could be planned and criticism would be unnecessary.

I am inclined to think that wrong signs, maybe low R^2 values, and data displays might stimulate me to think of a better model. But I share with Hill (1988) the opinion that 'No theory that I know of attempts to answer [this question], which is a formal way to facilitate scientific creativity.' Thus, when you see claims of automated methods of criticism and hypothesis discovery: *caveat emptor*!

Bayesian Criticisms

There clearly cannot be a fully acceptable formal Bayesian solution to the choice of criticisms because the Bayesian statistical theory is limited to comparing alternative explicit models. A Bayesian, by selecting a prior distribution, say $f(\theta)$, and a sampling distribution, say $f(y \mid \theta)$,

claims to know the distribution from which a statistic $t(y)$ is drawn: $f(t) = \int f(t(y) \mid \theta) f(\theta)\, d\theta$. When the data come from the 'extreme tail' of this distribution, it seems unlikely that the assumptions ($f(y \mid \theta)$ and $f(\theta)$) are correct. But which statistic t should be used and when is $f(t)$ small? I am reminded of the old joke: When asked 'How's your wife?' he replied, 'Compared to what?' The point is that a Bayesian can only say one model is better than another. Formally, the odds in favour of an alternative hypothesis, say H_a, compared with this initial hypothesis, say H_0, are

$$\frac{P(H_a \mid y)}{P(H_0 \mid y)} = \frac{f(y \mid H_a) P(H_a)}{f(y \mid H_0) P(H_0)}$$

This posterior odds ratio depends on both the prior odds ratio $P(H_a)/P(H_0)$ and also the Bayes factor defined as the *ratio* of the density under the alternative to the density under the null. The null density is large or small only in comparison with the density value for alternatives with adequate prior probability. A data set may come from the tail of the null distribution, but it may come even more remotely from the distributions corresponding to sensible alternative hypotheses with reasonably large prior probability!

The problem of the alternative is not satisfactorily resolved by attempting to define the distribution of the data given the vague alternative that 'something else' is happening Barnard (in Savage, 1962, pp. 75–86; quoted in Hill, 1988) argues: "Professor Savage says in effect, 'add at the bottom of the list $H_1, H_2, \ldots$, "something else"'. But what is the probability that a penny comes up heads given the hypothesis 'something else'? We do not know." Furthermore, I ask rhetorically, what is the prior probability of 'something else'? I am inclined to think that a sensitivity analysis could be helpful here. More on this below.

Bayesian Encompassing Diagnostics

One of the popular statistics in the LSE–Hendry tradition tests for 'encompassing' by embedding a pair of non-nested models into a general composite model and testing to see which if either outperforms the composite model. When a model does not perform well compared with the composite model it is said not to 'encompass' the other model. I will make two comments about these encompassing tests from the Bayesian perspective for the special case of the linear model. First, the failure or ability of model 1 to 'encompass' model 2 in the sense of Hendry and Mizon is irrelevant for the choice between models 1 and 2. Second, the encompassing statistics can be used as criticisms of the pair of models, though there is a substantial problem in choosing an appropriate significance level.

Consider the simple setting in which there are three competing regression hypotheses; y depends on X_1, y depends on X_2, and y depends on X_3:

$$H_i\colon\ y \sim N(X_i \beta_i, \sigma_i^2 I) \quad i = 1, 2, 3$$

where y is an $n \times 1$ observable vector, X_i is an $n \times k_i$ observable matrix, β_i is a $k_i \times 1$ unobservable vector, and σ_i is an unobservable scalar.

The Bayesian problem of discriminating among these three hypotheses rests on a straightforward application of Bayes rule, beginning with the prior probabilities of the three hypotheses and a prior distribution over the parameter space. It is then a straightforward application of Bayes rule to compute the posterior probability of each of the hypotheses:

$$P(H_i \mid y) = f_i(y \mid X) P(H_i) / \Sigma_j\, f_j(y \mid X) P(H_j)$$

where f_i is the marginal likelihood of hypothesis i

$$f_i(y \mid X) = \int f_i(y \mid X, \beta_i, \sigma_i) g_i(\beta_i, \sigma_i)\, d\beta_i\, d\sigma_i$$

where $g_i(\beta_i, \sigma_i)$ is the prior distribution for the parameters under hypothesis i.

The posterior odds ratio of hypothesis 1 relative to hypothesis 2 is then the ratio of weighted likelihoods times the prior odds ratio:

$$\frac{P(H_1 \mid y)}{P(H_2 \mid y)} = \frac{f_1(y \mid X)}{f_2(y \mid X)} \frac{P(H_1)}{P(H_2)}.$$

This odds ratio that compares model 1 with model 2 has a very important feature: it does not depend at all on the existence or quality of the third hypothesis. *The performance of the third hypothesis can add or subtract to the total posterior probability of hypotheses 1 and 2, but cannot affect the division of the posterior probability between them.* Thus the ability to 'encompass' is irrelevant to the choice between a pair of models.

Next, suppose that there are only two fully specified competing hypotheses. Though no specific alternatives to these two hypotheses may be identified, it is unlikely that we could have enough confidence in a pair (or finite set) of hypotheses that we would not want to reserve at least a little probability for 'something else'. In order to select between H_1, H_2, and 'something else' we need to specify the distribution of the data if they are generated by 'something else'. One way to think about this, suggested in Leamer (1974), is to suppose that there is a third hypothesis

$$H_3\colon\ y \sim N(X_3\beta_3, \sigma_3^2 I)$$

for which the relevant explanatory variables X_3 are not observed. These unobserved variables must be marginalized from the likelihood function. If, for example, all the explanatory variables come from a multivariate normal distribution, then this marginalization produces the alternative hypothesis[4]

$$H_a\colon\ y \sim N(X_1\theta_1 + X_2\theta_2, \sigma_a^2 I)$$

This composite model that includes both X_1 and X_2 may be theoretically meaningless. This model is formed only as a surrogate for the unspecified alternative that y depends on X_3.

We now have two well-defined hypotheses and a vague alternative. The performance of the vague alternative can cast doubt on the pair of well-defined hypotheses in the sense of lowering the posterior probability assigned to them. A Bayesian diagnostic is therefore a measure of the performance of the composite hypothesis with all the explanatory variables compared with the maintained hypotheses. But, of course, in order to form this measure, one requires a prior distribution for θ_1 and θ_2. How one might do this is something of a mystery, which is Barnard's point in the quotation above. When I feel fatherly, I am inclined to insist that one make a commitment to the choice of prior for these parameters, even though one cannot know what they represent. This commitment allows one later to correct for successful criticisms. See Section 6.

One feature of the distribution for θ_1 and θ_2 that might be selected by convention is the mean. The implicit prior for the unstated model with unobserved variables X_3 is that it does not explain the data. This must mean that β_3 is implicitly revealed to be small, and consequently

[4] By the way, this approach leaves a lot of freedom in forming the alternative hypothesis, even if we commit to the normal regression model if X_3 is observed:

$$f_a(y \mid X_1, X_2) = \int \sigma^2 \exp[-(y - X_3\beta_3)'(Y - X_3\beta_3)/2\sigma^2] f(X_3 \mid X_1, X_2)\, dX_3$$

so are θ_1 and θ_2. Measuring the performance of the alternative model requires not just a prior mean but also a prior variance matrix. For reasons discussed below, this prior variance matrix is required to adjust the inferences for successful and unsuccessful criticisms. In practice, however, it is awfully difficult to submit to this kind of discipline and to commit to a particular choice of this prior covariance matrix. A possible compromise is a sensitivity analysis which allows the prior variance V to be free. A Bayesian diagnostic with known V is the Bayes factor in favour of the alternative hypothesis relative to hypothesis i:

$$B(H_a\colon H_i \mid V) = \frac{\int f_a(Y \mid X, \theta) f(\theta \mid V)\, \mathrm{d}\theta}{f_i(y \mid X)} \quad i = 1, 2$$

A Bayesian diagnostic when V is difficult to select is the maximum Bayes factor in favour of the vague alternative:

$$\max_V B(H_a\colon H_i \mid V)$$

I do not pretend to be able to tell you what should be the critical value of these statistics, since the posterior odds ratio depends both on the Bayes factor and also the prior odds ratio. The question that must be answered is: When is this Bayes factor so high that the search for a new model is likely to be successful? Is it 10 : 1 in favour of the alternative. Or 100 : 1? I don't know. For that matter, I don't even know if this Bayes factor is useful information. It would take a lot of experience before that could be established. This is clearly not a matter of theory, since to get to this point we have made a number of assumptions that are questionable at best. Furthermore, this form of criticism applies only if there are competing non-nested hypotheses with non-zero prior probability, a setting which in my opinion is rare in economics.

Classical Criticisms

Classical criticisms are usually 'goodness-of-fit' tests which also indicate whether the data come from the tail of the assumed distribution. These 'goodness-of-fit' tests have a shaky logical foundation. One problem, pointed out by Berkson (1938), is that unless the model is perfectly correct, the model will surely be rejected as sample size grows. Diagnostics that are tests of the model against unspecified alternatives thus amount only to elaborate schemes for measuring sample size.

It thus seems unlikely that the finding that the data come from the tail of the distribution is properly regarded to be a criticism of the assumed model. But here is a counter-example: suppose that you point out to your class of ten students that two of them have the same birthday. Many of these students would be surprised and might start wondering if there was some nonrandom sorting that has occurred. Then you point to Feller (1957, p. 32) that the probability of no matches is only 0·883, so that an event with rather high probability has occurred. This will probably dissuade the students from looking for another explanation. Note that this sequence of events refers repeatedly to the probability of the data under the assumed model of randomness and never to any alternative. First the probability of a match was thought to be very small, and the data seemed sufficiently anomalous to justify the search for an alternative. Then the miscalculation was pointed out and the higher probability did not seem to justify any further search. Perhaps in this setting one has an intuitive sense of the probability of this kind of data under the alternative that might be constructed, and also the prior probability of this alternative. It is possible, but it does seem doubtful.

7. CORRECTIONS FOR SUCCESSFUL AND UNSUCCESSFUL CRITICISM

Both successful and unsuccessful criticisms have implications for the inferences that are properly drawn from a data set. The attempt to criticize, even when it does not lead to a revision, reveals that there is a lack of complete commitment to the assumptions that underlay the original plans. This lack of complete commitment requires some alteration of the plans—for example, enlargement of the standard errors of the coefficients to reflect the fact that there surely are omitted variables that cause bias in the estimates. When the criticism is successful, there is a double-counting problem because the data are used once to alter the assumptions, and then again to make inferences as if these were the assumptions that were used from the beginning. Something needs to be done to limit the double-counting and to minimize the chance of overfitting.

The corrections for both successful and unsuccessful criticism that I proposed in Leamer (1974) treat the phenomenon of hypothesis discovery as if it were a traditional problem of sequential observation with an initial decision not to observe some of the variables. Suppose that the full model has two explanatory variables:

$$y_i = \alpha + \beta x_i + \gamma z_i + u_i$$

where y, x and z are observables and u_i is a normally distributed serially uncorrelated error term with mean zero and variance σ_u^2. Suppose further that z given x is generated also by a regression:

$$z_i = s + rx_i + e_i$$

where e_i is a normally distributed serially uncorrelated error term with mean zero and variance σ_e^2. Then, if interest focuses on β, it is possible to make the decision to observe only y and x and to estimate the function:

$$\begin{aligned} y_i &= \alpha + \beta x_i + \gamma(s + rx_i + e_i) + u_i \\ &= (\alpha + \gamma s) + (\beta + r\gamma)x_i + (u_i + \gamma e_i) \\ &= \alpha + \alpha^* + (\beta + \beta^*)x_i + (u_i + u_i^*) \end{aligned}$$

where: $\alpha^* = \gamma s$
$\beta^* = \gamma\sigma$
$u_i^* = \gamma e_i$

I assume that the prior distribution for β^* is located at the origin, meaning that the expected bias of the least-squares estimate of β is zero. The presence of the 'experimental bias' β^* reduces the effective sample information about β from $x'x/\sigma^2(x'x/\sigma^2)(vx'x/\sigma^2 + 1)$ where v is the prior variance of β^*. Thus the possibility of misspecification requires a discount of the data evidence that depends on the quality of the experiment measured by v.

When a criticism is successful, and it is decided to observe z, the sample information is properly summarized by the regression of y on x and z; no adjustment is necessary for the fact that the data were observed in stages. However, there is a restriction that must be made on the processing of the data. The prior distribution that is used for γ must be consistent with the prior that was used for β^* since $\beta^* = r\gamma$. With r and γ independent this implies the moments:

$$\begin{aligned} E(r)E(\gamma) &= 0 \\ E(r^2)E(\gamma^2) &= v. \end{aligned}$$

These equations place restrictions on the prior distributions for r and γ. Usually they will be

interpreted to mean that the original decision to omit z reveals a prior for γ that is located at zero with a variance that is limited depending on the size of the prior variance for β^*. The effect of this prior is to shrink the estimate of γ to zero and thus to discount the inferences implied by the regression of y on x and z. The amount of discounting that is required is a decreasing function of the prior variance v.

To summarize, there are a sequence of discount rates that apply at different stages of a data analysis if there is criticism and potential revision. An initial discount applies if the criticism is unsuccessful. If this discount is great, meaning that there is a substantial chance of successful criticism because the model is probably poorly specified, then the discount applying to subsequent models will be less. If, on the other hand, the initial discount is small because the initial model is thought to be pretty good, then the results from data-instigated models are more heavily discounted.[5]

ACKNOWLEDGEMENTS

The author gratefully acknowledges the support of NSF grant SES-8708399, and helpful comments by participants at the 1988 Australian Economics Congress, especially Mike McAleer.

REFERENCES

Baba, Y., D. F. Hendry, and R. M. Starr (1987), 'U.S. money demand, 1960–1984', University of California at San Diego, December.

Berger, J. (1984), 'The robust Bayesian viewpoint', in J. Kadane (ed.), *Robustness of Bayesian Analysis*, North Holland, Amsterdam, pp. 63–124.

Berger, J. O., and A. O'Hagan (1987), 'Ranges of posterior probabilities for unimodal priors with specified quantiles', (Invited paper) Third Valencia International Meeting on Bayesian Statistics, Altea (Spain), 1–5 June.

Berkson, J. (1938), 'Some difficulties of interpretation encountered in the application of the chi-squared test', *Journal of the American Statistical Association*, **33**, 526–542.

Feller, W. (1957), *An Introduction to Probability Theory and its Application*, John Wiley & Sons, New York.

Gilbert, C. L. (1988), 'Alternative approaches to time series methodology in economics', Oxford University, May.

Hendry, D. F. (1987), 'Econometric methodology: a personal perspective', in T. F. Bewley (ed.), *Advances in Econometrics*, vol. II, Cambridge University Press, Cambridge, pp. 29–48.

Hill, B. M. (1980), Review of *Specification Searches*, by Edward E. Leamer, *Journal of the American Statistical Association*, **75**, 252–253.

Hill, B. M. (1985), 'Some subjective Bayesian considerations in the selection of models' (with discussion), *Econometric Reviews*, **4**, 191–288.

Hill, B. M. (1988), 'A theory of Bayesian data analysis', University of Michigan, February.

Keynes, J. M. (1921), *A Treatise on Probability*, Harper & Row, New York.

Leamer, E. E. (1974), 'False models and post-data model construction', *Journal of the American Statistical Association*, **69**, 122–131.

Leamer, E. E. (1975), 'A result on the sign of restricted least squares estimates', *Journal of Econometrics*, **3**, 387–390.

Leamer, E. E. (1978), *Specification Searches*, John Wiley & Sons, New York.

Leamer, E. E. (1982), 'Sets of posterior means with bounded variance priors', *Econometrica*, **50**, 725–736.

[5] Because these formulae restrict jointly the moments of r and γ, they can be satisfied for any choice of the moments of γ. In particular, they are satisfied with $E(r) = 0$ and $\text{var}(r) = v/E(\gamma^2)$. But for some values of $E(\gamma^2)$ this will imply incredible values for $\text{var}(r)$. A certain amount of self restraint will probably be required.

Leamer, E. E. (1983), 'Let's take the con out of econometrics', *American Economic Review*, **73**, 31–43.
Leamer, E. E. (1985), 'Sensitivity analyses would help', *American Economic Review*, **75**, 308–313.
Leamer, E. E. (1987), 'Econometric metaphors', in T. F. Bewley (ed.), *Advances in Econometrics*, Cambridge University Press, Cambridge, pp. 1–29.
Leamer, E. E. (1989), 'Bayesian elicitation diagnostics', University of California at Los Angeles (mimeo).
Leamer, E. E., and G. Chamberlain (1976), 'matrix weighted averages and posterior bounds', *Journal of the Royal Statistical Society, Series B*, **38**, 73–84.
McAleer, M., A. R. Pagan, and P. A. Volker (1985), 'What will take the con out of econometrics?', *American Economic Review*, **75**, 293–307.
Pagan, A. (1987a), 'Three econometric methodologies: a critical appraisal', *Journal of Economic Surveys*, **1**, 3–24.
Pagan, A. (1978b), 'Twenty years after: econometrics, 1966–1986', University of Rochester, Working Paper No. 94.
Smith, A. F. M. (1986), 'Some Bayesian thoughts on modelling and model choice', *Statistician*, **35**, 97–102.
Smith, C. A. B. (1965), 'Personal probability and statistical analysis', *Journal of the Royal Statistical Society, Series B*, **128**, 469–499.