CHAPTER 1

INTRODUCTION

"Data mining," "fishing," "grubbing," "number crunching." These are the value-laden terms we use to disparage each other's empirical work with the linear regression model. A less provocative description would be "specification searching," and a catch-all definition is "the data-dependent process of selecting a statistical model." This definition encompasses both the estimation of different regression equations with different sets of explanatory variables and also the estimation of a single equation using different subsets of the data.

The fact that specification searching invalidates the traditional models of statistical inference is implicit in the pejorative content of the word "fishing," but the industrious implication of the word "mining" suggests that the activity may, in fact, be productive.[1] Although "fishing" too might seem to be a productive activity, the term is usually used in a derogatory way to indicate both the fisherman's great uncertainty over the quantity and quality of fish that might appear in his net and his willingness to accept anything that shows up. Mining, in contrast, is an activity intended to bring to the surface a specific valuable commodity whose existence is likely to be relatively well established before mining commences.[2]

[1]Computer programs for data analysis are given names that reflect this use and abuse of the power of the computer: RAPFE (regression analysis program for economists), ESP (econometric software program), TROLL (time-shared reactive on-line laboratory).

[2]Commercial fishing that involves greatly reduced uncertainty is sometimes called "mining the sea."

This book is about "data mining." It describes how specification searches can be legitimately used to bring to the surface the nuggets of truth that may be buried in a data set. The essential ingredients are judgment and purpose, which jointly determine where in a data set one ought to be digging and also which stones are gems and which are rocks. Without judgment and purpose, a specification search is merely a fishing expedition, and the product of the search will have a value that is difficult or impossible to assess.

The subtitle of this book, "Ad Hoc Inference with Nonexperimental Data," was chosen to suggest that the phenomenon of specification searching is an order of magnitude more common in nonexperimental inference. This can be made definitionally true by asserting that an experiment defines a model. When a specification search occurs, the researcher reveals that he does not think an experiment was conducted. Given this definition, I offer both a descriptive and a prescriptive theory of nonexperimental inference. My observations of economists have led me to the conclusion that there are six logically distinct varieties of specification searches, and each is discussed in this book. The resultant theory is descriptive, in the sense that it springs from observation of nonexperimental scientists at work, but it is also prescriptive, in that it offers alternatives to what seems to be going on now.

A Bayesian approach is used almost exclusively. Anyone who is familiar with the extent to which judgment is used in the analysis of nonexperimental data should have no difficulty in accepting the Bayesian, personal view of inference that is espoused here. Arguments concerning Bayesian versus classical inference are implicit in much of this book, but the battle over the proper philosophical foundations for inference is largely ignored. That battle is intellectually stimulating, and, as far as I am concerned, decidedly one-sided. But it is a battle evidently of little interest to analyzers of real data, perhaps because the practical consequences of accepting the Bayesian view are either ambiguous or minor.

I offer here a different argument in favor of the Bayesian position. The phenomenon of specification searches completely invalidates the traditional models of inference, both Bayesian and classical. But the Bayesian approach is sufficiently flexible that, with suitable alterations, specification searches can be made legitimate, or at least understandable. This does not seem to be the case with the classical model of inference. I am definitely not arguing that one must be a formal Bayesian. I am only claiming that the Bayesian view yields insights. A formal Bayesian encounters insurmountable difficulties in constructing meaningful prior distributions. Thus, most uncertain judgments elude precise quantification. But a way to deal with the fuzziness of quantified probability judgments is to explore

the implications of many different, precisely described judgments, a procedure which seems to me to be better than the other approaches that compound the judgment fuzziness with methodological fuzziness. The myth that inference with nonexperimental data (or any data) could be judgment-free creates an insidious and a counter productive goal.

To the extent that I have been successful in identifying all the reasons for specification searches, this book offers a nearly complete normative theory of personal learning with the linear regression model. It parallels to a great degree the commonsense "ad hoceries" that are characteristic of nonexperimental inference. In this book there are, however, several aspects of learning that are either not mentioned or incompletely discussed. First, no mention is made of the simultaneous-equations problem that plagues nonexperimental inference. The simultaneous equations model does bring up the interesting problem of inferring causality, but from the standpoint of specification searches, it is a formal variant of the simple linear model and therefore implies no interesting methodological issues that are not discussed herein. A second shortcoming is the brief treatment of memory failures. The shortcomings of memory seem quite important for any positive theory of personal inference, although a normative theory may proceed usefully with a perfect memory assumption. The usual Bayesian model implicitly does make this assumption, and memory failures may cast doubt on any of its implications. A third shortcoming of this book is its neglect of social learning. It is obvious that the accumulation of opinion is partly, if not largely, a social phenomenon. Unfortunately, the currently available mathematical models of social learning are primitive and are hardly worth discussing, except that they, rightly, remind us of the social-learning phenomenon. It is useful to observe that the social-learning problem is a special memory problem. Social memory is simply the accumulated set of experiences of *all* individuals, and your access to the totality of experiences depends on your contact and communication with the people who had or who heard of the particular experiences. There are, of course, various distortions for various reasons in the communication of these experiences, just as there are features of personal memory that make some events more memorable than others. Thus the significant shortcoming of this book is its inadequate treatment of memory problems, personal and social.

1.1 The Axiom of Specification

In searching for a model of nonexperimental inference, we may easily discard the textbook version of classical inference. It makes implicit use of the following unacceptable specification axiom.

The Axiom of Correct Specification

(a) The set of explanatory variables that are thought to determine (linearly) the dependent variable must be
  (1) unique,
  (2) complete,
  (3) small in number, and
  (4) observable.
(b) Other determinants of the dependent variable must have a probability distribution with at most a few unknown parameters.
(c) All unknown parameters must be constant.

If this axiom were, in fact, accepted, we would find one equation estimated for every phenomenon, and we would have books that compiled these estimates published with the same scientific fanfare that accompanies estimates of the speed of light or the gravitational constant. Quite the contrary, we are literally deluged with regression equations, all offering to "explain" the same event, and instead of a book of findings we have volumes of competing estimates.

The phenomenon of specification searches thus represents an unambiguous rejection of the axiom of specification and literally pulls the foundation from under classical inference. This book presents an alternative theory of inference that either formally allows specification searches or suggests alternatives. The theory rests on the firm (but fuzzy) foundation of probabilistic judgments. It makes use of formal decision theory in those cases in which a specification search seems to be solving a decision problem.

I am certainly not the first to notice the discrepancy between inference as it is described in the textbooks and inference as it is practiced at the computer center. There is a wide spectrum of opinions concerning the effect of specification searches on inference. "Believers" use ad hoc techniques to search for specification, throwing out insignificant variables here and there, for example, but they continue to regard the end result of such a methodology to be identical to the end result obtained in the experimental sciences (or at least cynically to act that way). Believers report the summary statistics from the nth equation as if the other $n-1$ were not tried, as if the nth equation defined a controlled experiment.

At the other extreme are the agnostics, who gladly admit the irrelevance of classical inference. They argue that a nonexperimental scientist is merely identifying relationships that exist in the historical data. He is describing the salient features of the data accurately but economically. Ideally, the data analysis generates hypotheses that need new data to be tested. Agnostics may thus discount any statistical result until it has been employed in a prediction outside the data period. We might interpret such

statements in a statistical context as the absence of information concerning the standard errors. A point estimate without an associated standard error does not imply an hypothesis test, nor can it determine unambiguous inferences.

Somewhere between these two extremes is a group of pragmatists. They feel that the believers' contentment stems only from ignorance but that the agnostics have gone too far. This group argues that estimated standard errors are properly enlarged by a specification search but not to the extent that they become infinite. Theil (1961), for example, writes:

> The obvious result is that, if a 'maintained' hypothesis [a specification, in our terms] gives unsatisfactory results, it is not maintained but rejected, and replaced by another 'maintained' hypothesis; etc. It is hardly reasonable to say that this kind of experimentation is incorrect, even if it affects the superstructure built on such 'maintained' hypotheses. [In a footnote, he explains that he is referring specifically to the standard errors calculated by classical formulae.] It is especially unreasonable to reject such an experimental approach, because... the statistical theory which forbids the rejection of a 'maintained' hypothesis is not fully satisfactory either in view of the difficulty of its application.
>
> What *is* incorrect, however, is to act as if the final hypothesis presented is the first one, whereas in fact it is the result of much experimentation.

Although Theil is rejecting classical inference as unworkable and berating the naïveté of the believers, he does not offer a procedure that would allow valid inferences in the context of a specification search. By how much are the standard errors to be enlarged? And which of the many estimates are we to choose? A theory of specification searches is needed to answer these important questions.

1.2 The Six Varieties of Specification Searches

A theory of specification searches can be constructed first by identifying the reason a researcher engages in a search, and second by building formal inferential models that properly carry out his legitimate intentions. By observation of economists analyzing data, I have come to the conclusion that there are six different reasons for specification searches. Each is discussed in a separate chapter of this book. The six searches are listed with chapter references in Table 1.1.

For illustrative purposes, imagine a researcher interested in exploring empirically the theory of demand. In its simplest form the theory may be stated as follows: "*Ceteris paribus*, an individual's purchases of some commodity depends on his income and on the price of the commodity." The problem of the empirical worker is to translate this theoretical assertion into a statement about observable phenomena. He must identify the

Table 1.1

Specification Searches

Name of Search	Designed to	Chapter
Hypothesis-testing search	choose a "true model"	4
Interpretative search	interpret multidimensional evidence	5
Simplification search	construct a "fruitful model"	6
Proxy search	find a quantitative facsimile	7
Data-selection search	select a data set	8
Postdata model construction	improve an existing model	9

observable counterparts of the theoretical variables, he must select other variables that may significantly affect purchases, he must choose a particular functional relationship between the variables, and he must decide which individuals are actually to be observed. Because he cannot make these decisions with complete confidence, the researcher is willing to change his mind if his original choices seem not to work out as well as he might have liked. He does so by changing the specification of his statistical model. He may include more explanatory variables; he may omit certain variables; he may substitute one variable for another; he may discard observations, or he may include new observations.

Suppose the initial model is $\log D_i = \alpha + \beta \log Y_i + \gamma \log P_i + u_i$, where D_i is the purchases of oranges by household i, Y_i is monetery income, P_i is the the price of the commodity, and u_i is a "random disturbance" assumed to be normally distributed, independent of u_j, for $i \neq j$. The variables are observed by asking a random selection of heads of households, "How much did you earn last month, how many oranges did you purchase, and how much did they cost?" Using the replies of 150 households, the following regression equation is estimated:

$$\log D_i = \underset{(1.1)}{6.2} + \underset{(.21)}{.85} \log Y_i - \underset{(.13)}{.67} \log P_i, \qquad R^2 = .15,$$

with standard errors in parentheses. For a variety of reasons, it is likely that other equations would be estimated with the same data set. Without endorsing the procedures, I now describe a typical search program.

Of special interest is the hypothesis that the fraction of income spent on oranges is not a function of price, $\gamma = -1$. To test this hypothesis, the equation is reestimated with the constraint applied:

$$\log D_i + \log P_i = \underset{(1.0)}{7.2} + \underset{(.20)}{.96} \log Y_i, \qquad R^2 = .14.$$

Using a standard F test, this hypothesis is rejected at the .05 level, and it is inferred that the data cast doubt on the hypothesis $\gamma = -1$. This is an example of an hypothesis testing search in which different specifications describe different hypotheses about the phenomenon.

The theory of demand describes the behavior of a single individual, but this sample varies across individuals. The nutritional importance of oranges is greatest in areas with the least sunlight, and it may be inappropriate to treat southerners as if they were identical to northerners in their taste for oranges. Separate regressions are therefore computed for southerners and northerners:

$$\log D_i^N = \underset{(1.9)}{7.3} + \underset{(.41)}{.89} \log Y_i^N - \underset{(.25)}{.60} \log P_i^N \qquad R^2 = .18,$$

$$\log D_i^S = \underset{(2.2)}{7.0} + \underset{(.31)}{.82} \log Y_i^S - \underset{(.26)}{1.10} \log P_i^S \qquad R^2 = .19.$$

These regressions suggest that in the North, income is the relatively more important variable and price the relatively less important variable, but the hypothesis that the coefficients are different is not rejected at the .05 level. This is an example of a *data-selection search*. The same theoretical hypothesis underlies all three specifications: the one estimated with all the data and the pair estimated with subsets. The specifications differ in their choice of data sets.

Next it must be observed that the answer to the income question may be a very poor measurement of the household's true income. As it turns out, households were asked to report their expenditures on a fairly inclusive list of other commodities, and it may be that their total expenditures E_i is a better measurement of income than Y_i. The variable E_i is substituted for Y_i, and the estimated equation becomes

$$\log D_i = \underset{(1.0)}{5.2} + \underset{(.18)}{1.1} \log E_i - \underset{(.16)}{.45} \log P_i, \qquad R^2 = .18.$$

The R^2 has increased, and the coefficient on the income variable has become more significant, which suggests that E_i is the better measurement of income. This is a *proxy variable search*. Competing specifications in a proxy variable search all derive from the same underlying hypothesis. Different estimated regressions reflect different ways of measuring a common set of hypothetical variables.

The R^2s in all these equations are unhappily low. Perhaps there are other variables that might be added to the specification to improve the fit. After all, the theory makes use of the Latin phrase, *ceteris paribus*, other things constant, yet it is the nature of nonexperimental research that other

things are not held constant. Although I prefer oranges, if grapefruit are on sale, I will sometimes buy them instead. Adding the price of grapefruit π_i to the equations yields the result

$$\log D_i = \underset{(1.0)}{3.1} + \underset{(.20)}{.83}\log E_i + \underset{(.15)}{.01}\log P_i - \underset{(.60)}{.56}\log \pi_i \qquad R^2 = .20.$$

This specification represents the broader theory: "*Ceteris paribus*, an individual's purchases of some commodity depends on his income, on the price of the commodity, and on the price of 'similar' commodities." The process of revising the underlying theory in response to the data evidence is called *post data model construction*, and the resulting hypothesis is called a *data-instigated* hypothesis. Whereas all other specifications are implicit in the original theoretical statement, a data-instigated hypothesis is not.

In the regression last reported, the coefficients on the price variables are insignificant and of the "wrong" sign. Furthermore, the sum of the coefficients ($.83 + .01 - .56 = .28$) is rather far from zero. The presumption that these coefficients sum to zero derives from the homogeneity postulate that asserts the following. "There is no money illusion: if money income and all prices are multiplied by the same constant, purchases will not change." Applying this homogeneity constraint yields the regression

$$\log D_i = \underset{(.9)}{4.2} + \underset{(.19)}{.52}\log E_i - \underset{(.14)}{.61}\log P_i + \underset{(.31)}{.09}\log \pi_i, \qquad R^2 = .19.$$

The R^2 has fallen only slightly, and the coefficients all have the right sign, two of them significantly so. Thus the constraint seems to improve the specification. This is an example of an *interpretive search*. The underlying hypothesis is taken as given. Restrictions are imposed in the hopes that the estimates may be "improved."

The regression equation now includes three variables, one with a very small coefficient and the other two with coefficients approximately the same size in absolute value. A simple equation would result if π were omitted and the other two coefficients set equal to each other (but opposite in sign):

$$\log D_i = \underset{(.8)}{3.7} + \underset{(.18).}{.58}\log(E_i/P_i) \qquad R^2 = .18.$$

The R^2 is only slightly smaller, and this simple equation is selected. This sixth and final search is a *simplification search*, the function of which is to find a simple but useful model.

The six kinds of specification searches may not yet be clearly different in your mind. In practice, there is little effort made to distinguish one from the other, and it is unsurprising that at first consideration it is difficult to discern the real differences. Moreover, since the searches differ sometimes only in the intent of the researcher and not in his actions, it may be difficult to infer which kind of search actually occurred. By this I do not mean to imply that it is *unimportant* to identify the type of search. Quite the contrary, the effectiveness of a search must be evaluated in terms of its intentions. An apparently successful simplification search may be judged completely unsuccessful as an interpretive search, and so forth.

It is always possible for a researcher to know what kind of search he is employing, and it is absolutely essential for him to communicate that information to the readers of his report. The differences in the searches will perhaps be most clear after this book is read in its entirety, but more may be said in this introductory chapter. Hypothesis-testing searches involve alternative models that have "truth value." It is difficult to find non-Bayesian language that can make such a statement less ambiguous, but in the Bayesian language, hypothesis-testing searches make use of alternative specifications that are assigned positive subjective prior probability. This can be contrasted with an interpretive search, in which only the most general specification is assigned positive probability. In an interpretive search an hypothesis, say, "$\gamma = 0$," is thought to be false with probability one, but the hypothesis, "γ is close to zero," is thought to be quite likely. That is to say, the prior distribution for γ concentrates the probability mass in the neighborhood of $\gamma = 0$ but assigns zero probability to zero. Given any such probability distribution, it is always possible to find a good approximation to it that does allocate positive probability to $\gamma = 0$, and the distinction between hypothesis-testing searches and interpretive searches is thereby blurred. But in a large sample the value $\gamma = 0$ almost certainly becomes uninteresting unless it is allocated a positive prior probability. Thus one practical difference between interpretive searches and hypothesis testing searches is that the former are strictly small-sample phenomena. To put this in the language of classical hypothesis testing, the significance level of a test should be a decreasing function of sample size in an hypothesis-testing search but should be relatively constant in an interpretive search. Incidentally, real examples of hypothesis-testing searches are extremely rare. The most general models used in nonexperimental inference are themselves not regarded to be complete descriptions of the phenomena under study. Restrictions on these "false" models could hardly lead to potentially true models. Hypothesis-testing searches are discussed first in this book only because the formal theory of hypothesis testing is most familiar, not because it solves an important problem.

A simple formal example contrasts three of the searches. A researcher may estimate the pair of equations $Y = x\beta + z\gamma + u$ and $Y = x\beta + u$ where β and γ are uncertain parameters, Y and x are observable variables, and u is the "residual error." From this fact alone it is impossible to determine

whether he has engaged in an hypothesis-testing search, an interpretive search, or a simplification search. In an hypothesis-testing search, the hypothesis $\gamma=0$ means "the model $Y=\beta x+u$ is true." In an interpretive search the same hypothesis implies only that "it is better to estimate β acting as if $Y=\beta x+u$ were the true model than to estimate β using the more complete model." An analogous simplification hypothesis might be "if prediction of Y is the goal, the value of γ is usefully set to zero."

The motivation for hypothesis-testing searches and interpretive searches is prior information. Hypotheses are conjectured to be true or to be approximately true. The motivation for simplification searches is a loss function that penalizes complexity. Hypotheses are not conjectured to be true, or even approximately true. It is only hoped that a simple model would turn out to be adequate.

The data-selection search described above apparently is also an interpretive search, or possibly postdata model construction. In that example, the data set was split into two subsets, and separate regressions were estimated for each. An interpretive search might make use of a general model with two sets of parameters and might test the hypothesis that better estimates would result if the parameters in the two regimes were treated as if they were identical. But since the more general model was not explicitly stated in the beginning, it might be better to think of the search as post data model construction. Data-selection searches could thus be treated as special cases of these and possibly other searches, but the category is nonetheless useful. A theory rarely indicates an experiment that could be used to test it or to estimate its uncertain parameters. A researcher must construct his own experiment, or in the nonexperimental sciences, he must select observations from the set of recorded nonexperiments. The problems he confronts in doing so lead to a data-selection search, even though these problems may be formally similar to the problems associated with other searches.

To understand this more clearly, consider again the theoretical statement, "Y depends linearly on x and z: $Y=\alpha+x\beta+z\gamma$." To estimate this model, a researcher must select a data set over which the parameters α, β, and γ can be thought to be constant, or he must append to this model some description of how the parameters change from observation to observation. In practice, he will often treat the slope parameters β and γ as constants and try several different probabilistic descriptions of the variability of the level α from observation to observation. By definition, a data-selection search deals with the variability of unobservables (parameters and "error" term). An interpretive search introduces prior information about the means of the unobservables. Postdata model construction adds new unobservables to the model.

There is little difficulty in identifying a proxy variable search, but there is great difficulty in determining the inferences that may be legitimately made in the context of a proxy search. At one extreme, the theory is taken as given, and the data are used to construct a quantitative facsimile of the unquestioned theory. The evidence is completely spent to select a proxy, and no evidence is left over for inference about the theoretical parameters. At the other extreme, perfect measurement is assumed, and none of the evidence is spent to select a proxy. Real proxy searches lie somewhere between, with the evidence partly spent to estimate the theoretical parameters. It is difficult to position a search very precisely between the two extremes.

The last search is what I have called postdata model construction. I also like to call it "Sherlock Holmes inference." Sherlock solves the case by weaving together all the bits of evidence into a plausible story. He would think it indeed preposterous if anyone suggested that he should construct a function indicating the probability of all possible configurations of evidence for all possible hypotheses about the crime. In response to a question from Dr. Watson concerning the likely perpetrators of the crime, Holmes replied, "No data yet.... It is a capital mistake to theorize before you have all the evidence. It biases the judgments."[3]

Sherlock avoids formulating the hypotheses because the set of viable alternatives is immense and any attempt to formulate it completely will involve intolerable costs. If an incomplete set of hypotheses is formulated before the data are observed, there is a great risk of not realizing that the data favor some yet unspecified hypothesis. Instead, evidence is used to direct the construction of a set of "empirically relevant" hypotheses, thereby reducing both the cost of formulating hypotheses and the risk of not identifying the "best" hypothesis. There is, unfortunately, an opportunity cost to this process: the data may not also be used in any obvious way to discriminate among the data-instigated hypotheses. This dilemma is most excruciating when the data set is strictly limited, as in astronomy.

Because statistical inference requires a well-specified theory in advance of the data, Sherlock regards statistical inference to be a "capital mistake." Unlike most of us, however, Sherlock has the luxury of the ultimate extra bit of data—the confession. Even under the greatest coercion, our data sets usually resist our efforts to force a confession from them. Without the confession, it is no longer possible to be confident that any inferences are legitimate.

A solution to this dilemma is to act as if Sherlock Holmes inference solved a certain statistical decision theory problem. Given the model

[3]Doyle (1888), *A Study in Scarlet*.

$Y = x\beta + z\gamma + u$, it is possible to determine before seeing any data whether it is necessary to observe z. The variable z may be thought to be uncorrelated with x, or γ may be thought to be small. Then inferences about β may be made without observing z, in the context of the model $Y = x\beta + u$. If the resulting estimate of β is the wrong sign, or if the pattern of estimated residuals is peculiar, one may legitimately change his mind and observe z.

This formal decision theory problem mimics Sherlock Holmes inference in that the data may induce the use of a more general model, but there is a very important difference. In the decision theory problem, the second model must have been explicitly defined before the data were observed. In sharp contrast, Sherlock Holmes admonishes Dr. Watson against formulating models too completely: "It biases the judgments." Although Sherlock Holmes inference is not, and cannot be, a formal statistical decision theory problem, it is nonetheless desirable to act as if Sherlock were solving the decision theory problem, since legitimate statistical inferences are then implied by a Sherlock Holmes procedure.

Consider again the example of postdata model construction. After getting a low R^2 in a regression of demand for oranges on the price of oranges and monetary income, the price of grapefruit is added to the equation. Any economist will explain that the price of close substitutes surely influences purchases of a commodity, and the use of the price of grapefruit does not reflect a new theory but only a more complete version of the theory that was available all along. Excluding the price of grapefruit cannot be sensible theoretically, although it may be desirable practically. This sounds just like the formal decision theory problem, in which it was first determined that observation of the price of grapefruit was unnecessary. Although the researcher did not explicitly solve this problem, I think he did so implicitly. As a result, it is possible to broaden statistical inference to encompass Sherlock Holmes inference.

In general, the consequence of a specification search is what you might expect. There is greater uncertainty over the parameters than is suggested by the final specification. The data evidence is spent partly to specify the model, and only a part of it is left over to estimate parameters or to discriminate among competing models. The one exception to this rule is a simplification search. Simplification is a decision problem that properly occurs after inferences have already been drawn. It is not necessary to discount the evidence because of the search, but it is quite important to understand that the simplified specification is a tool for some anticipated decision problem and is not a model for inference with the given data set.

With the exception of postdata model construction, the other searches produce an equation that tends to understate the uncertainty, because the equation is estimated as if some parameter were known with certainty, when in fact the parameter remains uncertain. The equation is estimated as if the specification were given, whereas the very fact that a search occurred reveals that there is uncertainty over the specification. Loosely speaking, the apparent statistical evidence implied by the final equation must be discounted; the greater the range of search, the greater must be the discount.

The discount applying to a data-instigated model is somewhat different. A data-instigated model is treated as if it were the model the researcher always believed in. As a result, the final specification is certainly better than the original specification, and there can be no discounting because of uncertainty in the specification. A discount nonetheless applies to the final specification. In estimating the original specification the researcher reveals something about his prior information. He thinks that the variables he has omitted are not important. When he decides to add them to the specification, he is obligated to retain his original prior. This prior tends to adjust his estimates back toward the estimates obtained with the simple model. In that sense, the evidence implied by the final specification is discounted.

The various specification searches can be connected with the axiom of correct specification described in Section 1.1. When the set of explanatory variables is not unique, an hypothesis-testing search occurs. The incompleteness of the list of variables leads to postdata model construction. When the list of variables is excessively long, interpretive and simplification searches may be used. Unobservable variables imply proxy variable searches. Finally, data-selection searches are a response to the researcher's uncertainty over the choice of error distribution or to his concern that parameters may have shifted.

1.3 Data in Economics

There is a growing cynicism among economists toward empirical work. Regression equations are regarded by many to be merely stylistic devices, not unlike footnotes referencing obscure scholarly papers. It is the phenomenon of specification searches that has made the profession uneasy, and a theory of specification searches may help.

Distinguishing the various kinds of specification searches is a step in the right direction. Researchers currently do not distinguish one kind of search from another. Casual examination of papers in economics suggests that interpretive searches are the most prevalent, although these are hardly distinguishable from simplification searches. Hypothesis-testing searches are certainly the least common. Regardless of the type, the rules of search are informal and rarely stated explicitly. As a consequence, there is

considerable doubt whether the average nonexperimental scientist is getting what he wants from his techniques. There is even some doubt that he knows what he wants.

Readers of this book are strongly urged not to conclude either that real learning processes could be fully mathematized and therefore trivialized, or that actual learning should be altered to meet fully the mechanical features of any mathematical model. To paraphrase an analogy of Polanyi's (1964), this is a book about violin playing, and while mastery of the technical/mechanical aspects of violin playing is essential, no one would suggest that studying a book alone would lead to great artistry; nor must a great artist completely conform to mechanical standards, the functions of which are primarily to improve the performances of the great mass of lesser artists.

One mathematical model of learning can be discarded, but it is not clear that the other should be retained. I refer, respectively, to classical and Bayesian inferences. Classical inference apparently allows judgments that are either completely certain or "completely uncertain." We are asked to be certain about the parameter spaces but peculiarly uncertain about the choice of parameters within those spaces. Typically, when selecting a parameter space economists also formulate judgments about the likelihood of various values of the parameters within that space. If they do have these uncertain judgments, they may want to make use of Bayesian tools.

It is perhaps more accurate to describe the classical judgmental inputs relative to the strength of the sample evidence, rather than in an absolute sense. The judgments are not absolutely certain or absolutely uncertain; rather they either overwhelm or are overwhelmed by the sample evidence. The process of learning is a "herky-jerky" reaction to the sample evidence, consisting of phases of complete disregard of sample evidence (failure to reject a null hypothesis?) and phases of complete disregard of nonsample evidence (rejection and discarding of the null hypothesis?). A Bayesian approach can obviously deal with the trivial cases of overwhelming sample evidence or overwhelming nonsample evidence, but it considers also the nontrivial problem of mixing two sources of information.[4]

No one who has worked with economic data or who has watched others work with it could retain the notion that economists have either overwhelming sample or overwhelming nonsample information. If this were so, economic research would often involve fitting a single regression equation. No reference to the process that generated the data would be made, no discussion of peculiar coefficients. There would be no collinearity problem, no proxy variables.

[4]The adjective trivial applies to the data-interpretation problem, obviously not to the mathematics that has been developed to solve these problems.

Necessarily, practicing economists have discarded the formal constraints of classical inference, and they have added the essential bits of subjective uncertain information through ad hoc specification searches. This involves trying not two or three different equations but literally thousands. Curiously, they retain a verbal commitment to classical inference, talking about such irrelevant things as "best linear unbiased estimators," "t-ratios," and the like.

When these specification searches are most effective, the final result may be an appropriate mixture of sample and nonsample information, essentially a posterior distribution. This process may be accurately captured by a Bayesian learning model, according to which we begin with a well-specified set of certain and uncertain judgments and enlist the data to encourage or discourage subsets of those judgments. (Another possibility, discussed below, is that economists are not doing statistical inference at all.)

It is highly unlikely that an ad hoc specification search could be as effective in implementing quantifiable uncertain judgments as the Bayesian approach. Even if the two techniques were to yield identical descriptions of the postdata uncertainty, the specification search approach has the critical defect that it cannot clearly distinguish sample from nonsample information, and the researcher thus has no way of effectively communicating the judgments that were required to analyze the data. It is then impossible for a reader to evaluate the reported results.

The inferential problems of nonexperimental scientists thus seem to be especially well suited to Bayesian inference. It is apparently astounding that Bayesian theory, which has been available in rudimentary forms for two centuries and in highly developed forms for several decades, has had so little impact on real data analysis. Can it be that the Bayesian philosophy attracts poor salesmen? Or is the product better in theory than in practice? I'm afraid it may be the latter.

In practice the Bayesian model has two defects. The first is that it requires the researcher actually to select a prior probability function. There is no doubt in my mind that uncertain prior information is used to analyze nonexperimental data. But there is also no doubt in my mind that uncertain prior information is impossible to quantify precisely. Ad hoc procedures may, in fact, be efficient methods of using imprecisely defined priors. I like to comment on this suggestion with a slogan: "The mapping is the message." The meaning of a data set is that it changes opinions. It takes particular prior opinions into particular posterior opinions. A data set may thus be fully described in terms of the mapping that it implies from prior distributions into posterior distributions. It is not necessary and it is even undesirable for a researcher to select a particular prior distribu-

tion. That task properly belongs to the reader of the report. A researcher should instead describe as completely as possible the mapping from priors into posteriors. He may properly recommend a particular prior, but he has no business forcing it on his reader.

An interesting Bayesian analysis of data, therefore, need not use a single, precisely specified prior distribution, and this hurdle need not be surmounted. But in deflecting our course from this one hurdle, we are forced to surmount another: How can the mapping of priors into posteriors be economically analyzed? This is a question that has not been asked often; it seems to me to be the major issue involved in practical use of the Bayesian tools.

There is, unfortunately, a second, potentially insurmountable difficulty with the Bayesian approach. Whereas I am confident that economists interested in statistical inference should be Bayesians, I question whether they should be interested in statistical inference at all. If, instead, they are doing Sherlock Holmes inference, the choice between a distorted Bayesian and a distorted classical approach is ambiguous. The Bayesian approach encourages more careful formulation of the model space, and to the extent that this is the right direction for the profession to move, the approach seems desirable. But Sherlock warns us against excessive theoretical development before seeing the facts. The process of assigning probabilities to models tends to make a researcher believe and cling to his original set of hypotheses. This straitjackets his Sherlock Holmes instincts, and he may ignore important evidence simply because the relevant hypothesis is outside his immediate field of vision.

I hope that this book will make clear the contribution Bayesian inference can make toward understanding the processes of research with nonexperimental data. In many cases it is enough that we understand what we are doing. In several cases specific alternatives are suggested—alternatives that are unambiguously superior to current procedures. Also, a long chapter is devoted to Sherlock Holmes inference, and it is hoped that the reader will understand the importance of this problem in real research as well as its implications for models of inference.

1.4 A Schematic Model of Inference

The model of inference that is being suggested in this book is indicated schematically in Figure 1.1. Inputs into the inferential process occur at ovals 1, 2, and 3. Major elements of the data analysis are indicated in rectangles 4 to 8. The specification search decisions to redo the analysis with a different set of models or propositions are indicated in diamonds 9 and 10. Solid line linkages may be discussed as problems in statistical

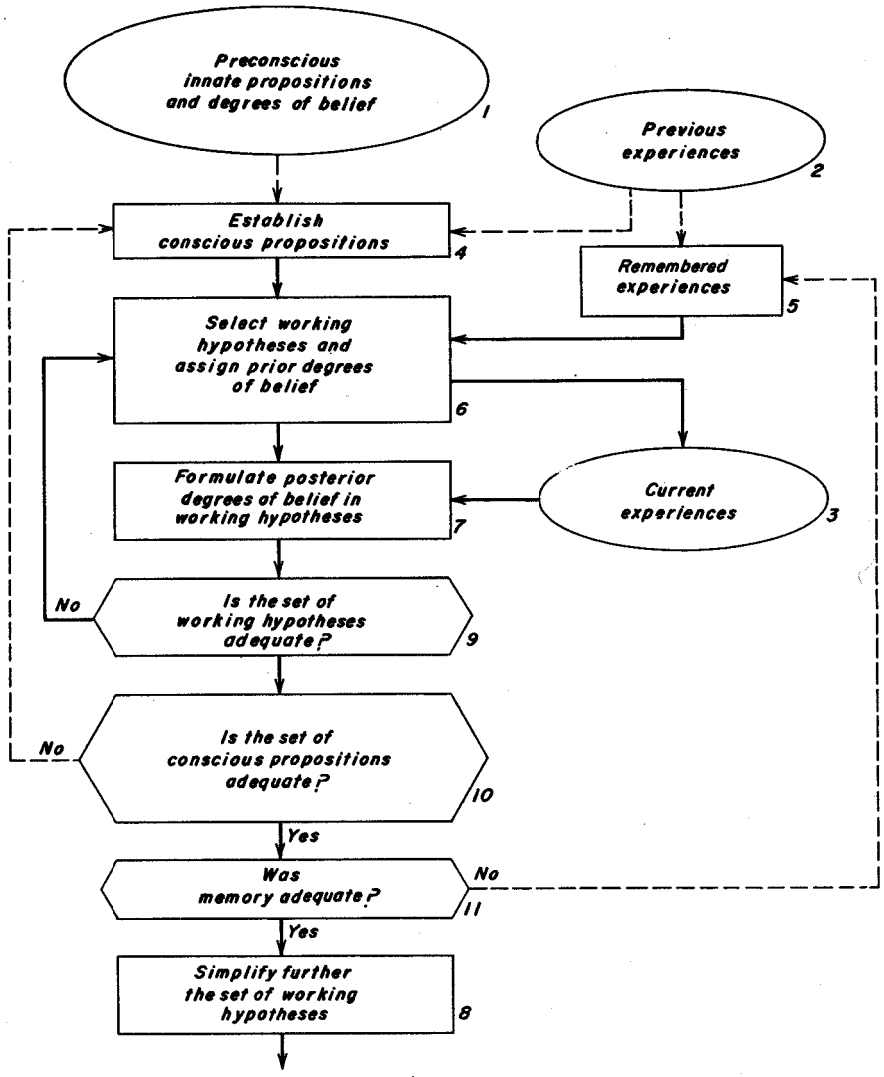

Fig. 1.1 A schematic diagram of inference.

inference as it is currently conceived. Dotted line linkages are philosophically outside the scope of statistical theory.

An individual is thought to have an enormous set of innate but preconscious propositions (oval 1), with innate degrees of belief assigned to them. These are determined, for example, by inherited sensory apparatus. Propositions about heat, hardness, taste are not learned but rather are built

into the nervous system. The experiences indicated in oval 2 are used to select from this set of propositions a relatively tiny set of conscious beliefs (rectangle 4). Remembered experiences (rectangle 5) may then be used to determine which of the still large set of conscious propositions are to be used as a basis for a data analysis. The resulting set of working hypotheses (rectangle 6) is only remotely connected to the set of innate propositions, and the degrees of belief assigned to these propositions must be at best crude approximations to the degrees of belief of a Bayesian with unlimited memory and cognitive skills.

The preobservation "theoretical" work terminates temporarily when the set of working hypotheses is established and degrees of belief are assigned to them. Data are then observed (oval 3). The line linking the working hypothesis rectangle to the current experiences oval allows the choice of data to be a function of the working hypotheses. The data and the prior are mixed in rectangle seven to form tentative posterior degrees of belief.

Peculiarities in the data may then force a reconsideration of the several decisions that were explicitly or implicitly made earlier. The data may suggest that one of the excluded conscious hypotheses should now be included in the set of working hypotheses (diamond 9); or the data may induce the researcher to think up "new" hypotheses (diamond 10); or memory may be searched again (diamond 11). If the researcher changes his mind about one or more of his decisions, he will reanalyze the same data set with different working hypotheses or different priors.

Having satisfied himself with his analysis of the given data set, he must either make use of his newly formed opinions for the imminent decision toward which his efforts had been aimed, or he must make ready for future inference and decision problems as yet ill-defined. In either case he may wish to simplify the set of working hypotheses (rectangle 8).

Most treatments of statistical inference deal with a much more restricted description of learning. The set of hypotheses is ordinarily treated as if it

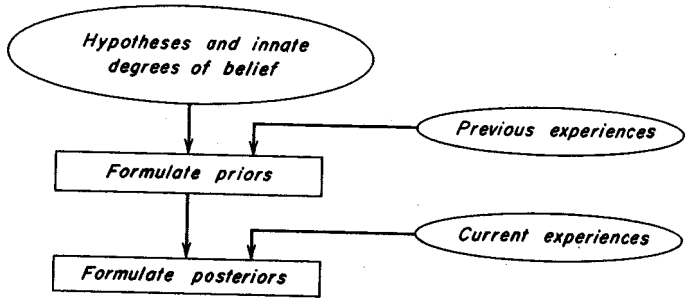

Fig. 1.2 Simple statistical inference.

were complete; that is, the set of working hypotheses, the set of conscious propositions, and the set of preconscious propositions are implicitly treated as if they were identical. Furthermore, the problem of fallible memory is ignored, and if prior information is used, it is assumed to represent accurately all previous experiences. The postdata simplification problem (rectangle 8) is also not discussed. The result is Figure 1.2.

It is possible to extend the logic of statistical decision theory to include both predata and postdata simplification problems. In the case of predata simplification we may formally ask the question: given the costs associated with working with a complete set of hypotheses, is it not better to use a restricted set of hypotheses? Given current experiences, it is logically proper to re-evaluate that decision and therefore to redo the analysis with an enlarged set of hypotheses. The result is Figure 1.3.

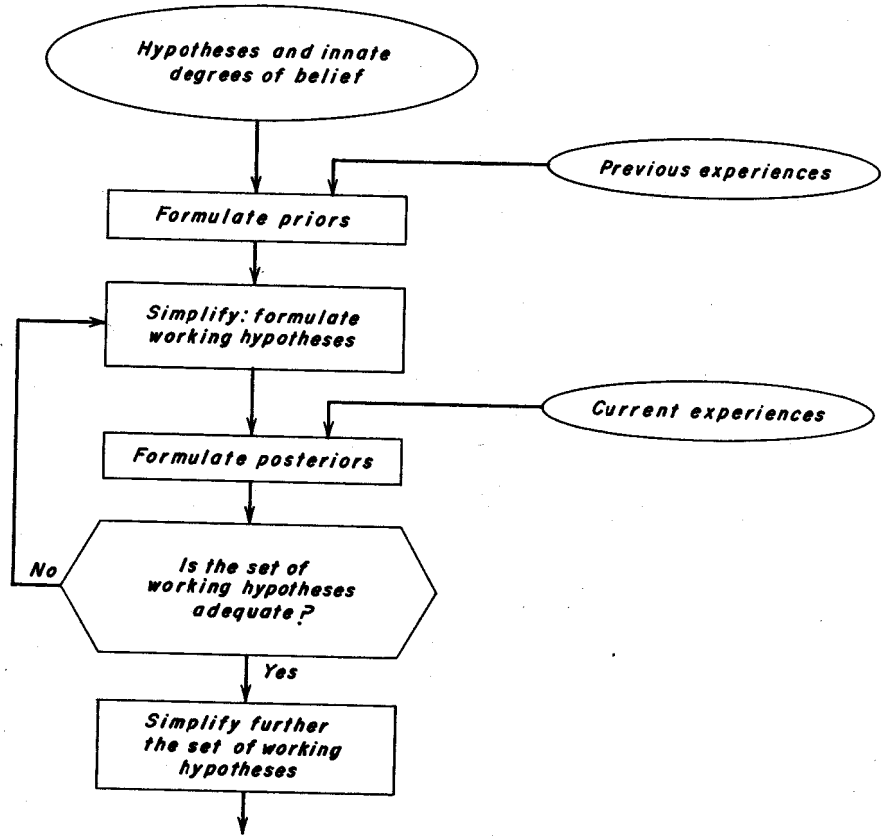

Fig. 1.3 Complete statistical inference.

That part of Figure 1.1 that is not also part of Figure 1.3 is philosophically outside the range of statistical inference. There is, first of all, the problem of memory failure associated with rectangle 5 and diamond 11. Second, there is the problem that the set of conscious hypotheses or propositions is a small subset of the complete set of propositions. This leads to rectangle 4 and diamond 10, which deal with the elicitation of hypotheses from the enormous file of innate propositions.

Four of the six kinds of specification searches lie within the framework of simple statistical inference: interpretive searches, hypothesis-testing searches, proxy searches and data-selection searches. Simplification searches (postdata) also are a straightforward problem in statistical inference, albeit not in the simple versions. The sixth search—postdata model construction—is either within the framework of statistical inference or not, depending on whether the models that are instigated by the data were conscious or preconscious before the analysis began. If the models were preconscious, inference may usefully proceed as if they were, in fact, conscious, and an inference problem that is necessarily outside the framework of statistical inference can be treated as if it were within.