interesting circumstances. Since the massaging concepts and principles seem little affected by the uncertainty in the hyperparameter, the treatment to this point is adequate for the purpose at hand.

Nor do I wish now to deal with the tedious algebra that would be required to treat uncertain hyperparameters. Typically, this involves assigning a hyperparameter some diffuse distribution and either integrating it out of the posterior analytically or writing the equations that would be jointly solved to find the modes of the posterior distribution. In some cases, particularly with the time-varying parameter models, it is still an open question as to which parameters may be assigned diffuse priors and which may not, if a proper posterior is desired.

For treatments of an uncertain autocorrelation coefficient the reader may consult Zellner (1971, Chap. 7), for multivariate regressions with an uncertain covariance matrix, see Zellner (1971, Chap. 8). Lindley and Smith (1972), Geisser (1966), and Box and Tiao (1964) deal with many different multivariate models. Swamy (1971) and Hildreth and Houck (1968) also discuss inference about the parameters of a (prior) distribution. For time-varying parameters there are many papers and references in a special volume of the *Annals of Economic and Social Measurement*, National Bureau of Economic Research (1973).

Another model of time-varying parameters—switching regressions—has been analyzed by Quandt (1958). For a review see Brown et al. (1975).

CHAPTER 9

DATA-INSTIGATED MODELS[1]

The theory of statistical inference takes as given a fixed set of maintained hypotheses. A critical feature of many real learning exercises is, however, the search for *new* hypotheses that explain the given data. An example is a judicial proceeding in which the lawyers for the defense spend their time looking for hypotheses that are plausible given the available facts and that discredit the prosecutor's hypothesis of their client's guilt. Once the proceeding gets to the court, it may concentrate on the statistical inference issue of identifying the data evidence in favor of a set of fairly well-defined hypotheses. But before it gets there, the participants scramble for hypotheses that explain the given evidence. When the search for new hypotheses is successful, the following dilemma must be confronted: how can we say whether the data favor or cast doubt on the new hypothesis, when the new hypothesis was, in fact, constructed to explain the data?

A fictitious example illustrates this dilemma. In a large survey involving many questions it is discovered that coffee drinking and heart disease are correlated, a fact which suggests some control of coffee consumption. The lawyers for the defense, the American Coffee Institute, argue that coffee drinkers tend to fill their tea-

[1]This chapter is taken from Leamer (1974).

pots with the first water out of the tap in the morning, that this water is brackish from sitting in the pipes overnight, and that it is brackish water not coffee that causes heart disease. They recommend plastic pipes. After allowing for the consumption of brackish water by the individuals surveyed, the correlation between coffee consumption and heart disease is greatly reduced, to the point of insignificance. Is coffee guilty or innocent?

A real example of the phenomenon is reported in an interview by Jones (1974), subtitled "Princeton Professor Charles Westoff Finds Twenty Percent Increase in Frequency of Sexual Activity Among Married Americans." (Reprinted from *People Weekly*, © 1974 Time Inc.)

What did you find out?

We started by looking at the relationship between coital frequency and method of contraception in order to determine whether methods with high and low frequency were the same in 1965 and 1970. Then I noticed, almost in passing, that there seemed to be a 20 percent increase in frequency of sexual intercourse between 1965 and 1970.

What was your reaction?

The figures excited my curiosity. I tried to explain the increase at first by looking at obvious reasons, such as the fact that the entire population was younger in 1970 than in 1965, and we know that young people have a high coital frequency and that it declines steadily with age. That explained only a small part of the increase. I then checked our hypothesis that the increased use of the modern birth control methods might explain the increase, but that explained only about a third of it.

Did you then accept the fact of increased sexual activity among married couples?

I was still not sure. It could have been that the apparent increase was not real but rather a reporting phenomenon. That is, because of the more permissive atmosphere surrounding sex, people talk about it much more freely than before and perhaps even feel a pressure to be "with it." The reported increase in sexual activity could be a matter of exaggeration in 1970 and/or under-reporting in 1965.

How did you resolve this "exaggeration factor"?

There was only one test I could think of, and it is hardly definitive. Since the same 20 percent increase in coital frequency showed up more or less among women using different birth control methods, or no method at all, I reasoned that among women who did not practice any contraception, either because they wanted to get pregnant or for other reasons, that a 20 percent increase in sexual intercourse might be reflected in a decrease in the length of time it took such women to become pregnant. So we looked at that and, much to my surprise, saw a substantial change which, quite fortuitously, also showed up as a 20 percent reduction in the time required to conceive in the absence of contraception.

In the end, I was forced to two conclusions. First, that there is a striking relationship between frequency of sexual intercourse and type of contraceptive method, with the greatest frequency associated with the pill, the IUD, and male sterilization. Second, that there has been an increase of about 20 percent in the frequency of sexual intercourse between married couples under the age of 45 between 1965 and 1970.

How do you account for the increase?

I can only speculate about that. I have already mentioned the influence of more effective and more convenient contraceptive methods. The increasing availability of legal abortion has also reduced anxieties among many women about unwanted pregnancies. There has been an increase in openness and permissiveness about sex in our society during this period. Another possible cause results from the fact that divorce rates have been going up, with the consequence that the average duration of marriage was lower in 1970 than in 1965. To exaggerate it, there were more women on their honeymoons in 1970 than in 1965.

This is a delightful example of how research with nonexperimental data frequently proceeds. The observed fact of 20% increase in sexual activity sequentially stimulated the three hypotheses:

1. A younger population.
2. Greater use of birth control devices.
3. Reporting problems.

The first two could not fully account for the increase, and the third was eliminated by the clever use of outside information. This led to two additional hypotheses, not actually examined:

4. Increased permissiveness. (How do we measure it?)
5. Fewer years of marriage.

This is a very clear case of observations in search of hypotheses. If one of the hypotheses turned out to be "successful," can we say that it is favored by the same data? This problem is quite outside the scope of Bayesian statistical theory. The formal Bayesian learning model describes a superbeing who begins his existence with a joint probability function on all uncertain events. Empirical learning amounts to nothing more than the transformation of a marginal to a conditional distribution. In contrast, much of our informal nonnumerical day-to-day learning, and at least some of the more formal statistical-numerical learning, begins without any explicit joint distribution. If the Bayesian learning model is used, it must, therefore, make use of a joint distribution that is constructed given the observed data. This is both philosophically and practically questionable, since it clearly risks double-counting the data evidence.

I like to describe this as 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 the particular evidence at hand for all possible hypotheses and then assign prior probabilities to the hypotheses. He advises instead, "No data yet.... It is a capital mistake to theorize before you have all the evidence. It biases the judgments."[2]

There is, incidentally, a tendency among social scientists, particularly those most trained in statistical inference, to disparage Sherlock Holmes inference. "Boy, he really went on a fishing expedition that time, didn't he?" The fact that Sherlock Holmes procedures invalidate statistical inference is even sometimes taken to mean that Sherlock Holmes inference is "unscientific." Nothing could be further from the truth. In fact, a strong argument can be made that statistical inference, not Sherlock Holmes inference is unscientific. The nineteenth century French physiologist Bernard (1927, pp. 137–138) writes (quoted by Cornfield, 1975)

> A great surgeon performs operations for stones by a single method; later he makes a statistical summary of deaths and recoveries, and he concludes from these statistics that the mortality law for this operation is two out of five. Well, I say that this ratio means literally nothing scientifically and gives no certainty in performing the next operation. What really should be done, instead of gathering facts empirically, is to study them more accurately, each in its special determinism...by statistics, we get a conjecture of greater or less probability about a given case, but never any certainty, never any absolute determinism...only basing itself on experimental determinism can medicine become a true science....

Of course, this overstates the case, but there can be no doubt that an essential part of the scientific method is a careful examination of the anomalies of the data, with the intent of finding plausible explanations if possible. Kuhn (1969, pp. 9–10) makes this point forcefully in explaining why astronomy is a science and why astrology is not:

> Compare the situations of the astronomer and the astrologer. If an astronomer's prediction failed and his calculations checked, he could hope to set the situation right. Perhaps the data were at fault: old observations could be re-examined and new measurements made, tasks which posed a host of calculational and instrumental puzzles. Or perhaps theory needed adjustment, either by the manipulation of epicycles, eccentrics, equants, etc., or by more fundamental reforms of astronomical technique. For more than a millennium these were the theoretical and mathematical puzzles around which, together with their instrumental counterparts, the astronomical research tradition was constituted. The astrologer, by contrast, had no such puzzles. The occurrence of failures could be explained, but particular failures did not give rise to research puzzles, for no man, however skilled, could make use of them in a constructive attempt to revise the astrological tradition. There were too many possible sources of difficulty, most of them beyond the astrologer's knowledge, control, or responsibility. Individual failures were correspondingly uninformative, and they did not reflect on the competence of the prognosticator in the eyes of his professional compeers... In short, though astrologers made testable predictions and recognized that the predictions sometimes failed, they did not and could not engage in the sorts of activities that normally characterize all recognized sciences.

[2]Doyle (1888).

An implication of both of these quotations is that Sherlock Holmes procedures are an essential feature of scientific learning. But when models are instigated by the data, the traditional theories of inference are, regretably, invalidated. It does seem intuitively clear that the data evidence is weaker than it would have been if a complete set of models had been hypothesized before observation commenced. It thus seems desirable to have a method by which evidence can be formally discounted when postdata model construction occurs. This would have the desirable benefit of putting a price on this kind of data mining. Researchers would then be encouraged more carefully to consider the cost of hypothesis specification relative to the costs of data evidence deterioration through Sherlock Holmes procedures.

I propose in this chapter a method of discounting evidence that parallels a formal decision-theoretic analysis of a presimplification problem in which models are simplified before observation to avoid observation or processing costs. During the analysis, relatively inexpensive tests may indicate that the simplification is undesirable, and the full model may be resurrected. Postdata model construction may thus be interpreted as the data-dependent decision that presimplification is undesirable.

For example, given the two-variable linear regression model $\mathbf{Y}=\mathbf{x}\beta+\mathbf{z}\gamma+\mathbf{u}$ and the auxiliary regression $\mathbf{z}=\mathbf{x}r+\boldsymbol{\varepsilon}$, it is not necessary to observe $\mathbf{z}$ in order to make inferences about β, if either γ or r is zero. Even if neither is identically zero, it may be uneconomical to suffer the costs of observing $\mathbf{z}$. However, once $\mathbf{Y}$ and $\mathbf{x}$ are observed, you may change your mind about observing $\mathbf{z}$, possibly because the sample correlation between $\mathbf{Y}$ and $\mathbf{x}$ is too low or the wrong sign.

This formal decision theory proolem requires a supermind, capable of fully specifying an unsimplified model and the relevant prior distributions. But the principal reason most of us use presimplified models is to avoid the (unlimited?) cost of a full probability assessment. Once a full assessment is made, it seems likely that the true ("believed") complete model would be used. Although a simplified model thus cannot usefully result from the

formal decision-theory apparatus, we can think of our models as if they were so derived. In fact, the informal construction of a "working hypothesis" parallels closely the formal decision-theory problem. Models are constructed not as reality but rather as simplifications useful for some implicit or explicit decisions.

The reason for adopting this attitude toward models is that it implies constraints on priors for models constructed after the data analysis commences. The implication of these constraints is that data evidence is discounted in an appealing way when it results from a postdata model search. In the two-variable model mentioned previously, the conditional mean of $\mathbf{Y}$ given $\mathbf{x}$ is $\mathbf{x}(\beta+r\gamma)$. The regression of $\mathbf{Y}$ on $\mathbf{x}$ thus yields an estimate of $\beta+r\gamma$. If you interpret this as an estimate of β, then you have revealed that you think $r\gamma$ is small. If you then decide to observe $\mathbf{z}$, and if you do not improperly alter your prior, you will shrink the estimate of γ toward the revealed prior mean of zero. Thus the data evidence will have to be strong enough to overcome this prejudice, and in that sense the evidence is discounted.

This chapter is divided into six sections. The phenomenon of concept formation is further introduced in section 9.1. The implications (or rather the nonimplications) of stopping rules for inference are discussed in Section 9.2. A surprising conclusion of this chapter is that suspicion of postdata model construction should derive not from the rule used to add variables to an equation, but rather from the improper alteration of one's original implicit priors.

The idea that is being introduced in this chapter is presimplification of models. Concept formation is interpreted as the decision to use a more complex model that was at least implicitly known all the time. Inference with presimplified models is discussed in Section 9.3. A presimplified model necessarily involves an uncertain misspecification error, which causes us to discount any evidence implied by it. Inference with models that are constructed after data are observed—data-instigated models—is discussed in Section 9.4. Quite simply, in using the simple model a researcher reveals certain things about his priors for the more complex models. We are merely suggesting that he stick with those judgments. The fifth section reports an example and the sixth some concluding remarks.

9.1 Concept Formation

The problem of concept formation may be illustrated in a simple example. A mythical kingdom is inhabited only by (green) parakeets, (green) crocodiles and (white) swans. A newly arrived visitor named Richard first meets two swans and two crocodiles. The latter, being in a nasty mood, proceed

to bite Richard on the leg. That evidence suggests to Richard the slogan "green bite, white all right," a theory that seems to work well enough when he meets a third swan and a third crocodile. However, the seventh being he confronts is a friendly parakeet, who forces Richard to alter his slogan to "white all right, green usually bite." Being a good Bayesian with a uniform prior for p, the probability that a green being will bite, Richard assigns degree of belief $4/6$ to the proposition "The next green being I meet will bite me," and he furthermore anticipates that he will accumulate evidence about p, the proportion of green beings that bite. Richard's wife, who is little awed by the mathematical and logical bases of Richard's statement, proclaims "You fool! In truth, 'Four legs bite, two legs all right'."

Well, that is a theory that indeed "predicted" the data with certainty. The probability that three out of four green beings bite given p, the proportion of green beings that bite, is only $\binom{4}{3}p^3(1-p)=4p^3(1-p)$. Richard approximated his prior for p with a uniform distribution and computes the "Bayes factor" in favor of the "legs" hypothesis relative to the "color" hypothesis as

$$\frac{P(\text{data}\,|\,\text{legs theory})}{\int P(\text{data}\,|\,\text{color theory with proportion } p) f(p)\,dp}$$

$$= 1 \Big/ \left(4\int_0^1 p^3(1-p)\,dp\right)$$

$$= 1 \Big/ \left(1-\frac{4}{5}\right) = 5.$$

He concludes that his wife's hypothesis is favored by the ratio five to one relative to his own. In fairness to his wife, Richard supposes that he had equal prior degrees of belief in each hypothesis, from which he calculates the probabilities of each hypothesis as

$$P(\text{legs hypothesis}|\text{data}) = 5/6$$

$$P(\text{color hypothesis}|\text{data}) = 1/6.$$

With these he can calculate his degree of belief $(4/6)(1/6)+1(5/6)=34/36$ in the proposition "The next green being with four legs that I meet will bite me."

At this point Richard, who is used to changing his degrees of belief only in response to data evidence, observes confusedly that his degree of belief in this proposition increased from $4/6$ to $34/36$ in response only to his wife's observation, "Four legs bite, two legs all right." "Can this be data evidence?" he asks himself. In retracing his steps, Richard discovers that

the assumption that the legs hypothesis receives zero prior probability would mean that he would not change his degrees of belief. He tentatively suggests that what he has done is to alter the relative prior probability of the two hypotheses, and that more generally his current degrees of belief depend on that relative probability. Perplexed, he asserts, "If I alter the relative prior probability of the two hypotheses it is because my wife pointed out certain compelling regularities in the data evidence not because of any new experiments. My rules of inference are designed to prevent me from making inferential errors, in particular from double-counting the evidence. It is obvious double-counting to let the data first alter the prior odds ratio from zero to one as I change the prior odds in response to my wife's suggestion and secondly alter the odds ratio from one to five as I would if I applied Bayes' rule to the new prior odds. I shall stick to my original assessment." To this his wife replies, "You fool. Can you not see that your original odds ratio was a mistake? You surely never held degree of belief zero in the legs hypothesis. Better to admit your mistake now than to perpetuate your error." Richard sighs in response, "Yes, I suppose it was a mistake. But I don't see how I can without self-deceit assess any new prior odds ratio which is legitimately unpolluted by the data, and I don't see therefore how I can correct my old mistake without making a new one. Besides, you are so skilled at 'explaining' observations that, regardless of the data you would have come up with some plausible and compelling hypothesis. Why should I believe this one?"

This example aptly illustrates the dilemma of concept formation. In any real learning situation, data evidence strongly compels us to alter our prior, but if we do so, we risk double-counting the data and placing excessive faith in the current evidence. An appropriate model of inference would necessarily allow hypothesis discovery but would also discount the data evidence when it occurs. We propose a method that does just that. It rests on the observation that Richard's p, the probability that a green being will bite, is not conditional on all other features of the being. It is rather a marginal probability such as

$$\begin{aligned} p &= P(\text{bite}|\text{green}) \\ &= P(\text{bite}|\text{green and two legs})P(\text{two legs}|\text{green}) \\ &\quad + P(\text{bite}|\text{green and four legs})P(\text{four legs}|\text{green}). \end{aligned}$$

In particular, let us suppose that conditional on the two hypotheses, the probabilities of biting are as given in Table 9.1. Furthermore, let $1-f$ be the proportion of green beings with two legs, f the proportion with four legs. Assume also that Richard observes randomly selected beings.

Table 9.1
Probability of Bite

Hypothesis	H_0	H_1
Green, two legs	p_0	0
Green, four legs	p_0	1

Letting $\pi = P(H_0)$, $1-\pi = P(H_1)$, we can then calculate p as

$$\begin{aligned} p &= P(\text{bite}|\text{green}) \\ &= \begin{cases} p_0 & \text{with probability } \pi \\ f & \text{with probability } 1-\pi \end{cases} \end{aligned} \tag{9.1}$$

We suppose that Richard has a prior distribution for p_0 and f and also that he assigns a number to π, which, together, imply the mixed distribution (9.1) for p. When he first arrives in our mythical kingdom he makes the judgment that he will observe only color, partly because he is not so good at counting legs and partly because he does not have much faith in H_1 (π is close to one). His prior for p together with the likelihood function implied by three bites in four trials imply a posterior distribution for p. At this point, since his wife implicitly lowers the cost of counting the legs, Richard proceeds to observe the number of legs and therefore to condition on that data as well in applying Bayes' rule. He uses his prior for p_0 and the likelihood evidence of three bites in four trials to compute a posterior for p_0 conditional on H_0. He also uses the Bayes factor as above to compute the posterior probabilities of the two hypotheses.

What, then, is the problem of concept formation? We have just described a perfectly valid application of Bayes' rule with sequential construction of theories. Most of us would be suspicious of the new concept because it was "constructed" only when the first one failed to predict the data perfectly. But the decision to observe the number of legs when the color concept "fails" is what is known as a noninformative stopping rule that to a Bayesian has no implications for inference. (More is said about noninformative stopping rules shortly.) The problem of concept formation lies not in the stopping rule but rather in the failure to observe the constraints implied by the probability function (9.1). In particular, p and p_0 are not the same parameter unless $\pi = 1$, the trivial case in which the observation of the number of legs, is ignored.

Of course, in constructing his prior probability function for p, Richard did not consciously have in mind the alternative hypothesis H_1. It is thus

not possible to treat the problem of concept formation strictly as described above with the distribution for p derived from distributions for f and p_0. Inference in the context of concept formation is, therefore, necessarily not a topic that can be handled by formal methods of statistical inference. Richard can, however, act *as if* he were deriving his distribution for p from the more basic distributions. Several arguments may be made that this is a desirable approach. The fact that Richard wants to change his mind when his wife makes her suggestion is evidence that Richard does not assign to H_0 a degree of belief equal to one. Unless π were one, a prior for p is necessarily a derivative distribution, and Richard could not apply Bayes' rule to make inferences about p unless he could implicitly derive the prior distribution for p from the more basic distributions. And finally, quite pragmatically, if he behaves as if he so derived his distribution for p, we can solve the inferential issues raised by the phenomenon of concept formation; otherwise they remain entirely beyond our reach.

9.2 Stopping Rules and Inference

At first blush the problem of concept formation appears to be associated with the fact that new hypotheses are data instigated. How could we possibly claim that the data favor H_1 relative to H_0 when the only reason H_1 is examined at all is that H_0 did not work? To be more specific, imagine a researcher who adds variables to his regression equation until a favorite coefficient is significantly positive. We would all chastise him for prejudicing his conclusions in such an obvious way, and we would want to discount his results because of his biased rule for observing the data.

For me, the greatest surprise of the Bayesian logic is that these instincts are simply wrong. This rule and, practically speaking, any rules for observing the data are noninformative stopping rules. They have no implications whatsoever for inference. This counterintuitive assertion needs considerable argument before it can be accepted. Thus we consider in this section the inferential problems implied by stopping rules.

As an example of the kind of problem raised by optional stopping, consider binomial sampling with a sample size equal to one if there is a success on the first trial and equal to two otherwise. The sampling distribution and a hypothetical estimator of p are given below.

Sample	Probability	Estimator
S	p	a
FS	$p(1-p)$	b
FF	$(1-p)^2$	c

If we wanted our estimator of p to be unbiased, we would have to choose a, b, and c to satisfy for all p

$$p \equiv ap + bp(1-p) + c(1-p)^2$$

or

$$0 \equiv c + (a+b-2c-1)p + (c-b)p^2.$$

This is satisfied in the interval $0 \leqslant p \leqslant 1$ if and only if all three coefficients in this polynomial are zero. The only solution is $c=b=0$, $a=1$. This is equivalent to throwing out the second observation.

The usual estimator—the number of successes r divided by the number of trials n—has $a=1$, $b=1/2$, and $c=0$. The expected value of this estimator

$$E\left(\frac{r}{n}\right) = p + \frac{p(1-p)}{2}$$

exceeds p. That is, the sampling scheme prejudices the sample in favor of high values of p, since there is a tendency to observe samples with too many successes.

But let us look at this from a Bayesian point of view. The posterior distribution of p is, of course, the product of the likelihood function times the prior. But the likelihood function (the second column in the table) is exactly the same for every sample as the likelihood function derived under a sampling rule with fixed sample size. Thus from a Bayesian point of view the meaning of the sample FS does not depend on the stopping rule, and the fact that you might have stopped on the first trial is quite irrelevant for the interpretation of this particular sample.

A more concrete example illustrates why the stopping rule should not matter. Suppose that boys are born with probability one-half, and that all families stop having children if their first is a boy, otherwise, they have two children. Family composition and probability would then be

Family	Probability
B	$p=1/2$
GB	$p(1-p)=1/4$
GG	$(1-p)(1-p)=1/4$

The average proportion of boys per family would be $1/2+1/4\cdot 1/2=5/8$, more than one-half. Apparently, the stopping rule has biased the population in favor of boys. But the proportion of boys in the whole population is still one-half. Apparently, the stopping rule has failed.

Concern over the stopping rule derives from the following proposition: the mean proportion of boys per family exceeds one-half; if you estimate p

by taking an average over all families of the family proportions, you will necessarily exceed one-half. The error being made, however, is not that families are reporting "biased" numbers; rather, it is that you have not allowed for family size. If you weight families by family size, you will obtain the right number. Exactly the same thing can be said about the estimator (r/n). The fact that (r/n) is a biased estimator is not because r/n is a "biased" summary of the data for any sample; rather, it is because the expected value operator does not weight by sample size.

It is easy to see that stopping rules dependent on the data only are noninformative. Let θ be the parameter of interest, let $X_1, X_2, \ldots$ be a sequence of observations, and let N be the sample size. Given the sample $X_1 = x_1, X_2 = x_2, \ldots X_n = x_n$ and $N = n$, we may write the likelihood function as

$$\begin{aligned} L(\theta; \mathbf{x}, n) &\propto P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n, N = n|\theta) \\ &= P(N = n|X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n, \theta) \\ &\quad P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n|\theta). \end{aligned}$$

But the stopping rule is assumed to terminate sampling with probability one given the sample $x_1, x_2, \ldots, x_n$; thus $P(N = n|\mathbf{X} = \mathbf{x}, \theta) = 1$, independent of θ, and the likelihood function is proportional to $P(\mathbf{X} = \mathbf{x}|\theta)$ regardless of the stopping rule.[3]

[3]The apparent danger of a stopping rule is that a researcher can prejudice the sample in any way he sees fit. He may even be able to sample to a foregone false conclusion. Suppose that we sample from a normal population with mean zero and variance one. If we wish to prove that the mean is, in fact, not zero, we may want to continue sampling until the sample mean is "significantly different from zero." That is, let

$$m_n = \sum_{i=1}^{n} x_i/n,$$

the mean of the sample of size n, and adopt a stopping rule

if $|m_n|n^{1/2} > 1.96$, stop

otherwise, continue sampling.

The statistic $m_n n^{1/2}$ is, of course, the normal statistic typically used to test the hypothesis $\mu = 0$ against the alternative $\mu \neq 0$. A value of $|m_n|n^{1/2} > 1.96$ is taken as evidence against the point-null hypothesis.

Surprisingly enough, this inequality will eventually be satisfied with probability "essentially one." It is apparently possible to sample to a foregone false conclusion. The paradox is completely resolved, however, by noting the discussion in Chapter 4.2, that from a Bayesian point of view a value of $m_n n^{1/2}$ equal to 1.96 may, in fact, be overwhelming evidence in favor of $\mu = 0$, if the sample size is large enough. That is the significance of a "statistically significant" result depends on sample size. If it takes a large sample to get the result, this should be taken as evidence in favor of the null hypothesis. For references and further discussion see Cornfield (1969).

To sum up, we have argued by analogy with the problem of estimating the proportion of boys in a population that the bias of the usual estimator when there is a stopping rule should cause uncomfortableness not with the estimator but with the concept of bias. Constructing an unbiased estimator is roughly equivalent to passing a law that families with one boy may tell the truth, whereas larger families must report to the census taker that they have no boys at all. Bias, being a property of a sampling distribution, is not of direct interest to a Bayesian. Concern over bias from a sampling theory point of view apparently derives from the following proposition. If $\hat{\theta}_i$ is a biased estimator of θ, if n independent values $\hat{\theta}_1, \ldots, \hat{\theta}_n$, are observed, and if a composite estimator $\theta_n^* = \Sigma\hat{\theta}_n/n$ is computed, then as n grows, θ_n^* will converge (in a probability sense) to a value different from θ. Thus accumulation of evidence will not lead to the truth. The counter-argument, as we have seen, is that the appropriate pooling of evidence does not lead to θ_n^*. Instead, we should maximize the composite likelihood function formed by multiplying the individual likelihoods together. Bias should, therefore, concern us only if for some peculiar reason we are compelled to pool information from different samples in this undesirable way.

Most practical stopping rules that lead to biased estimators are from the Bayesian point of view noninformative and therefore irrelevant to the inference problem. In particular, the class of rules dependent on the data alone is noninformative.

9.3 Inference with Presimplified Regression Models

The idea on which our solution to the problem of concept formation rests is that inferential models are highly simplified versions of the learner's inherent set of beliefs. In this section we discuss inference with regression models that are simplified versions of more complete models.

A regression function in a linear nonstochastic world may be written

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} \tag{9.2}$$

where $\mathbf{X}$ and $\mathbf{Z}$ are observable matrices, $\mathbf{Y}$ an observable vector, and $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are unobservable parameter vectors. As in the example in the introduction, inferences about $\boldsymbol{\beta}$ may be made by observing $\mathbf{Y}$ and $\mathbf{X}$ alone,

$$P(\boldsymbol{\beta}|\mathbf{Y}, \mathbf{X}) \propto \int_{\boldsymbol{\gamma}} \int_{\mathbf{Z}} P(\mathbf{Y}, \mathbf{X}, \mathbf{Z}, \boldsymbol{\beta}, \boldsymbol{\gamma})\, d\mathbf{Z}\, d\boldsymbol{\gamma}.$$

The integral in $\mathbf{Z}$ is easily computed by writing the linear regression function (an assumption)

$$\mathbf{Z}|\mathbf{X} = \mathbf{X}\mathbf{R} + \mathbf{U}$$

where $\mathbf{U}$ is a matrix of random variables subjectively independent of $\mathbf{X}$ and

where $\mathbf{R}$ is assumed known. The resulting working hypothesis is then

$$\mathbf{Y}|\mathbf{X}=\mathbf{X}\boldsymbol{\beta}+\mathbf{X}\mathbf{R}\boldsymbol{\gamma}+\mathbf{U}\boldsymbol{\gamma} \tag{9.3}$$

where we have written $\mathbf{Y}|\mathbf{X}$ to emphasize the point that other observables have been marginalized out.

The usual regression model

$$\mathbf{Y}=\mathbf{X}\boldsymbol{\beta}+\mathbf{u} \tag{9.4}$$

with $\mathbf{u}$ independent of $\mathbf{X}$ is derivable from the working hypothesis (9.3) only if $\mathbf{R}\boldsymbol{\gamma}=\mathbf{0}$. Thus the usual analysis involves a (well-known) specification assumption that left-out variables have either zero effect ($\boldsymbol{\gamma}=\mathbf{0}$) or are uncorrelated with included variables ($\mathbf{R}=\mathbf{0}$). Postdata model construction in response to peculiarities in the least-squares estimate of $\boldsymbol{\beta}$ constitutes a de facto rejection of this assumption.

The model

$$\mathbf{Y}=\mathbf{X}\boldsymbol{\beta}+\mathbf{X}\boldsymbol{\beta}^c+\mathbf{u} \tag{9.5}$$

with $\mathbf{u}\sim N(\mathbf{0},\sigma^2\mathbf{I})$ offers a closer approximation to the probability assignments implied by the working hypothesis (9.3). It admits the possibility of left-out variables ($\boldsymbol{\beta}^c$ plays the role of $\mathbf{R}\boldsymbol{\gamma}$) but does not require us actually to identify them. It also fits, with minor modification, into the traditional statistical theory.

The parameter vector $\boldsymbol{\beta}^c$ summarizes the bias in the information about $\boldsymbol{\beta}$ due to excluded variables. It is called either a *contamination vector* or an *experimental bias* vector. The usual regression model (9.4) is called a *false model*, since it unbelievably sets the contamination vector to zero and since it yields reasonable results only if that approximation is adequate. The amended model (9.5) is called the *working hypothesis*, indicating that degrees-of-belief are not allocated directly to it but rather are derived implicitly from a true model or *"world view"* such as (9.2). A working hypothesis includes a statement about the quality of the experiment (a prior on $\boldsymbol{\beta}^c$); a false model does not.

We may "identify" model (9.5) by specifying $\boldsymbol{\beta}^c$ and making inferences about the theoretical coefficient $\boldsymbol{\beta}$ or by specifying $\boldsymbol{\beta}$ and making inferences about the experimental bias $\boldsymbol{\beta}^c$. Informative priors offer a range of intermediate inferences. Analysis of this model from a Bayesian point of view is a straightforward generalization of Pratt, Raiffa, and Schlaifer's (1965) biased sampling. Let the prior be normal with mean and variance

$$E\begin{bmatrix}\boldsymbol{\beta}\\ \boldsymbol{\beta}^c\end{bmatrix}=\begin{bmatrix}\mathbf{b}^*\\ \mathbf{0}\end{bmatrix} \tag{9.6}$$

$$V\begin{bmatrix}\boldsymbol{\beta}\\ \boldsymbol{\beta}^c\end{bmatrix}=\sigma^2\begin{bmatrix}\mathbf{N}^* & \mathbf{0}\\ \mathbf{0} & \mathbf{B}\end{bmatrix}^{-1}, \tag{9.7}$$

with $\sigma^2=\mathrm{Var}\,u_t$. With $\mathbf{N}=\mathbf{X}'\mathbf{X}$, the posterior precision of the coefficient vector is

$$\sigma^{-2}\begin{bmatrix}\mathbf{N}^* & \mathbf{0}\\ \mathbf{0} & \mathbf{B}\end{bmatrix}+\sigma^{-2}\begin{bmatrix}\mathbf{N} & \mathbf{N}\\ \mathbf{N} & \mathbf{N}\end{bmatrix}$$

with posterior variance

$$V\left(\begin{bmatrix}\boldsymbol{\beta}\\ \boldsymbol{\beta}^c\end{bmatrix}|\mathbf{Y}\right)=\sigma^2\begin{bmatrix}\mathbf{N}^*+\mathbf{N} & \mathbf{N}\\ \mathbf{N} & \mathbf{B}+\mathbf{N}\end{bmatrix}^{-1}$$

$$=\sigma^2\begin{bmatrix}\mathbf{D}^{-1} & -\mathbf{D}^{-1}\mathbf{N}(\mathbf{B}+\mathbf{N})^{-1}\\ -\mathbf{E}^{-1}\mathbf{N}(\mathbf{N}^*+\mathbf{N})^{-1} & \mathbf{E}^{-1}\end{bmatrix} \tag{9.8}$$

with

$$\mathbf{D}=\mathbf{N}^*+\mathbf{N}-\mathbf{N}(\mathbf{B}+\mathbf{N})^{-1}\mathbf{N}$$

and

$$\mathbf{E}=\mathbf{B}+\mathbf{N}-\mathbf{N}(\mathbf{N}^*+\mathbf{N})^{-1}\mathbf{N}.$$

Similarly, the posterior mean is (with $\mathbf{b}$ a solution to $\mathbf{X}'\mathbf{X}\mathbf{b}=\mathbf{X}'\mathbf{Y}$)

$$E\left(\begin{bmatrix}\boldsymbol{\beta}\\ \boldsymbol{\beta}^c\end{bmatrix}|\mathbf{Y}\right)$$

$$=\begin{bmatrix}\mathbf{D}^{-1} & -\mathbf{D}^{-1}\mathbf{N}(\mathbf{B}+\mathbf{N})^{-1}\\ -\mathbf{E}^{-1}\mathbf{N}(\mathbf{N}^*+\mathbf{N})^{-1} & \mathbf{E}^{-1}\end{bmatrix}\begin{bmatrix}\mathbf{N}^*\mathbf{b}^*+\mathbf{N}\mathbf{b}\\ \mathbf{N}\mathbf{b}\end{bmatrix}$$

$$=\begin{bmatrix}\mathbf{D}^{-1}\left(\mathbf{N}^*\mathbf{b}^*+\left[\mathbf{N}-\mathbf{N}(\mathbf{B}+\mathbf{N})^{-1}\mathbf{N}\right]\mathbf{b}\right)\\ \mathbf{E}^{-1}\mathbf{N}(\mathbf{N}^*+\mathbf{N})^{-1}\mathbf{N}^*(\mathbf{b}-\mathbf{b}^*)\end{bmatrix}. \tag{9.9}$$

Notice, first, that the posterior mean of $\boldsymbol{\beta}$ is, in the usual way, a weighted average of the prior mean $\mathbf{b}^*$ and the sample mean $\mathbf{b}$. Whereas ordinarily the sample mean receives weight $\mathbf{N}$, its weight is here reduced to $\mathbf{N}-\mathbf{N}(\mathbf{B}+\mathbf{N})^{-1}\mathbf{N}=\mathbf{N}(\mathbf{B}+\mathbf{N})^{-1}\mathbf{B}$. That is to say, we discount the evidence provided by contaminated (or potentially contaminated) experiments. The discount depends on $\mathbf{B}$, the prior precision of the experimental bias $\boldsymbol{\beta}^c$. As $\mathbf{B}$ grows the posterior parameters converge to their values in an uncontaminated experiment.

The posterior mean of $\boldsymbol{\beta}^c$ is a matrix-weighted average of zero and $(\mathbf{b}-\mathbf{b}^*)$. When $\mathbf{b}$ exceeds $\mathbf{b}^*$ we conclude in part that $\boldsymbol{\beta}$ exceeds $\mathbf{b}^*$ (in the matrix-weighted average sense) but we prejudice the posterior distribution toward $\mathbf{b}^*$ more than in the case of a true model. We adjust for this by moving the distribution of $\boldsymbol{\beta}^c$ from the origin; that is, part of the excess of $\mathbf{b}$ over $\mathbf{b}^*$ is attributed to experimental bias, part to large $\boldsymbol{\beta}$.

Given a diffuse prior for $\boldsymbol{\beta}$, with $\mathbf{N}^* = \mathbf{0}$, the posterior parameters are

$$E\left(\begin{bmatrix}\boldsymbol{\beta}\\ \boldsymbol{\beta}^c\end{bmatrix}|\mathbf{Y}\right)=\begin{bmatrix}\mathbf{b}\\ \mathbf{0}\end{bmatrix}$$

$$V\left(\begin{bmatrix}\boldsymbol{\beta}\\ \boldsymbol{\beta}^c\end{bmatrix}|\mathbf{Y}\right)=\sigma^2\begin{bmatrix}\mathbf{N}^{-1}+\mathbf{B}^{-1} & -\mathbf{B}^{-1}\\ -\mathbf{B}^{-1} & \mathbf{B}^{-1}\end{bmatrix}$$

where we have used the fact that

$$\left[\mathbf{N}-\mathbf{N}(\mathbf{B}+\mathbf{N})^{-1}\mathbf{N}\right]^{-1}=\mathbf{N}^{-1}+\mathbf{B}^{-1}.$$

This result is of interest, since it suggests that the least squares estimator $\mathbf{b}$ is the best we can do when the direction of the bias is unknown ($E\boldsymbol{\beta}^c=\mathbf{0}$). The posterior variance matrix, however, differs importantly from the usual OLS result $\sigma^2\mathbf{N}^{-1}$. We must add to this matrix another matrix $\sigma^2\mathbf{B}^{-1}$, which is just the prior variance of $\boldsymbol{\beta}^c$. Most importantly, as sample size increases, this second term does not decay away and thus becomes a lower bound to the variance of $\boldsymbol{\beta}$. In words, if you arrive at a sampling experiment with no knowledge of $\boldsymbol{\beta}$, you can never know more about $\boldsymbol{\beta}$ than you claim to know about $\boldsymbol{\beta}^c$. Of course, you cannot measure more accurately than your measuring instrument is capable of. More than that, the capability of the measuring instrument is not disclosed in the process of measuring (since $\mathbf{B}$ is fixed before measurement commences).

Once the sampling uncertainty $\sigma^2\mathbf{N}^{-1}$ becomes small relative to the misspecification uncertainty $\sigma^2\mathbf{B}^{-1}$, continued sampling of this process is, essentially, a waste of time. Additional information may be gathered only by improved experimentation, that is, by smaller σ^2 or larger $\mathbf{B}$. Larger $\mathbf{B}$ is a pure prior concept, whereas reduction in σ^2 is evidenced through smaller error sum of squares; thus the latter, when it is not offset by smaller $\mathbf{B}$, seems to be the only unambiguous method of improving our knowledge of the process parameters.

In the nonexperimental sciences, the possibility of improving an "experiment" is, by definition, excluded. Researchers implicitly do what they regard to be the next best thing: they treat the R^2 as an indicator of the quality of experimental control and discount results when R^2s are small. The extent to which this discounting is appropriate depends on how it is done. Since R^2 does map into an estimate of σ^2, R^2 may give an indication of the absolute misspecification uncertainty $\sigma^2\mathbf{B}$. However, the percentage understatement of the uncertainty is a function of sample size ($\mathbf{N}^{-1}$) and not of σ^2. In this sense, the R^2 is not an indicator of experimental control. Independent of R^2, the OLS variance $\sigma^2\mathbf{N}^{-1}$ accurately summarizes the uncertainty for small samples but understates the uncertainty for large samples. This is simply because it ignores the misspecification uncertainty $\sigma^2\mathbf{B}^{-1}$, which is negligible compared to the sampling uncertainty $\sigma^2\mathbf{N}^{-1}$ in small samples, but not in large.

9.4 Inference with Data-instigated Models

In this section we discuss the inferences that are legitimate when new variables are added to regression equations. As we have suggested previously, it is not the stopping rule that should cause suspicion. Rather, the error that is potentially made is that in adding a new variable to the equation the researcher implicitly changes his priors about various parameters. He left the variable out in the first place because he thought it did not belong, and to be consistent he must have a prior on the new coefficient that concentrates the probability in the neighborhood of zero. This automatically "discounts" the evidence implied by the new regression model in the sense that the posterior distribution of the new regression coefficient is pushed toward the origin.

Consider the following hypothesis:

$$\mathbf{Y}|\mathbf{X},\mathbf{Z}=\mathbf{X}\boldsymbol{\beta}+\mathbf{X}\boldsymbol{\beta}^c+\mathbf{Z}\boldsymbol{\gamma}+\mathbf{Z}\boldsymbol{\gamma}^c+\boldsymbol{\varepsilon} \tag{9.10}$$

with $\boldsymbol{\varepsilon}\sim N(\mathbf{0},\sigma_\varepsilon^2\mathbf{I})$ and where $\mathbf{X}$ and $\mathbf{Z}$ may be matrices. A simpler working hypothesis may be obtained by marginalizing out $\mathbf{Z}$. Assume a linear regression function

$$\mathbf{Z}=\mathbf{X}\mathbf{R}+\mathbf{U} \tag{9.11}$$

with $\mathbf{U}$ having a matrix-normal distribution such that if $\mathbf{u}_i$ is a row of $\mathbf{U}$, $\mathbf{u}_i\sim N(0,\boldsymbol{\Sigma}_{UU})$, $\mathbf{u}_i$ independent of $\mathbf{u}_j$ and $\boldsymbol{\varepsilon}$. The model conditioned on $\mathbf{X}$ would then be

$$\begin{aligned}\mathbf{Y}|\mathbf{X}&=\mathbf{X}\boldsymbol{\beta}+\mathbf{X}\boldsymbol{\beta}^c+\mathbf{X}\mathbf{R}\boldsymbol{\gamma}+\mathbf{X}\mathbf{R}\boldsymbol{\gamma}^c+\mathbf{U}\boldsymbol{\gamma}+\mathbf{U}\boldsymbol{\gamma}^c+\boldsymbol{\varepsilon}\\ &=\mathbf{X}\boldsymbol{\beta}+\mathbf{X}\boldsymbol{\Gamma}+\mathbf{e}\end{aligned} \tag{9.12}$$

which is in the form of the usual contaminated model but with the constraints

$$\boldsymbol{\Gamma}=\boldsymbol{\beta}^c+\mathbf{R}\boldsymbol{\gamma}+\mathbf{R}\boldsymbol{\gamma}^c \tag{9.13}$$

$$\mathbf{e}=\mathbf{U}\boldsymbol{\gamma}+\mathbf{U}\boldsymbol{\gamma}^c+\boldsymbol{\varepsilon}, \tag{9.14}$$

with parameters

$$\sigma_e^2=(\boldsymbol{\gamma}+\boldsymbol{\gamma}^c)'\boldsymbol{\Sigma}_{UU}(\boldsymbol{\gamma}+\boldsymbol{\gamma}^c)+\sigma_\varepsilon^2 \tag{9.15}$$

where $\boldsymbol{\Sigma}_{UU}$ is a contemporaneous covariance matrix of the (matrix) random variable $\mathbf{U}$.

Inferences about the parameters may be made as implied by either (9.10) or (9.12), depending on whether $\mathbf{Z}$ is observed. This is true even when the

analysis proceeds sequentially with $\mathbf{Z}$ observed depending on the least-squares outcome $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$. [There is a tendency to think that the data evidence is contaminated by the stopping rule. Suppose, for example, we decide to observe $\mathbf{Z}$ if $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ has any negative elements. Since this rule depends only on the data and not at all on the parameters, it is noninformative. Classically, of course, the resulting estimator is biased, and the stopping rule would be regarded as a source of contamination.]

If we begin our analysis with the simple model (9.12) and therefore a probability assignment to $(\boldsymbol{\beta},\boldsymbol{\Gamma},\sigma_e^2)$, the fact that the more complex working hypothesis (9.10) is lurking in the background is quite irrelevant. Within the confines of noninformative stopping rules, we may decide to observe $\mathbf{Z}$, and to expand to the fuller model at any time. We are *not* free, however, to assign any distributions to $(\boldsymbol{\beta},\boldsymbol{\beta}^c,\boldsymbol{\gamma},\boldsymbol{\gamma}^c,\sigma_e^2)$, since we already have an assignment on $(\boldsymbol{\beta},\boldsymbol{\Gamma},\sigma_e^2)$, functions (9.13) and (9.15) defined on the expanded parameter space. As long as we satisfy these constraints, we *are* free to alter the model as we choose. Thus postdata model construction becomes fully legitimate.

Although these constraints are conceptually straightforward, they are not easy to implement. When $\mathbf{R}$ is known, however, the implication of constraint (9.13) is straightforward. Under a normality assumption, only the first two moments are of interest.

$$E(\boldsymbol{\Gamma}) = E(\boldsymbol{\beta}^c) + \mathbf{R}E(\boldsymbol{\gamma}+\boldsymbol{\gamma}^c) \tag{9.16}$$

$$V(\boldsymbol{\Gamma}) = V(\boldsymbol{\beta}) + \mathbf{R}V(\boldsymbol{\gamma}+\boldsymbol{\gamma}^c)\mathbf{R}' \tag{9.17}$$

where we have assumed the independence of $\boldsymbol{\beta}^c$ and $(\boldsymbol{\gamma},\boldsymbol{\gamma}^c)$. We would typically set $E(\boldsymbol{\Gamma})$, $E(\boldsymbol{\beta}^c)$ and $E(\boldsymbol{\gamma}^c)$ to zero to indicate expected unbiased experiments, and (9.16) implies that $\boldsymbol{\gamma}$ must have prior mean zero. The extent to which we allow $\boldsymbol{\gamma}$ to wander from zero in response to the data evidence is determined by the variance $V(\boldsymbol{\gamma})$, which is constrained by (9.17). The larger $V(\boldsymbol{\Gamma})$ is, the more we discount the evidence derived from the regression of $\mathbf{Y}$ on $\mathbf{X}$ (see Section 9.3). But large $V(\boldsymbol{\Gamma})$ also allows us to assign large $V(\boldsymbol{\gamma})$, and this allows $\boldsymbol{\gamma}$ to wander from zero when $\mathbf{Y}$ is regressed on both $\mathbf{X}$ and $\mathbf{Z}$. Loosely speaking, if we are willing to discount the evidence collected when we regress $\mathbf{Y}$ on $\mathbf{X}$, we may believe the evidence collected when we regress $\mathbf{Y}$ on $\mathbf{X}$ and $\mathbf{Z}$. Note, by the way, that the constraints become inoperative for orthogonal data, $\mathbf{R}=\mathbf{0}$, and we are thus completely free to add in orthogonal variables.

See Leamer (1974) for further discussion of the implications of these constraints.

9.5 An Example: Bode's Law

An interesting example of a data-instigated model is the numerological relationship discovered by Titius describing the mean distance of a planet

from the sun as a simple function of the planet order, specifically, by the simple geometric progression

$$d_n = 4 + 3(2^n) \tag{9.18}$$

where d_n is the distance from the sun to the nth planet from the sun. For the first eight planets this implies the mean distances of 4, 7, 10, 16, 28, 52, 100, 196 (using $n = -\infty, 0, 1 \ldots$). The seven planets known in 1800 had mean distances of 3.9, 7.2, (10), 15.2, 52, 95, 192, with the earth's distance arbitrarily set to 10. These numbers fit the numerical sequence remarkably well, with the exception of a missing planet 28 units from the sun. The very real decision problem of astronomers at that time was whether this evidence was compelling enough to warrant a search for the missing planet in the region suggested by the relationship.

Surprisingly enough, Bode and five other German astronomers, searching the heavens at roughly 28 units from the sun, on January 1, 1801, did indeed find the small planet, Ceres, and since then dozens of other small planets have been found and are hypothesized to be the fragments of a single larger planet.[4] The "law" was given Bode's name perhaps because his discovery of the missing planet makes the law ever so much more believable than it would have been otherwise. The law was instigated by the observations available up to 1800, and we tend, properly I think, to discount the evidence implied by those observations. The single observation that was not known at the time and that could not have instigated the model is taken as essentially the only data point relevant for testing. It intuitively lends considerable believability to the law, and application of Bayes rule leads to the same conclusion (using as an alternative hypothesis almost any other plausible hypotheses about the dispersion of the planets around the sun).[5]

But it is not enough to observe that Bode's discovery considerably adds to the believability of the law. Beliefs prior to the observation of Ceres also partly determined the posterior belief. In order to determine if Bode's Law is believable today we must determine what degree of believability it had prior to 1801. We must assess the uncertainty, allowing for the fact that the law was instigated by the first seven observations.

It is interesting to observe that the statisticians Good (1969) and Efron (1971) seem to be concerned primarily with the construction of interesting alternative hypotheses, with little argument over the appropriateness of statistical theory in general. Blyth (1971, p. 566) comments on this, "The Efron and Good tests seem to me invalid because they are based on the same data that suggested both hypotheses." He takes a pessimistic posi-

[4]This is Polanyi's (1964) version of the facts. Good (1972) attributes the discovery to Piazzi and describes Bode as merely a publicist.

[5]Almost any other hypothesis places low probability on finding a planet in this region. See Good (1969) or Efron (1971).

tion, "And it would appear that any real test of this would have to be based on future observations."

We claim to have a way of characterizing the uncertainty about Bode's law that allows for the fact that it was data instigated. We may include all the observations available today, Bode's eight planets plus Neptune and Pluto. As it turns out, neither of these last two planets obeys Bode's law very well.

To perform the analyses described in the previous section, we must phrase the postdata model construction aspect of Bode's law as the addition of variables to a linear model. There are, apparently, three "discoveries" from the data set that are candidates for the postdata label:

1. That distance depends on the order
2. That distance depends nonlinearly on order
3. That there are three "outliers"

It seems to me that any model of planetary distance would include order as an explanatory variable. The significant postdata discovery is that distance depends nonlinearly on order. The discarding of the three outliers represents a second step we do not explore here.

Let us phrase the model in terms of our linear regression parameters of the previous section as

y_n = distance from planet $n-1$ to planet n (in units of Sun-Earth distance)

x_0 = constant

$x_1 = n$, the planet order

$z = n^2$

The first and second phase regressions are

Phase I: $y_n = \beta_0 + \beta_1 n + \Gamma_0 + \Gamma_1 n + e_n$
Phase II: $y_n = \beta_0 + \beta_1 n + \beta_0^c + \beta_1^c n + \gamma n^2 + \gamma^c n^2 + \varepsilon_n$
with the regression of Z on X being

$$n^2 = r_0 + r_1 n + u, \qquad n = 1, \ldots, 10. \tag{9.19}$$

Note, of course, that Bode's law is distorted to fit it into our framework. I do not think that the distortion has important substantive implications, however. I also substitute sample estimates for σ_e^2 and σ_ε^2.

In this case, it is possible to calculate r_0 and r_1 with certainty, and the linear system (9.13) becomes

$$\begin{bmatrix} \Gamma_0 \\ \Gamma_1 \end{bmatrix} = \begin{bmatrix} \beta_0^c \\ \beta_1^c \end{bmatrix} + \begin{bmatrix} r_0(\gamma + \gamma^c) \\ r_1(\gamma + \gamma^c) \end{bmatrix} \tag{9.20}$$

where $r_0 = -22$, $r_1 = 11$.

I choose to ignore the constraint on the variances, Equation (9.15), on the basis that the assumed vagueness of the prior distribution of σ_e^2 effectively eliminates the constraint. If we take all prior means to be zeroes, the constraints (9.20) under a normality and an independence assumption are satisfied when

$$\begin{aligned} V(\Gamma_0) &= V(\beta_0^c) + r_0^2(V(\gamma) + V(\gamma^c)) \\ V(\Gamma_1) &= V(\beta_1^c) + r_1^2(V(\gamma) + V(\gamma^c)). \end{aligned} \tag{9.21}$$

We require that the researcher who employs the phase I regression must at that time select $V\Gamma_0$ and $V\Gamma_1$, where Γ_0 and Γ_1 are the first-phase experimental-bias coefficients. Relatively small values of these variances imply relatively small discounting of the first phase result but also imply through (9.21) relatively tight distributions of γ and γ^c, or equivalently, relatively large "discounting" of the second-phase regression.

To begin, let us take a look at the unadorned regressions (with standard errors in parentheses)

$$\begin{array}{llll} y_n = -32.01 & +13.0n & & \\ \quad (13.5) & \quad (2.18) & & \\ R^2 = .82 & \bar{R}^2 = .80 & \text{d.f.} = 8 & \text{D.W.} = .91 \end{array} \tag{9.22}$$

$$\begin{array}{llll} y_n = 3.25 & -4.63n + 1.6n^2 & & \\ \quad (18.7) & \quad (7.8) \quad (.7) & & \\ R^2 = .90 & \bar{R}^2 = .87 & \text{d.f.} = 7 & \text{D.W.} = 1.6 \end{array} \tag{9.23}$$

where D.W. indicates the Durbin-Watson statistic.

We assume that the researcher first runs (9.22) and then notes peculiarities in the residual pattern, indicated especially by the Durbin-Watson statistic of .91. To rid his model of those peculiarities, he adds the n^2 term and refits. It is pretty clear from (9.23) that he obtains a substantially improved fit. The variable n^2 effectively wipes out any apparent influence of the variable n. Is this, however, real or manufactured evidence?

I claim that before the regression equation (9.22) is estimated, one must decide how much he will believe the result. He does this by selecting $V(\beta_0)$, $V(\beta_1)$, $V(\Gamma_0)$, $V(\Gamma_1)$. Consider the following three cases:

	$V(\beta_0)$	$V(\beta_1)$	$V(\Gamma_0)$	$V(\Gamma_1)$
Case 1	10^4	10^4	10^4	10^4
Case 2	10^4	10^4	10^2	10^2
Case 3	10^4	10^4	10	10

In all cases the researcher is very uncertain about the theoretical coefficients β_i. As we proceed from case 1 to case 3, he is increasingly confident about the quality of the experiment.

The phase I posterior means and standard errors implied by these priors may be found in Table 9.2. One is expected at this stage to choose one of these cases. If you wish to believe the sample result you must select Case 3. At the other extreme, you may select Case 1 and discount the sample very significantly.

Having committed oneself during phase I to one of the three cases, one has a restricted menu of things he can believe following phase II. These are given for the three cases in Table 9.3. The constraints (9.20) imply

$$V(\gamma)+V(\gamma^*)\leqslant m=\min\left(\frac{V\Gamma_0}{r_0^2},\frac{V\Gamma_1}{r_1^2}\right).$$

The small letters in Table 9.3 indicate

$$\begin{aligned}&(a)\,V(\gamma)+V(\gamma^*)=.01m\\&(b)\,V(\gamma)+V(\gamma^*)=.5m\\&(c)\,V(\gamma)+V(\gamma^*)=.99m.\end{aligned}$$

That is, for distribution (c), the coefficient $(\gamma+\gamma^*)$ has the largest variance and therefore the greatest freedom to vary from zero. In all cases we have set $V(\gamma)=99V(\gamma^*)$, that is, we are allocating almost all the evidence to the theoretical coefficient γ. (More on this point shortly.)

Notice in Table 9.3 that the phase II posterior distributions for case 1 are effectively the phase II sample regression function, whereas the distributions for case 3 are effectively the phase I sample regression function. In words, if you were willing to completely discount the evidence generated in phase I (case 1) you may now believe in the nonlinearity of the function. If, on the other hand, you thought you were getting evidence about the linear term in the first phase, no significant evidence about the nonlinearity of the function was generated during the second phase. The reason for

Table 9.2

Posterior Means and Standard Errors
(Standard errors in parentheses)

	β_0	β_1	Γ_0	Γ_1
Sample	−32.01	13.0		
	(13.5)	(2.18)		
Case 1	−15.8	6.4	−15.8	6.4
	(71)	(71)	(71)	(71)
Case 2	−31.1	12.8	−.31	.13
	(16.6)	(10.2)	(10.0)	(10.0)
Case 3	−31.4	12.9	−.03	.013
	(13.7)	(3.8)	(3.16)	(3.16)

Table 9.3

Phase II Posterior Means and Standard Errors

		β_0	β_1	γ	β_0^c	β_1^c	γ^c
Sample		3.25	−4.63	1.6			
		(18.7)	(7.8)	(.7)			
Case 1	*a*	−1.7	−.67	1.3	−1.6	−.64	.013
		(69)	(71)	(.63)	(69)	(71)	(.14)
	b	1.1	−2.1	1.5	.54	−1.79	.015
		(58)	(69)	(.74)	(58)	(69)	(.32)
	c	2.3	−2.4	1.55	0	−1.77	.016
		(18.3)	(66.8)	(.81)	(.089)	(66)	(.045)
Case 2	*a*	−29	12.1	.067	−.27	.12	0
		(15)	(10)	(.14)	(9.4)	(9.8)	.014
	b	−25	9.7	.29	−.13	.085	.003
		(14)	(9.9)	(.29)	(7.1)	(9.3)	(.003)
	c	−21	7.5	.48	0	.06	.005
		(13)	(9.7)	(.38)	(.006)	(8.6)	(.05)
Case 3	*a*	−31	12.9	.007	−.03	.01	0
		(11)	(3.6)	(.045)	(2.99)	(3.12)	(.005)
	b	−31	12.5	.03	−.015	.011	0
		(11.2)	(3.6)	(.10)	(2.2)	(2.96)	(.01)
	c	−30	12.2	.07	0	.009	0
		(11)	(3.6)	(.14)	(.0009)	(2.7)	(.014)

this seems clear. Since n and n^2 are highly correlated, the only way a regression of y on n alone could give us evidence about how n affects y is if n^2 simply does not belong in the equation. The sample evidence that n^2 does belong simply is not enough to overcome that prejudice.

Let us now suppose that you did discount the evidence in phase I, that is, that you selected case 1. As I have indicated, the phase II posteriors assign almost all of the evidence to γ rather than γ^*. Thus although it is possible to believe in the nonlinearity of the function you may instead decide to discount this evidence by reallocating the prior variance from $V(\gamma)$ to $V(\gamma^*)$. The advantage of doing this is clear: it greatly increases your flexibility in phase III. In the absence of a "deep" model that encourages me to commit myself to this peculiar nonlinearity, it seems wise to maintain as much flexibility as possible.

9.6 Conclusion

It is possible to construct a formal decision-theoretic solution to the problem of choice of variables that allows for reconsideration of that choice *after* data have been observed. Such a solution requires us first to

identify all the potential variables. We then must provide subjective probability distributions for both the parameters that govern the generation of these variables and also for the parameters that link these variables to the dependent variable under study. For many problems, unfortunately, the identification and assessment problems jointly constitute the most significant costs of dealing with other variables. The observation costs are trivial in comparison, and once we bear the former costs, we almost certainly want to observe and process the complete data set.

I am proposing, therefore, that we behave only *as if* we were formally solving this decision problem. We identify through economic theory and/or introspection certain variables that are potentially important. This is, essentially, the first phase in the formal decision problem. The left-out variables are not, however, formally identified. Instead, we summarize their influence in a contamination parameter β^c, the prior on which essentially determines the extent to which we are committed to "believe" the regression result.

Just as if we were solving the formal decision problem, we may decide to observe other variables because of either low R^2, peculiar residuals, or peculiar coefficient estimates. At the very least, the probability distribution over the new parameters must imply the original on β^c.

This constraint prejudices the coefficients on the new variables to zero; that is, by leaving the variable out of the equation to begin with, thereby expressing interest in a "false" model, we have revealed that we think the variable will not significantly distort inferences on the other parameters. This is the case if the regression coefficients are negligible or if the added variables are orthogonal to the original set (or a combination).

This analysis can obviously be carried on to additional stages. At each stage the constraints on the new variables become more severe. Incidentally, the order in which variables are added to the equation influences the interpretation of the evidence. For example, two researchers may end up with the same set of explanatory variables. If these variables have been added to the equations in different orders, then the researchers have revealed different priors and must also make different interpretations of the data evidence.

CHAPTER **10**

SYSTEMATIC JUDGMENTAL ERRORS

A theme of this book is that judgment is the critical input into the analysis of nonexperimental data. Systematic errors in the formation of judgment may lead to significant systematic errors in the interpretation of evidence. The elimination of systematic judgmental errors is thus highly desirable. As a first step in that direction, we may identify in this chapter what seem to be the more consequential systematic errors.

10.1 "Explaining Your Results" as Access-Biased Memory

QUESTION. What do the following quotations have in common?

The stock market reacted today to the favorable news released by the Commerce Department that our fourth-quarter trade surplus established a new record.

Casey Stengel demonstrated again his lack of managerial talent by replacing pitcher Whitey Ford by a wild Ryne Duren, who proceeded to walk in the winning run.

The negative estimated effect of the price of butter on the consumption of wheat is fully consistent with the fact that bread and butter are jointly consumed.

Answer. (a) All three statements are "explanations" of certain events. In the terminology of probability, where A and B are events, an explanation of an event