CHAPTER 6

SIMPLIFICATION SEARCHES

6.1 Simplification for Conditional Prediction 208
6.2 Causally Constrained Conditional Predictions 214
6.3 Simplification for Control 217
6.4 Conclusion 223

In the two previous chapters we have considered specification searches that are intended to introduce into a data analysis uncertain prior information. Hypothesis-testing searches arise when more than one model or hypothesis receive positive a priori probability. Interpretive searches involve prior density functions that, although allocating zero probability to all but one hypothesis, do concentrate the prior probability in certain regions of the parameter space. In the case of hypothesis-testing searches the statistical testing selects among a set of hypotheses with no presumption that in a large sample one of the hypotheses will be favored. In contrast, interpretive searches recognize that in a sufficiently large sample the most general hypothesis will necessarily be favored. The intent is not to select among legitimately competing models but rather to "improve" the estimate of the parameters by using an a priori estimate when the data evidence is too weak to yield a reliable sample estimate.

In this chapter we discuss a variety of search that has yet another motivation: simplification. The most general models appropriate for inference with nonexperimental data are usually so cluttered with variables of an incidental nature that they are nearly impossible to comprehend directly. It is thus incumbent on the researcher to find vehicles for communication of his results. He might, for example, focus his discussion on a particular parameter of special interest or perhaps on a linear combination of parameters. Alternatively, the researcher might seek from the data an indication

of the "important" variables. We call this a simplification search.

Thus the function of simplification search is not to ask if a restricted specification is true, nor to ask if a restricted specification might lead to better parameter estimates, but rather to ask if a restricted specification that is undeniably simpler and more easily understood is not also "significantly" inferior to the more general model for some hypothetical or real decisions. If it is, we reject the hypothesis that the benefits of the restriction outweigh the costs.

Formal analysis of simplification problems requires a precise definition of the costs and benefits of simplicity. The costs of simplicity may be assessed in the context of some hypothetical decision problems, but the benefits are likely to elude precise definition. Consequently we concentrate our formal attention on the cost side, but we first comment informally on the likely benefits from simplification.

Justifications for simplicity can usefully be divided into two categories. The first makes a "metaphysical" reference to the inherent simplicity of Nature, or at least to man's belief in such. The second category of justifications accepts a complex Nature but rests simplicity on the finiteness and fallibility of Man's perceptive and reasoning faculties. Briefly, simplicity is preferred because "Nature is simple" or because "Man is simple."

The "Nature is simple" hypothesis has, I think, little support among philosophers and statisticians. Jeffreys' is a widely cited exception. He writes [1961, p. 4] "It is asserted, for instance, that the choice of the simplest law is purely a matter of economy of description or thought, and has nothing to do with any reason for believing the law... I say, on the contrary, that the simplest law is chosen because it is the most likely to give correct predictions; that the choice is based on a reasonable degree of belief; ..."

Jeffreys is asserting not only that constrained hypotheses should be assigned positive probability but also that they ought to be assigned greater prior probability than any alternative, more complex hypotheses. Such a preference for simple models might be inductively derived. Simple hypotheses could usually yield better predictions. But it is not enough to observe merely that people act as if simple models had a greater degree of believability. Any observed preference for simple models may derive not from the inherent superior believability of parsimonious models but rather from the undeniable difficulties encountered in working with complex descriptions of reality. Nor do I know of any proper empirical evidence to support the assertion that simpler models generally yield better predictions. There is the oft-told story of overfitting in which a naive researcher fits a polynomial of degree $T-1$ given T pairs of observations (y_i, x_i). This

; yield inferior predictions relative to a polynomial of
e. But that can be fully remedied by assigning a proper
to the parameters of the higher-degree polynomial. I
nple as an illustration of the illogic of using a prior that is
ated by the data evidence when the data evidence simply
it. It is hardly evidence in favor of simpler models.
d in the chapter on hypothesis testing that I know of few
would assign positive probability to a restricted (simple)
I can find nothing that compels me to favor the simpler
gnment of probability. It is the other set of reasons for
find persuasive: Simplicity is desirable because it is
transmission and accumulation of knowledge. It greatly
nication between and among observers and theorists. A
heory that might take years to filter accurately to other
transmit rapidly (but inaccurately) if it is simplified.
at they regard to be superior knowledge for their own
likely to engage in this kind of marketing activity. Many
duct may never realize that there is more to the theory
logans used to advertise it.
ave argued in various ways that simplicity encourages
(1972) favors simpler models because they are more easily
ich might at first glance seem to hasten the rejection of
This would be true if the simple model were assigned
ity, but if such a model is derived from a more complex
apparently falsifying evidence can be taken to mean only
version does not work under all conditions. In that
ity protects a system of belief from falsification and
tly impedes progress. On the other hand, protection of a
from potential falsification is an essential feature of
according to Kuhn (1962). Filling in the details of a theory
t all its implications requires a vast amount of tedious
r would hardly be performed by doubters or even agnost-
nagine the value of their efforts overnight crashing to zero.
Nature is simple" nor the "Man is simple" hypothesis
nbiguous definitions or methods of measuring the benefits
e number of uncertain parameters is a possible mechanical
licity, but it cannot be generally satisfactory. If we take (as
as a consequence of man's and society's shortcomings, the
nplicity necessarily changes from social millieu to social
us impossible and even undesirable to define simplicity
e instead must content ourselves with the satisfaction that
in any social information process can know themselves
is and what it is not.

The prototypical example of this is the construction of a map (Polan
1964). We may take as a theory of the world an enormously detailed glo
which identifies every object down to the smallest grain of sand. T
complexity of this theory effectively prevents us from using it for a
purpose whatsoever. Instead, we simplify it in the form of a set of maps
use one map to find my way to the subway station, another to select t
station at which to depart. The pilot of the airplane uses yet another
navigate from Boston to Washington. Each map is a greatly simplifi
version of the theory of the world; each is designed for some class
decisions and works relatively poorly for others.

The construction of a language is another good example of a simplific
tion problem. The number of aurally and visually discernible words a
word patterns is absolutely enormous, perhaps limitless. With as f
characters as are in our alphabet we could form $26^5 > 10^7$ distinct fi
letter words. Such a vocabulary would be beyond the reach of even t
most verbally talented, and the mistaken use of words used infrequen
would greatly distort intended communications. A highly limited voc
bulary likewise distorts communications by not distinguishing one compl
communication from another, for example, the American overuse of t
word "nice" to describe a wide variety of generally pleasing responses
environmental stimuli. An optimal vocabulary ideally solves the trade
between miscommunications from too few words and miscommunicatio
from too many.

Incidentally, there is a great danger that a simple language is not only
vehicle for communication but that it also creates an impoverished real
of its own. The art of communication forces an awareness of reality, a
the more subtle is the language, the more practice one obtains in d
tinguishing subtleties. Conversely, a coarse language creates no situatic
for exercising one's capacities to distinguish subtleties, and those facult
may atrophy like any unused muscle. We may, in fact, be unable anym
to distinguish the great variety of sensations we refer to as "nice." T
may also be the case in the communication of scientific theories. We m
come erroneously to believe in the simplicity of Nature because that is t
way scientific theories are communicated.

I do not think it is possible to define simplicity, which is to say in t
language of decision theory that it is difficult to compute precise benef
or precise costs from any simplification. In this chapter the cost
simplification is measured in the context of several simple decision prc
lems, but the benefits are not quantified at all. We hope that what we lea
can have implications for more complex and more realistic decisions.

One thing that is important to understand is that simplification is
decision problem which uses as an *input* the current information about t
parameters. When a current sample is available, simplification logica

and is confused with the inferential process, at great peril of a statistical analysis. I would recommend making as ı as possible between inference and decision, by discuss- :tions of a research report first the inferential question of · distributions are influenced by the data and second the of how given various posterior distributions the model l. Incidentally, since a data set is taken as given, any :nts reported in this chapter are necessarily conditional ; thus notationally convenient to suppress the data when ıl probability statements, and it is hoped that this will not For example, the statement $E(\boldsymbol{\beta})=(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ implies nal mean of $\boldsymbol{\beta}$, $E(\boldsymbol{\beta}|\mathbf{Y},\mathbf{X})$, is equal to the least-squares $\mathbf{X}'\mathbf{Y}$.

merits repeating is that a simplified model that might :ly for some decision-making circumstances will be unam- :ptable in others. It is, therefore, essential to identify)blem that is considered. Three examples suggest the diversity of decision problems.

regate consumption of apples C_a and aggregate consump- C_b depend on aggregate GNP Y through the functions ıd $C_b=\alpha_b+\beta_b Y$. If we wish to predict future levels of apples and bananas, may we without great detriment to)nstrain the marginal propensities to consume to equal $_b$ and therefore "remember" only the marginal propensity rather than separate propensities for each fruit?

IP Y is thought to depend on the government deficit G :ock M, $Y=\alpha+\beta G+\gamma M$. If we wish GNP to attain some /e effectively assure that goal by selecting an appropriate rnment deficit G^* while treating money M as if it had no :onversely, might we better control money M and act as if

:onstant-elasticity-of-substitution production function ex- ; a function of capital and labor inputs. If the elasticity of ıual to one, the investment function assuming profit-maxi- is a function of one explanatory variable rather than two. ıation generated by observing the production process, may ces about the investment process acting as if certain of its on special values?

These three examples illustrate, respectively, a prediction problem, a control problem, and an inference problem. Relative to a model of the form $y=z\gamma+w\delta+u$ they ask if we may act *as if* δ were zero (1) if we wanted to predict y, (2) if we wanted to control y, or (3) if we wanted to make inferences about γ. The inference problem is distinguished from the others only in that more data is to be gathered before decisions are to be made. The actual, ultimate decision may, in fact, be either a prediction or a control problem. This is called a presimplification problem, referring to the fact that simplification occurs prior to observation. We make much use of the presimplification notion in Chapter 9, when we discuss postdata model construction.

It is easy to demonstrate the inappropriateness of classical hypothesis testing at a fixed level of significance for the simplification problem. Suppose the prior distribution were diffuse. The only information conveyed by the fact that the hypothesis $\gamma=0$ is or is not rejected at the 5% level of significance is the information that the posterior 95% credible interval includes or does not include the point $\gamma=0$. Thus you may reject the hypothesis $\gamma=0$ even though with near certainty γ is infinitesimal (Figure 6.1a) And you may accept the hypothesis even though with high probability γ is enormous. (Figure 6.1b) It is thus important to distinguish the words "statistically significant" from the words "economically signif

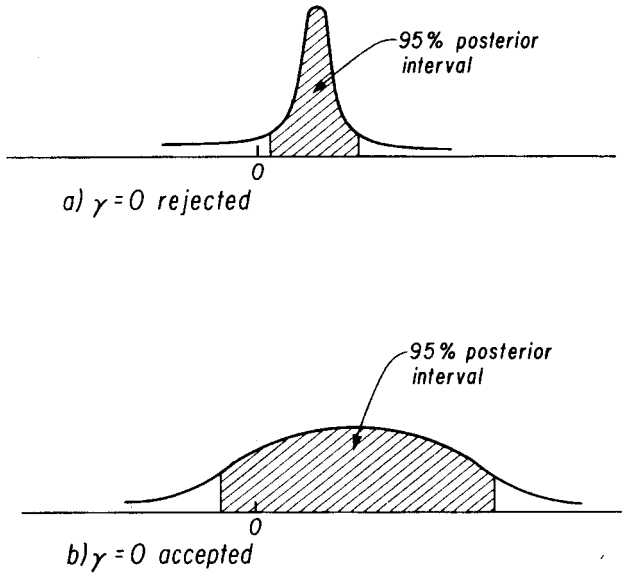

Fig. 6.1 Posterior distributions.

measures the amount of information in the data; the
size of the coefficient in the context of some decision

oint is that classical tests have built into them rather
unwarranted assumptions about the behavior of the
les. Consider again the model $y = z\gamma + w\delta + u$ with δ and
exactly, and with the explanatory variables z and w
iary relationship $w = rz + \varepsilon$, where u and ε are indepen-
bles and r is known. The hypothesis $\delta = 0$ can be used to
l, yielding either $H_0: y = z\gamma$ or $H_0': y = z(\gamma + r\delta)$, where
llows the included variable to play partly the role of the

re the only goal, the hypothesis H_0' is unambiguously
elds a lower expected loss. But for other reasons H_0 may
fication. A simplification is intended to facilitate com-
$'_0$ may be difficult to communicate, since it seems to say
effect of z on y is $(\gamma + r\delta)$ when, in fact, it is only γ. It
least to distinguish the hypotheses "we may act as if δ
may act as if w_T were zero" from the hypothesis "we
for not observing w_T," the former pair implying the
and the latter implying H_0'. Classical hypothesis testing
second-form H_0' with r implicitly estimated in a special
sed subsequently. The other form of simplification is
n 6.2.

of this chapter consists of three sections and a conclu-
ction we report Lindley's (1968) formal decision-theore-
rediction-simplification problem. Among the lessons to
great importance of assumptions about the process that
lanatory variables. In fact, the simplification problem
f not more on the process that generates these variables
regression process linking the dependent variable to the
les. That observation is used in Section 6.2 to argue in
of simplification that makes fewer demands on our
xplanatory variable process and that also communicates
The third section emphasizes the dependence of the
cess on the decision problem under consideration by
s (1968) analysis of a control-simplification problem and
t solution with the prediction-simplification problem.

Simplification for Conditional Prediction

a conditional-prediction problem, consider the two-vari-
ion model

$$y = \alpha + z\gamma + w\delta + u \qquad (6.1)$$

where α, δ and γ are unobservable scalar parameters, u_t is an unobservable error, and y_t, z_t, and w_t are observable variables. Suppose that u_t $(t = 0, \ldots, T)$ is a sequence of independent normal random variables with zero means and known variance σ^2. Let a set of T previous observations of the process be $(\mathbf{Y}, \mathbf{z}, \mathbf{w})$, which together with a multivariate prior distribution for the parameters (α, γ, δ) imply a multivariate posterior distribution with mean

$$E\left([\alpha, \gamma, \delta] | \mathbf{Y}, \mathbf{z}, \mathbf{w}\right) = [\bar{\alpha}, \bar{\gamma}, \bar{\delta}].$$

In making a conditional prediction of the next outcome, say, Y_T, we assume that both the explanatory variables z_T and w_T are potentially observable prior to the announcement of the prediction, hence the adjective "conditional" modifying prediction. It is perhaps obvious, but it is demonstrated here that if the penalty for prediction error is quadratic, the optimal prediction given both z_T and w_T is

$$\hat{y}_T = \bar{\alpha} + z_T\bar{\gamma} + w_T\bar{\delta} \qquad (6.2$$

where $\bar{\gamma}$ and $\bar{\delta}$ are the posterior means of γ and δ. There will, of course, be prediction errors, partly because of the residual error process u_t and partly because the actual values of the parameters α, γ, and δ are not known.

Suppose, now, that we wished to determine if it is worth the expense to observe the second variable w_T. If w_T is not observed, we must estimate it by say, $\hat{w}_T$, and predict y_T as a function of z_T only as

$$\hat{y}_T^* = \bar{\alpha} + z_T\bar{\gamma} + \hat{w}_T\bar{\delta}. \qquad (6.3$$

The squared discrepancy between Equations (6.2) and (6.3) is a measure of the error induced by not observing w_T:

$$(\hat{y}_T - \hat{y}_T^*)^2 = (w_T - \hat{w}_T)^2\bar{\delta}^2 \qquad (6.4$$

Note especially that this error depends on the mean of $\bar{\delta}$ but not on its variance. Note also that in testing the hypothesis $\delta = 0$ in the sense of this chapter, that is, by computing numbers like (6.4), we are partly asking the question "is δ small?" but more importantly we are asking also "how well can we forecast w_T?" To answer the latter question, we must model the process that generates w_T and z_T—there is no way the simplification question can be answered without such a model.

One model (that should, I think, be of little interest to economists operating with time-series data) is the multivariate random model, in which the explanatory variables are treated as if they were drawn randomly from a population with fixed mean vector and covariance matrix. In particular

natrix

$$\mathbf{V}=\begin{bmatrix} v_{11} & v_{12} \\ v_{21} & v_{22} \end{bmatrix}.$$

arameters of this distribution, a prediction of w_T would be conditional regression

$$\hat{w}_T = E(w_T|z_T) = \mu_w + v_{21}v_{11}^{-1}(z_T - \mu_z)$$

variance

$$E\left[(w_T - \hat{w}_T)^2|z_T\right] = v_{22} - v_{21}v_{11}^{-1}v_{12}.$$

parts of these unknown parameters can be used if the prior diffuse and if the number of observations is large.[1] We

$$E(w_T|z_T, \mathbf{z}, \mathbf{w}) = \bar{w} + (\mathbf{w}'\mathbf{M}\mathbf{z})(\mathbf{z}'\mathbf{M}\mathbf{z})^{-1}(z_T - \bar{z}) \tag{6.5}$$

$$-\hat{w}_T)^2|z_T, \mathbf{z}, \mathbf{w}\Big] = \frac{\mathbf{w}'\mathbf{M}\mathbf{w} - (\mathbf{w}'\mathbf{M}\mathbf{z})(\mathbf{z}'\mathbf{M}\mathbf{z})^{-1}(\mathbf{z}'\mathbf{M}\mathbf{w})}{T} \tag{6.6}$$

are the sample means of $\mathbf{w}$ and $\mathbf{z}$ and $\mathbf{M}$ is the matrix that $\mathbf{M} = \mathbf{I} - \mathbf{1}_T T^{-1}\mathbf{1}_T'$. The predicting equations (6.3) and error (6.4) thus become

$$\bar{w} - (\mathbf{w}'\mathbf{M}\mathbf{z})(\mathbf{z}'\mathbf{M}\mathbf{z})^{-1}\bar{z}\bar{\delta} + z_T\left(\bar{\gamma} + (\mathbf{w}'\mathbf{M}\mathbf{z})(\mathbf{z}'\mathbf{M}\mathbf{z})^{-1}\bar{\delta}\right) \tag{6.7}$$

$$= E(w_T - \hat{w}_T)^2\bar{\delta}^2 = \frac{\left[\mathbf{w}'\mathbf{M}\mathbf{w} - \mathbf{w}'\mathbf{M}\mathbf{z}(\mathbf{z}'\mathbf{M}\mathbf{z})^{-1}\mathbf{z}'\mathbf{M}\mathbf{w}\right]\bar{\delta}^2}{T} \tag{6.8}$$

ns may now be made. If the prior for γ and δ were diffuse, eans $\bar{\alpha}$, $\bar{\delta}$, and $\bar{\gamma}$ would be just the least-squares estimates, b_z. The coefficient of z_T in Equation (6.7) would then be $[\mathbf{z})^{-1}b_w$ which is just the estimated coefficient of a regres- alone. Furthermore, the penalty (6.8) can be written as χ^2 is the χ-square value for testing the restriction $\delta = 0$,

$$\chi^2 = \frac{b_w^2\left[\mathbf{w}'\mathbf{M}\mathbf{w} - \mathbf{w}'\mathbf{M}\mathbf{z}(\mathbf{z}'\mathbf{M}\mathbf{z})^{-1}\mathbf{z}'\mathbf{M}\mathbf{w}\right]}{\sigma^2}.$$

dure just described measures the increase in the expected r when w_T is not observed in terms of the usual χ^2 variable 0. As is discussed in detail subsequently, it differs from hesis testing in implicitly defining the significance level as a

rial from Section 3.4, and the diffuse prior assumption with $T^* = 0$, $\mathbf{S}^* = \mathbf{0}$, riance of w_T given z_T is not (6.6) but rather (6.6) times the adjustment

decreasing function of the sample size. What is perhaps more important is the fact that the decision theory logic makes unambiguous the otherwise implicit assumptions about the process that generates the explanatory variables. In particular, classical tests are appropriate only if the explanatory variable vectors are independently drawn from the same population.

Let us now repeat this logic for a general model and for general linear restrictions. Write the linear regression process as

$$\begin{bmatrix} \mathbf{Y} \\ y_T \end{bmatrix} = \begin{bmatrix} \mathbf{X} \\ \mathbf{x}_T' \end{bmatrix}\boldsymbol{\beta} + \begin{bmatrix} \mathbf{u} \\ u_T \end{bmatrix}$$

where $\mathbf{Y}$ and $\mathbf{X}$ are $(T \times 1)$ and $(T \times k)$ matrices and are already observed, where y_T is a future outcome of the process and $\mathbf{x}_T$ is a $k \times 1$ vector of future explanatory variables, and where $[\mathbf{u}', u_T]$ is a $(1 \times (T+1))$ vector of errors with mean zero and covariance $\boldsymbol{\Sigma}$. We are asked to predict y_T given $\mathbf{Y}$, $\mathbf{X}$, and $\mathbf{x}_T$ and in particular to minimize squared prediction error $[y_T - \hat{y}(\mathbf{Y}, \mathbf{X}, \mathbf{x}_T)]^2$ with prediction $\hat{y}$. The expected prediction error can be written as

$$E\left[y_T - \hat{y}(\mathbf{Y}, \mathbf{X}, \mathbf{x}_T)\right]^2 = E\left(E\left(\left[y_T - \hat{y}(\mathbf{Y}, \mathbf{X}, \mathbf{x}_T)\right]^2|\mathbf{Y}, \mathbf{X}, \mathbf{x}_T\right)\right),$$

where the expression in the internal brackets is straightforwardly minimized for every value of $(\mathbf{Y}, \mathbf{X}, \mathbf{x}_T)$ by setting

$$\begin{aligned}\hat{y}(\mathbf{Y}, \mathbf{X}, \mathbf{x}_T) &= E(y_T|\mathbf{Y}, \mathbf{X}, \mathbf{x}_T) = \mathbf{x}_T' E(\boldsymbol{\beta}|\mathbf{Y}, \mathbf{X}, \mathbf{x}_T) + E(u_T|\mathbf{Y}, \mathbf{X}, \mathbf{x}_T) \\ &= \mathbf{x}_T' E(\boldsymbol{\beta}|\mathbf{Y}, \mathbf{X}) + E(u_T|\mathbf{Y}, \mathbf{X})\end{aligned}$$

which is a linear function of $\mathbf{x}_T$. If some part of $\mathbf{x}_T$ is not observed, we assume that the complete vector is predicted as a linear function of that which is observed. That is, letting $\mathbf{x}_T' = (\mathbf{x}_T^I, \mathbf{x}_T^J)$, we assume that $E(\mathbf{x}_T|\mathbf{Y}, \mathbf{X}, \mathbf{x}_T^I) = \mathbf{A}\mathbf{x}_T^I$, and thus the optimal predicting equation becomes

$$\hat{y}(\mathbf{Y}, \mathbf{X}, \mathbf{x}_T^I) = \mathbf{x}_T^{I\prime}\mathbf{A}' E(\boldsymbol{\beta}|\mathbf{Y}, \mathbf{X}) + E(u_T|\mathbf{Y}, \mathbf{X}).$$

Or, to make a long story short, we wish to restrict our attention to predictions linear in $\mathbf{x}_T$

$$\hat{y}(\mathbf{Y}, \mathbf{X}, \mathbf{x}_T) = \mathbf{x}_T'\boldsymbol{\theta}(\mathbf{Y}, \mathbf{X}) \tag{6.9}$$

where the function $\boldsymbol{\theta}$ may be completely free, in which case it is just the posterior mean of $\boldsymbol{\beta}$, or it may be constrained to have certain elements zero to reflect the fact that certain elements of the vector $\mathbf{x}_T$ are not observed prior to the prediction of y_T. Incidentally, Equation 6.9 implicitly includes the $E(u_T|\mathbf{Y}, \mathbf{X})$ term, since $\mathbf{x}_T'$ is assumed to have one element equal to one.

For ease of notation we write the conditional expected value operator $E(\boldsymbol{\theta}|\mathbf{Y}, \mathbf{X})$ henceforth as just $E(\ )$. If $\mathbf{Y}$ and $\mathbf{X}$ are given, $\boldsymbol{\theta}$ is just a vector

the expected loss can be written as

$$\beta'\mathbf{x}_T + u_T - \theta'\mathbf{x}_T)^2$$

$$_T(\beta-\theta)(\beta'-\theta')\mathbf{x}_T$$

$$_T(\beta - E\beta + E\beta - \theta)(\beta - E\beta + E\beta - \theta)'\mathbf{x}_T$$

$$\mathbf{x}_T'(\beta - E\beta)(\beta - E\beta)'\mathbf{x}_T + \mathbf{x}_T'(E\beta-\theta)(E\beta-\theta)'\mathbf{x}_T]$$

$$(\mathbf{x}_T)V(\beta) + (E\beta-\theta)'S(\mathbf{x}_T)(E\beta-\theta) \qquad (6.10)$$

ritten $S(\mathbf{x}_T) = E\mathbf{x}_T\mathbf{x}_T'$. The three terms in the last line of e the irreducible mean-square error Eu_T^2, a penalty for and an additional penalty for $\theta \neq E\beta$, this last term nt of the uncertainty in β.

xpected loss if θ lies in the linear subspace $\mathbf{R}\theta = \mathbf{r}$ is ed posterior loss (6.10) minimized over that linear sub-nization is a simple Lagrangian problem requiring the

$$= (E\beta-\theta)'S(\mathbf{x}_T)(E\beta-\theta) + 2\lambda'(\mathbf{R}\theta - \mathbf{r})$$

That is,

$$\frac{\partial f}{\partial \lambda} = \mathbf{R}\theta - \mathbf{r} = \mathbf{0} \qquad (6.11)$$

$$\frac{\partial f}{\partial \theta} = -S(\mathbf{x}_T)(E\beta-\theta) + \mathbf{R}'\lambda = \mathbf{0}. \qquad (6.12)$$

ed by premultiplying (6.12) by $\mathbf{R}S^{-1}(\mathbf{x}_T)$ and calculating

$$\lambda = (\mathbf{R}S^{-1}(\mathbf{x}_T)\mathbf{R}')^{-1}(\mathbf{R}E(\beta)-\mathbf{r})$$

rted into (6.12) to obtain

$$\beta) - S^{-1}(\mathbf{x}_T)\mathbf{R}'(\mathbf{R}S^{-1}(\mathbf{x}_T)\mathbf{R}')^{-1}(\mathbf{R}E(\beta)-\mathbf{r}). \qquad (6.13)$$

the mean-square error (6.10) becomes

$$(E\beta-\theta) = (\mathbf{R}E(\beta)-\mathbf{r})'(\mathbf{R}S^{-1}(\mathbf{x}_T)\mathbf{R}')^{-1}(\mathbf{R}E(\beta)-\mathbf{r}). \qquad (6.14)$$

om the positive definiteness of the third term in (6.10) ected posterior loss requires $\theta = E\beta$, or by (6.14) that a es expected loss unless $\mathbf{R}E\beta = \mathbf{r}$. A simplification thus ases expected prediction accuracy. We assume that a benefits also, and in the absence of any clear quantita- those benefits, a reasonable number to report is the

percentage increase in the expected posterior loss due to the restriction $\mathbf{R}\theta = \mathbf{r}$:

$$L^2(\mathbf{R},\mathbf{r}) = \frac{(\mathbf{R}E(\beta)-\mathbf{r})'(\mathbf{R}S^{-1}(\mathbf{x}_T)\mathbf{R}')^{-1}(\mathbf{R}E(\beta)-\mathbf{r})}{Eu_T^2 + \operatorname{tr}S(\mathbf{x}_T)V(\beta)}. \qquad (6.15)$$

With suitable definitions of prior vagueness we have simply the least-squares results (remembering that the expected value operator is conditional on $\mathbf{X}$ and $\mathbf{Y}$)

$$E(\beta|\mathbf{Y},\mathbf{X}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

$$V(\beta|\mathbf{Y},\mathbf{X}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}.$$

Further, if the explanatory variables are independent observations from a multivariate process, we would have the $\mathbf{x}_T$ moment matrix be approximately (see Section 3.4)

$$S(\mathbf{x}_T) = E(\mathbf{x}_T\mathbf{x}_T') = \frac{\mathbf{X}'\mathbf{X}}{T}.$$

Using these in (6.13), θ is seen to be simply the constrained least-squares estimate subject to $\mathbf{R}\beta = \mathbf{r}$. Inserting them into (6.14), we obtain the increase in the posterior expected loss to be T^{-1} times a factor that is well known to be the increase in the error-sum squares due to the restriction. The summary L^2 becomes

$$L^2(\mathbf{R},\mathbf{r}) = \frac{T^{-1}\Delta ESS}{\sigma^2\left(1+\frac{k}{T}\right)} \qquad (6.16)$$

where ΔESS is the increase in the error sum of squares, k is the number of coefficients, and T is the number of observations. This contrasts with the classical summary statistic $\Delta ESS/\sigma^2$, which is comparedwith $\chi_p^2(\alpha)$ where p is the rank of $\mathbf{R}$ and α is the significance level. Thus the classical counterpart of (6.16) is the ratio $\Delta ESS/\sigma^2\chi_p^2(\alpha)$. In addition to the nonoccurrence of the factor T^{-1} (which for large T necessitates a "significant" finding), the classical summary differs from the subjectivist summary in depending on p, the number of restrictions. The measure (6.16), incidentally, is just the difference in the multiple correlation coefficients of the two models times a factor that tends to a constant as sample size grows, $L^2(\mathbf{R},\mathbf{r}) = (R^2 - R_0^2)(\mathbf{Y}'\mathbf{M}\mathbf{Y}/T\sigma^2)/(1+kT^{-1})$. Thus if a restriction does not greatly affect the R^2 of an equation, it will not greatly increase the expected squared prediction error.

This rough coincidence of approaches usefully highlights the assumptions that are implicit in the use of classical tests to simplify models for

ourse, there is the diffuse prior assumption. But more vectors of explanatory variables are assumed to be $T+1$ lications of a multivariate process. Autocorrelation and ılar are assumed away. Few economists would find that also worth stating explicitly that the variance in the the t statistic does not measure the uncertainty in the ather the inverse of the conditional variance of the ex- ble. From (6.10), it is seen that uncertainty in the) does not influence choice of restrictions $\theta \neq E(\beta)$.

Causally Constrained Conditional Predictions

spect of the solution discussed in the previous section is ariables are used to forecast correlated unobserved vari- assumption that the correlation structure is maintained. effect (the coefficient) of an observed variable thus includes n estimated coefficient but also a part due to the effect of iables assumed to be correlated with it. Interpreted in terms sting, the change in the R^2 due to a restriction is calculated estricted equation with a reestimated set of coefficients. ıay make good sense if we intend the test to determine the of the restriction, it makes less sense for the simplification ve really mean to say that an effect of an explanatory ligible when it can be predicted well from observation of atory variable? This is the question implicit in a classical t le. A direct application of Webster suggests that a variable red negligible if we can neglect it without substantial loss. means not bothering to predict it or otherwise to make not observing it. As will be shown, this is the question sical beta coefficients and variants thereof.

w from semantics to metaphysics, we can find another same argument. To the extent that the full unconstrained rizes our beliefs about the causal nature of the world, the of the coefficients implicit in hypothesis testing constitutes a rtion of that causality. That is, since included variables play le of dropped variables, the constrained equation is causally less the included variables do, in fact, cause the excluded hey do not, the resulting equation is causally inaccurate. An ıde toward the causality within the explanatory variable set is reporting the original estimates of the coefficients of the ables calculated in the context of the unconstrained equation. described as the direct effects of the included variables on

the dependent variable. Indirect effects depend on other unspecified causal linkages.

An example is in order to make clear these relatively obscure notions. Suppose the equation of motion of a body falling from rest is

$$\frac{d^2y}{(dt)^2} = g(1-\beta z t)$$

where z measures the wind resistance and t the time since departure. The parameter g is acceleration in a vacuum, and terminal velocity is reached at time $t = 1/\beta z$. Suppose, further, that observations on a set of falling bodies are used to estimate the equation of location

$$y = \frac{\hat{g}t^2}{2} - \frac{\hat{\beta}\hat{g}zt^3}{6}, \qquad R^2 = .98$$

where the circumflexes indicate estimated parameters. For this particular sample of falling bodies (including feathers and bowling balls) the following auxiliary regression is also calculated

$$(zt^3) = \hat{r}t^2 - \hat{\alpha}.$$

The model may be simplified to exclude the wind resistance variable. Two alternative simpler models are

$$y = \hat{g}t^2/2 \qquad R^2 = .70 \qquad (6.17$$

$$y = \frac{(\hat{g} - \hat{\beta}\hat{g}\hat{r}/3)t^2}{2} + \frac{\hat{\alpha}\hat{\beta}\hat{g}}{6} \qquad R^2 = .95 \qquad (6.18$$

It is my contention that the first of these equations is the one that should be used to discuss simplification. It asserts that in a vacuum the estimated rate of acceleration is $\hat{g}$ and that for the class of bodies and for the time periods considered, we ought not to think of the experiment as if it were conducted in a vacuum, since one's ability to predict the location of the falling bodies is seriously affected by that assumption. (The R^2 drops from .98 to .7.) Contrast that perfectly clear statement with the statement appropriate for the second equation. "Wind resistance is 'negligible' since by adjusting the rate of acceleration to $\hat{g} - \hat{\beta}\hat{g}\hat{r}/3$ and by acting as if the initial location of the body were $\hat{\alpha}\hat{\beta}\hat{g}/6$ rather than zero, we can track the position of this class of falling bodies almost as well as we would if we actually observed the wind resistance."

In fact, wind resistance is not negligible; rather, it can be compensated for. At the very least we ought to make clear the distinction between the two statements. For reasons I have explained, I think simplification more appropriately interpreted as the problem of neglecting variabl

roblem of compensating for their effects. There is first the nt that if simplification were intended to compensate for neglect certain secondary influences, we might expect se a more appropriate adjective than "negligible." Second, for a secondary influence, the theory may be fundamen- isically distorted. Consider the gravity example. If we stance we assert what is completely true: a body falling cuum accelerates at the constant rate $\hat{g}$. Contrast that with results when wind resistance is compensated for: a body at the time of departure instantaneously falls to a height initial position, attaining thereby absolutely no velocity, alls, accelerating at the constant rate $\hat{g}-\hat{\beta}\hat{g}\hat{r}/3$. The vtonian mechanics is obvious and absurd.

tests with unrecomputed coefficients can be calculated formulas as tests with recomputed coefficients provided he constraint matrices $\mathbf{R}$ and $\mathbf{r}$ appropriately. If we write $=\alpha+\mathbf{z}_T'\boldsymbol{\gamma}+\mathbf{w}_T'\boldsymbol{\delta}+u_T$, a simplification hypothesis is $\boldsymbol{\delta}=0$. recomputation of the coefficients on the $\mathbf{z}$ variables by constraint that the coefficients must equal their posterior us a causally constrained simplification is implied by the es

$$\mathbf{R}=\begin{bmatrix}\mathbf{0} & \mathbf{I} & \mathbf{0}\\ \mathbf{0} & \mathbf{0} & \mathbf{I}\end{bmatrix} \qquad \mathbf{r}=\begin{bmatrix}E(\boldsymbol{\gamma})\\ \mathbf{0}\end{bmatrix}$$

olumn of $\mathbf{R}$ is a vector of zeroes multiplying the constant α

estriction matrices the mean-square–error penalty (6.14)

$$\mathbf{R}E(\boldsymbol{\beta})-\mathbf{r})'(\mathbf{R}S^{-1}(\mathbf{x}_T)\mathbf{R}')^{-1}(\mathbf{R}E(\boldsymbol{\beta})-\mathbf{r})$$
$$E(\boldsymbol{\delta})]'V(\mathbf{w}_T)[E(\boldsymbol{\delta})]. \tag{6.19}$$

bles without the causal constraint requires constraint

$$\mathbf{R}=[\mathbf{0} \quad \mathbf{0} \quad \mathbf{I}],\ \mathbf{r}=[\mathbf{0}],$$

uare–error penalty (6.14) becomes

$$[E(\boldsymbol{\delta})]'V(\mathbf{w}_T|\mathbf{z}_T)[E(\boldsymbol{\delta})] \tag{6.20}$$

is the conditional variance of $\mathbf{w}_T$, given $\mathbf{z}_T$. Penalty (6.20) penalty (6.19) depending on the correlation between the cluded variables, because the included variables are used ded variables.

Given the assumption of diffuse priors, and supposing that δ is a scalar, the criterion (6.19) is just the square of the least-squares coefficient times the sample variance of the variable. If this were divided by the square of the sample variance of the dependent variable, the resulting number would be just the beta coefficient, which can be computed as least squares with variables standardized to have unit variance. Although standardized coefficients are used in other disciplines, in the econometrics literature they are rarely even mentioned. Goldberger (1964, pp. 197–198) is an exception.

To conclude, criterion (6.20), which is equivalent to (6.16) under diffuseness assumptions, ranks variables considered individually for discarding in the same way as traditional t tests. Criterion (6.19), however, provides a ranking identical to the ranking implied by classical beta coefficients. It seems to me, therefore, that the rarely used beta coefficients could be usefully resurrected as indicators of significance when models are being simplified, although the variance of the explanatory variables ought at a minimum be trend and autocorrelated adjusted.

6.3 Simplification for Control

A point that may be obvious is that simplification is problem specific, and, for example, simplification for prediction may be quite different from simplification for control. The one-period control problem of Lindley (1968) illustrates this fact. Suppose a scalar variable y_T is determined by the linear-regression process

$$y_T=\alpha+\boldsymbol{\gamma}'\mathbf{z}_T+\boldsymbol{\delta}'\mathbf{w}_T+u_T \tag{6.21}$$

where $\boldsymbol{\gamma}$ and $\boldsymbol{\delta}$ are vector parameters, α is a scalar parameter, u_T is a residual error with mean zero and variance σ^2, and $\mathbf{z}_T$ and $\mathbf{w}_T$ are vectors of explanatory variables. The control problem is to select $\mathbf{z}_T$ and $\mathbf{w}_T$ in such a way that y_T is likely to be close to some target t. In particular, let us choose the explanatory variables to minimize expected loss where loss is quadratic

$$L(y_T,t)=(y_T-t)^2.$$

Writing the regression process as

$$y_T=\alpha+\boldsymbol{\beta}'\mathbf{x}_T+u_T \tag{6.22}$$

where $\boldsymbol{\beta}'=[\boldsymbol{\gamma}',\boldsymbol{\delta}']$ and $\mathbf{x}_T=[\mathbf{z}_T',\mathbf{w}_{cT}']$, the expected loss can be written as a function of $\mathbf{x}_T$ as

$$E(L(y_T,t)|\mathbf{x}_T)=E\left(\left[\alpha+\boldsymbol{\beta}'\mathbf{x}_T+u_T-t\right)^2\right]|\mathbf{x}_T\Big).$$

Setting the derivatives of this expression to zero to obtain the minimizing

$$\mathbf{0}=E\left[2\boldsymbol{\beta}\boldsymbol{\beta}'\mathbf{x}_T+2\boldsymbol{\beta}(\alpha-t)\right]$$

$$\mathbf{x}_T=(E\boldsymbol{\beta}\boldsymbol{\beta}')^{-1}E\boldsymbol{\beta}(t-\alpha).$$

alue of $\mathbf{x}_T$ into the expected loss, we obtain the minimum

$$\min_T E\left[L(y_T,t)|\mathbf{x}_T\right]$$
$$+E(\alpha-t)^2-E(t-\alpha)\boldsymbol{\beta}'(E\boldsymbol{\beta}\boldsymbol{\beta}')^{-1}E\boldsymbol{\beta}(t-\alpha). \qquad (6.23)$$

for the expected loss simplifies nicely in the case when α and $\boldsymbol{\beta}$ derives only from observation of the regression y. Letting $\mathbf{Y}$ be the T-dimensional vector of previous ıe process and $\mathbf{X}$ be the matrix of observations of the bles, the posterior moments are

$$E(\alpha)=\bar{Y}-\bar{\mathbf{X}}'(\mathbf{X}'\mathbf{M}\mathbf{X})^{-1}\mathbf{X}'\mathbf{M}\mathbf{Y}=b_0$$
$$E(\boldsymbol{\beta})=(\mathbf{X}'\mathbf{M}\mathbf{X})^{-1}\mathbf{X}'\mathbf{M}\mathbf{Y}=\mathbf{b}$$

dimensional vector of ones and $\mathbf{M}=\mathbf{I}-\mathbf{1}(\mathbf{1}'\mathbf{1})^{-1}\mathbf{1}'$, $\bar{\mathbf{X}}=$ '. Also, the variance matrix can be written as

$$\frac{1}{T^2}\begin{bmatrix}\mathbf{1}'\mathbf{1} & \mathbf{1}'\mathbf{X}\\ \mathbf{X}'\mathbf{1} & \mathbf{X}'\mathbf{X}\end{bmatrix}^{-1}$$
$$\frac{1}{T^2}\begin{bmatrix}T^{-1}\left(1+\mathbf{1}'\mathbf{X}(\mathbf{X}'\mathbf{M}\mathbf{X})^{-1}\bar{\mathbf{X}}\right) & -\bar{\mathbf{X}}'(\mathbf{X}'\mathbf{M}\mathbf{X})^{-1}\\ -(\mathbf{X}'\mathbf{M}\mathbf{X})^{-1}\bar{\mathbf{X}} & (\mathbf{X}'\mathbf{M}\mathbf{X})^{-1}\end{bmatrix}.$$

/ $\bar{Y}=b_0+\bar{\mathbf{X}}'\mathbf{b}$ we may write the regression process as

$$=\alpha+\boldsymbol{\beta}'\mathbf{x}_T-b_0-\mathbf{b}'\bar{\mathbf{X}}+u_T$$
$$=(\alpha-b_0)+(\boldsymbol{\beta}-\mathbf{b})'\bar{\mathbf{X}}+\boldsymbol{\beta}'\left(\mathbf{x}_T-\bar{\mathbf{X}}\right)+u_T$$
$$\equiv\alpha^*+\boldsymbol{\beta}'\mathbf{x}_T^*+u_T \qquad (6.24)$$

r and if α and β were known to equal $E\alpha$ and $E\beta$, then the instrument $E\alpha)/E\beta$, which is called the certainty equivalence control rule. Assument, the optimal rule can be written in terms of the certainty equivalence $+E^2(\beta))^{-1}E(\beta)(t-E(\alpha))=(1+t_\beta^{-2})^{-1}x_T^c$, where $t_\beta^2=E^2(\beta)/V(\beta)$. : is more conservative than the certainty equivalence rule in the sense ›le is not turned on as far. The shrinkage factor $(1+t_\beta^{-2})^{-1}$ is a function β as measured by t^2

where

$$\alpha^*=(\alpha-b_0)+(\boldsymbol{\beta}-\mathbf{b})'\bar{\mathbf{X}}$$
$$\mathbf{x}_T^*=\mathbf{x}_T-\bar{\mathbf{X}}.$$

Controlling y_T at t is equivalent to controlling $y_T^*=y_T-\bar{Y}$ at $t^*=t-\bar{Y}$, where y_T^* is generated by the process described in (6.24). The expected loss (6.23) attains a simple form since $E(\boldsymbol{\beta}\alpha^*)=\mathbf{0}$

$$\begin{aligned}L_1&=E(\alpha^*-t^*)^2-t^{*2}\mathbf{b}'\left(\mathbf{b}\mathbf{b}'+\sigma^2(\mathbf{X}'\mathbf{M}\mathbf{X})^{-1}\right)^{-1}\mathbf{b}+\sigma^2\\&=E\alpha^{*2}+t^{*2}\left(1-\mathbf{b}'\left(\mathbf{b}\mathbf{b}'+\sigma^2(\mathbf{X}'\mathbf{M}\mathbf{X})^{-1}\right)^{-1}\mathbf{b}\right)+\sigma^2\\&=\frac{\sigma^2}{T}+\sigma^2+\frac{t^{*2}}{1+\mathbf{b}'\mathbf{X}'\mathbf{M}\mathbf{X}\mathbf{b}/\sigma^2}\\&=\sigma^2(1+T^{-1})+\frac{t^{*2}}{1+\chi^2}\end{aligned} \qquad (6.25)$$

where we have used the inverse formula $(\mathbf{x}\mathbf{x}'+\mathbf{A})^{-1}=\mathbf{A}^{-1}-\mathbf{A}^{-1}\mathbf{x}(1+\mathbf{x}'\mathbf{A}^{-1}\mathbf{x})^{-1}\mathbf{x}'\mathbf{A}^{-1}$.

Thus the minimum expected loss is a quadratic function of the deviation of the target from the historical level of the process, $(t-\bar{Y})^2=t^{*2}$. The coefficient multiplying this term is $(1+\chi^2)^{-1}$ where χ^2 is the value of the chi-square statistic for testing $\boldsymbol{\beta}=\mathbf{0}$. A large χ^2 statistic thus implies that y_T can be pushed from its historical mean without incurring great expected loss. The part of the expected loss independent of the target is just the variance of y_T assuming that $\mathbf{x}_T$ is set to its historical level $\bar{\mathbf{X}}$,

$$V\left(y_T|\mathbf{x}_T=\bar{\mathbf{X}}\right)=E\alpha^{*2}+\sigma^2=\sigma^2(1+T^{-1}).$$

Next consider the possibility that none of the variables is controlled. To compute expected control error it is then necessary to "guess" what the explanatory variables will be. This means modeling the process that generates the explanatory variables. For our purposes it is enough to know the first two moments of $\mathbf{x}_T$, since the expected loss can be written as

$$\begin{aligned}E(y_T-t)^2&=E(\alpha+\boldsymbol{\beta}'\mathbf{x}_T+u_T-t)^2\\&=\sigma^2+E(\alpha-t)^2+2E(\alpha-t)\boldsymbol{\beta}'\mathbf{x}_T+E\boldsymbol{\beta}'\mathbf{x}_T\mathbf{x}_T'\boldsymbol{\beta}.\end{aligned}$$

Taking as we did in the previous section the assumption of an independent multivariate process for the explanatory variables, we have approximately $E\mathbf{x}_T=\bar{\mathbf{X}}=\mathbf{X}'\mathbf{1}/T$, $V\mathbf{x}_T=\mathbf{X}'\mathbf{M}\mathbf{X}/T$, where $\mathbf{M}=\mathbf{I}-\mathbf{1}(\mathbf{1}'\mathbf{1})^{-1}\mathbf{1}'$. These together with the least squares moments for α and $\boldsymbol{\beta}$ imply in the absence

$$
\begin{aligned}
t)^2 &= E\left(y_T-\bar{Y}+\bar{Y}-t\right)^2 \\
&= E\left(y_T-\bar{Y}\right)^2+(\bar{Y}-t)^2 \\
&= \sigma^2+E(\alpha^*+\boldsymbol{\beta}'\mathbf{x}_T^*)^2+t^{*2} \\
&= \sigma^2+E\alpha^{*2}+E\,\mathrm{tr}(\mathbf{x}_T^*\mathbf{x}_T^{*\prime}\boldsymbol{\beta}\boldsymbol{\beta}')+t^{*2} \\
&= \sigma^2+\frac{\sigma^2}{T}+\frac{k\sigma^2}{T}+\frac{\chi^2\sigma^2}{T}+t^{*2} \\
&= \sigma^2\left(1+\frac{k+1+\chi^2}{T}\right)+t^{*2} \qquad (6.26)
\end{aligned}
$$

nensionality of $\boldsymbol{\beta}$ and χ^2 is the chi-square value for testing
$\mathbf{Xb}/\sigma^2$.
), the expected loss with no control, is to be contrasted
).25), the expected loss with optimal control. Their dif-

$$
\frac{\sigma^2(k+\chi^2)}{T}+\frac{t^{*2}\chi^2}{1+\chi^2}, \qquad (6.27)
$$

entive to use what is known about the determinants of y_T
cise. If it is desired to assure that y_T attains its historical
econd term drops out ($t^*=0$). The percentage increase in
ue to decontrolling $\mathbf{x}_T$ is then

$$
\frac{[k+\chi^2]/T}{1+T^{-1}}=\frac{k+\chi^2}{T+1}, \qquad (6.28)
$$

minimum of $k/(T+1)$ when $\chi^2=0$. We are thus led to
with $T+1$ to determine if decontrolling $\mathbf{x}_T$ could be
ase expected losses substantially.
r hand, it is desired to control y_T at some value far from
n, the second term in (6.27) dominates the expected loss.
increase in expected loss would then be just χ^2, and we
mpare χ^2 with the number one to decide if controlling $\mathbf{x}_T$

examined the extreme cases in which either all or none of
he vector $\mathbf{x}_T'=(\mathbf{z}_T',\mathbf{w}_T')$ is under control. The intermediate
control affects only $\mathbf{z}_T$ is more difficult, since it requires a
how $\mathbf{z}_T$ affects the distribution of $\mathbf{w}_T$ or, more accurately,
he conditional distribution $f(y_T|\mathbf{z}_T)$. Both the prediction
ion 6.1 and the control problem of this section are most

the conditional moments of y_T:

$$
\begin{aligned}
E(y_T|\mathbf{z}_T) &= Ey_T+\mathbf{g}'(\mathbf{z}_T-E\mathbf{z}_T) \\
V(y_T|\mathbf{z}_T) &= a+(\mathbf{z}_T-E\mathbf{z}_T)'\mathbf{A}(\mathbf{z}-E\mathbf{z}_T). \qquad (6.29)
\end{aligned}
$$

These assumptions—that the mean is a linear function and that the variance is a quadratic function of $\mathbf{z}_T$—are implicit in the foregoing discussion. The prediction problem of minimizing $E(y_T-\hat{y})^2$ where $\hat{y}$ is a function of $\mathbf{z}_T$ is straightforwardly solved by letting $\hat{y}=E(y_T|\mathbf{z}_T)$ with resultant expected loss $E(y_T-\hat{y})^2=E[y_T-E(y_T|\mathbf{z}_T)]^2=EV(y_T|\mathbf{z}_T)=a+\mathrm{tr}\,\mathbf{A}V(\mathbf{z}_T)$.

The control problem is equally trivial. We wish to choose $\mathbf{z}_T$ to minimize

$$
\begin{aligned}
\min_{\mathbf{z}_T} E\left[(y_T-t)^2|\mathbf{z}_T\right] &= \min_{\mathbf{z}_T} E\left(\left[y_T-E(y_T|\mathbf{z}_T)^2\right]|\mathbf{z}_T\right)+\left[t-E(y_T|\mathbf{z}_T)\right]^2 \\
&= \min_{\mathbf{z}_T} V(y_T|\mathbf{z}_T)+\left[t-E(y_T|\mathbf{z}_T)\right]^2.
\end{aligned}
$$

With the foregoing moments the derivatives of this expression with respect to $\mathbf{z}_T$ are

$$
2\mathbf{A}(\mathbf{z}_T-E\mathbf{z}_T)-2\mathbf{g}(t-Ey_T-\mathbf{g}'[\mathbf{z}_T-E\mathbf{z}_T])
$$

which when set to zero yields the optimizing value of $\mathbf{z}_T$

$$
\mathbf{z}_T^*=E\mathbf{z}_T+(\mathbf{A}+\mathbf{gg}')^{-1}\mathbf{g}(t-Ey_T).
$$

The resulting expected loss is

$$
\begin{aligned}
E\left[(y_T-t)^2|\mathbf{z}_T=\mathbf{z}_T^*\right] &= a+(t-Ey_T)\mathbf{g}'(\mathbf{A}+\mathbf{gg}')^{-1}\mathbf{A}(\mathbf{A}+\mathbf{gg}')^{-1}\mathbf{g}(t-Ey_T) \\
&\quad +\left[t-Ey_T-\mathbf{g}'(\mathbf{A}+\mathbf{gg}')^{-1}\mathbf{g}(t-Ey_T)\right]^2 \\
&= a+(t-Ey_T)^2\left[1-\mathbf{g}'(\mathbf{A}+\mathbf{gg}')^{-1}\mathbf{g}\right] \\
&= a+\frac{(t-Ey_T)^2}{1+\mathbf{g}'\mathbf{A}^{-1}\mathbf{g}} \qquad (6.30)
\end{aligned}
$$

Note that this is a quadratic function of $(t-Ey_T)$, the discrepancy between the target and the expected value of y_T.

To be specific, let us again work with the diffuse prior assumption. After some minor manipulation, we may obtain for the constants in the moments (6.29) the following

$$
\begin{aligned}
&E(y_T)=\bar{Y}, \qquad E(\mathbf{z}_T)=\bar{\mathbf{Z}} \\
&\mathbf{g}=(\mathbf{Z}'\mathbf{MZ})^{-1}(\mathbf{Z}'\mathbf{MY}) \qquad \text{(the regression of } \mathbf{Z} \text{ on } \mathbf{Y}) \\
&\mathbf{A}=\sigma^2(\mathbf{Z}'\mathbf{MZ})^{-1}
\end{aligned}
$$

$\mathbf{W}'\mathbf{MW} - \mathbf{W}'\mathbf{MZ})(\mathbf{Z}'\mathbf{MZ})^{-1}\mathbf{Z}'\mathbf{MW})\mathbf{b}_w/\sigma^2$. Notice that e chi-square statistic for testing $\boldsymbol{\gamma}=\mathbf{0}$, given that $\boldsymbol{\delta}=$ pected loss is then

$$\min_{z_T} E\left[(y_T - t)^2 | \mathbf{z}_T\right]$$

$$\sigma^2(1+T^{-1}) + \frac{\chi^2_\delta \sigma^2}{T} + \frac{k_w\sigma^2}{T} + \frac{t^{*2}}{1+\chi^2_{\gamma|\delta=0}}.$$

und the historical mean, $t^{*2}=0$, the percentage increase ss if $\mathbf{w}_T$ is not controlled is thus

$$\frac{T^{-1}(\chi^2_\delta + k_w)}{1+T^{-1}} = \frac{\chi^2_\delta + k_w}{1+T}$$

compare $\chi^2_\delta + k_w$ with $(1+T)$ to determine if $\mathbf{w}_T$ can be little increase in expected error.

· from the historical mean the percentage increase in ıe to decontrolling $\mathbf{w}_T$ is

$$\frac{\left(1+\chi^2_{\gamma|\delta=0}\right)^{-1} - \left(1+\chi^2_\beta\right)^{-1}}{\left(1+\chi^2_\beta\right)^{-1}}$$

$$= \frac{\chi^2_\beta - \chi^2_{\gamma|\delta=0}}{1+\chi^2_{\gamma|\delta=0}}$$

$$= \frac{\chi^2_\delta}{1+\chi^2_{\gamma|\delta=0}}$$

compare χ^2_δ with $1+\chi^2_{\gamma|\delta=0}$.

epeated that these results involve the unlikely assumption ıg $\mathbf{z}_T$ we do not alter the process that generates the ıbles (in the sense that the conditional distribution ·ved). The assumption of known σ^2 can be altered by erior mean where relevant. Mathematically more ap- ay treat the vector $(\mathbf{y}_T, \mathbf{x}_T')$ as coming from a multivariate ɔn with unknown mean and unknown variance matrix. A 'or the uncertain parameters implies that the marginal $_T, \mathbf{x}_T')$ is a multivariate Student distribution with means isfying (6.28) and (6.29). We leave to the tenacious reader t calculation.

6.4 Conclusion

To conclude we may restate, first, the more important formal results of this chapter and then reiterate the more important informal lessons to be learned.

The results of this chapter listed in Table 6.2 make use of the assumptions listed in Table 6.1. If a variable y is generated by a linear regression process with explanatory variables $\mathbf{w}$ and $\mathbf{z}$, if $\mathbf{w}$ and $\mathbf{z}$ themselves come from a multivariate normal process, and if priors for the various parameters are appropriately diffuse, then: (1) for a conditional prediction problem, we need not observe w, if the χ^2 statistic for testing whether $\mathbf{w}$ can be omitted (χ^2_δ) is small relative to $(T+k)$ where T is the number of observations and k is the dimension of $\mathbf{x}'=(\mathbf{w}',\mathbf{z}')$; (2) for control with a target equal to the historical mean of y, $\mathbf{w}$ may be decontrolled if $\chi^2_\delta + k_w$ is small relative to $(1+T)$, where k_w is the dimension of $\mathbf{w}$; (3) for control far from the historical mean, $\mathbf{w}$ may be decontrolled if χ^2_δ is small relative to $1+\chi^2{}_{\gamma|\delta=0}$, one plus the χ^2 value for testing if $\mathbf{z}$ belongs in the equation given that $\mathbf{w}$ does not.

The principal caveat that has been repeated ad nauseam is that these results involve a very specific and often unwarranted assumption about the

Table 6.1

Assumptions for Simplification Analysis

Model	
	$y_t = \alpha + \mathbf{z}_t'\boldsymbol{\gamma} + \mathbf{w}_t'\boldsymbol{\delta} + u_t$
	$\equiv \alpha + \mathbf{x}_t'\boldsymbol{\beta} + u_t, \quad t=0,1,\ldots,T$
	$u_t \sim N(0,\sigma^2)$, σ^2 known
	$\mathbf{x}_t \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$
	$\boldsymbol{\mu}, \boldsymbol{\Sigma}, \alpha, \boldsymbol{\beta}$ have diffuse priors
Observations	
	$\mathbf{Y}(T\times 1), \mathbf{Z}(T\times k_z), \mathbf{W}(T\times k_w)$,
	$\mathbf{X}=(\mathbf{Z},\mathbf{W})(T\times k_x)$
Statistics	
	$\mathbf{b}=(\mathbf{X}'\mathbf{MX})^{-1}\mathbf{X}'\mathbf{MY}, \quad \mathbf{M}=\mathbf{I}-\mathbf{1}T^{-1}\mathbf{1}'$
	$\bar{Y}=\mathbf{1}'\mathbf{Y}/T, \bar{\mathbf{X}}=\mathbf{1}'\mathbf{X}/T$
	$b_0=\bar{\mathbf{Y}}-\bar{\mathbf{X}}\mathbf{b}$
	$\mathbf{g}=(\mathbf{Z}'\mathbf{MZ})^{-1}\mathbf{Z}'\mathbf{MY}, \quad \tilde{b}_0=\bar{Y}-\bar{Z}'\mathbf{g}$
	$\chi^2_\delta = \mathbf{b}_w'[\mathbf{W}'\mathbf{MW}-\mathbf{W}'\mathbf{MZ}(\mathbf{Z}'\mathbf{MZ})^{-1}\mathbf{Z}'\mathbf{MW}]\mathbf{b}_w/\sigma^2$
	$\chi^2_\beta = \mathbf{b}'\mathbf{X}'\mathbf{MXb}/\sigma^2$
	$\chi^2_{\gamma|\delta} = \mathbf{g}'\mathbf{Z}'\mathbf{MZg}/\sigma^2$

process that generates the explanatory variables. No simplification decisions can be made without either an implicit or explicit study of the behavior of the explanatory variables, and we hardly need say that it seems clear that an explicit study of their behavior is highly desirable.

For both prediction and control problems the effects of the excluded variables have been compensated for by adjustment of the included variables, and we have argued at length that it may be desirable not to adjust in this way. Semantically, adjustment is undesirable, because rather than asking if a variable can be neglected, in fact, we ask if it can be compensated for. Metaphysically, adjustment is undesirable, since it implies a causal link between the included and excluded variables. Statistically, the predictions and control that result may be quite inferior if anything happens to change the historical correlations between the variables. Control, especially, is likely to alter those correlations.

Simplification Analysis

	Decision Rules		Expected Losses	
	D_1: unconstrained	D_0: constrained	L_1	$L_0 - L_1$
Prediction	$\hat{y}_T = b_0 + \mathbf{z}_T'\mathbf{b}_z + \mathbf{w}_T'\mathbf{b}_w$	$\hat{y}_T = \tilde{b}_0 + \mathbf{z}_T'\bar{\boldsymbol{\beta}}$	$\sigma^2(1 + T^{-1}k_x)$	$T^{-1}\sigma^2\chi_\beta^2$
Causally Constrained Prediction	same as above	$\hat{y}_T = b_0 + \bar{\mathbf{W}}'\mathbf{b}_w + \mathbf{z}_T'\mathbf{b}_z$	same as above	$T^{-1}\mathbf{b}_w'\mathbf{W}'\mathbf{M}\mathbf{W}\mathbf{b}_w$
Control $t^* = t - \ddagger\bar{y}$	$\mathbf{x}_T = \bar{\mathbf{X}} - (\mathbf{b}\mathbf{b}' + \sigma_z^2[\mathbf{X}'\mathbf{M}\mathbf{X}]^{-1})^{-1}\mathbf{b}t^*$	$\mathbf{z}_T = \bar{\mathbf{Z}} + (\boldsymbol{\beta}\boldsymbol{\beta}' + \sigma_z^2[\mathbf{Z}'\mathbf{M}\mathbf{Z}]^{-1})^{-1}\boldsymbol{\beta}t^*$	$\sigma^2(1 + T^{-1}) + t^{*2}(1 + \chi_\beta^2)^{-1}$	$\sigma^2(\chi_\delta^2 + k_w)T^{-1} + t^{*2}([1 + \chi_{\gamma/\delta}^2]^{-1} - [1 + \chi_\beta^2]^{-1})$.
Control $t = \bar{Y}$	$\mathbf{x}_T = \bar{\mathbf{X}}$	$\mathbf{z}_T = \bar{\mathbf{Z}}$	$\sigma^2(1 + T^{-1})$	$\sigma^2(\chi_\delta^2 T^{-1} + k_w)$
Control (approx.) t^* very large	same as two lines above	same as two lines above	$t^{*2}(1 + \chi_\beta^2)^{-1}$	$t^{*2}([1 + \chi_{\gamma/\delta}^2]^{-1} - [(1 + \chi_\beta^2]^{-1})$