provided the variances are positive. Note that β is an instrumental variables estimator, with $\mathbf{z}$ used as an instrument for $\mathbf{x}$ in the regression of $\mathbf{Y} - \gamma\mathbf{z}$ on $\mathbf{x}$.

An appealing proposition is that if $\mathbf{x}$ is a "good" measurement of χ, then when we regress $\mathbf{Y}$ on $\mathbf{x}$ and $\mathbf{z}$, the coefficient on $\mathbf{z}$ should be close to the known value of γ. If it is not, then it seems unlikely that we could obtain much information about β. Intuition thus suggests that the uncertainty in β may be related to the difference between γ and $\hat{\gamma}$. This, like a long list of other interesting questions, will remain unanswered until some appropriate approximations of the posterior dispersion are available.

CHAPTER 8

DATA-SELECTION SEARCHES

Theoretical models are often vague or have nothing at all to say about the choice of particular observations. Even less frequently do they suggest the circumstances in which two or more observations can be regarded as independent pieces of relevant information. It is thus necessary for the empirical worker both to select a subset of potential observations and to determine the extent to which observations are correlated. To put it another way, the researcher must identify observations or transformations of observations that can be considered to be independent replications of an unchanging "experiment." In practice, this may mean estimating coefficients with different subsets or different transformations of the data set and selecting the result that appears best according to some criteria. We call this a data-selection search.

The fact that this process is data dependent obviously has consequences for the interpretation of the final result of a data-selection search. It seems clear that when the data evidence is partly spent to pick a data set, the regression equation that is finally selected to convey the data evidence at least overstates the precision of the evidence and likely distorts it as well. The function of this chapter is thus to describe the inferences that are appropriate when some of the data are discarded or when the data are transformed by a data-dependent function.

In the case of interpretive searches, a constrained regression can at best approximate the location of a

posterior distribution, which is, in fact, a mixture of many constrained regressions. Similarly, a Bayesian will necessarily discard none of the observations but will instead place relatively low weight on some. The extent to which one can approximate his posterior distribution by discarding altogether some observations and assigning equal weight to the remaining ones is not extensively discussed. Nor is the interplay between prior information and data selection extensively discussed. Instead, this chapter reports the likelihood functions implied by statistical models that generate outliers or interdependencies.

Although a data-selection search involves features of interpretive searches, the two may be distinguished in the following way. Let a linear model be written as $Y_t = \alpha_t + \boldsymbol{\beta}_t' \mathbf{x}_t$ where $\mathbf{x}_t$ is a vector of observable explanatory variables, Y_t is the observable dependent variable, and $(\alpha_t, \boldsymbol{\beta}_t)$ is a vector of unobservable parameters applying to the tth observation. Assume that the unobservables $(\alpha_t, \boldsymbol{\beta}_t)$, $t = 1, \ldots, T$ have a common mean $(\alpha, \boldsymbol{\beta})$. It is the business of a data-selection search to pick the multivariate distribution of the unobservables around their common mean $(\alpha, \boldsymbol{\beta})$. An interpretive search, on the other hand, is designed to make use of prior information about the means $(\alpha, \boldsymbol{\beta})$.

The usual least-squares logic results from the assumptions that $\boldsymbol{\beta}_t$ does not vary from observation to observation and that α_t is the sum of $\alpha + \varepsilon_t$, an independent normal random variable with constant variance. A first step toward relaxing this assumption is to allow for dependence among the ε_t random variables. This is discussed in Section 8.1 under the heading of "nonspherical disturbances." Alternatively, the assumption of normality may be relaxed. The consequences of fat-tailed nonnormal distributions are described in Section 8.2. The next two sections explore models that let the slope parameters $\boldsymbol{\beta}_t$ vary from observation to observation, but maintain the assumption of normality.

No attempt is made to relax all assumptions simultaneously, and this chapter does not suggest a mechanical approach to the data-selection problem. In that regard, it is like the chapter on interpretive searches, which takes the data distribution as given and analyzes the mapping of prior distributions into posterior distributions. The point is made in that chapter that if the prior could be uniquely determined, there would be a unique interpretation of the data, but ambiguity in the choice of prior implies ambiguity in the posterior distribution. In the case of data-selection searches, if the data distribution could be taken as given, the data would imply a unique likelihood function. But just as it is impossible unambiguously to select a prior, so too is it impossible unambiguously to select a data distribution. Not only must the interpretation of the data evidence thus remain elusive, but also the data evidence itself must be defined imprecisely. A researcher can only report features of the mapping of prior

8.1 Nonspherical Disturbances

The conditional distribution of the error vector has heretofore been assumed to be multivariate normal with mean vector zero and covariance matrix $\sigma^2\mathbf{I}$. The further assumption of a gamma distribution for σ^{-2} implies that the error vector is distributed (marginally) multivariate Student with a covariance matrix also proportional to the identity matrix. We now wish to consider the consequences of using a covariance matrix that is not proportional to the identity matrix. We refer to these errors as nonspherical disturbances, thereby implying that the isodensities of the random errors are not spheres, as they are when the assumptions above apply.

Some consideration of nonspherical errors seems absolutely essential with nonexperimental data. Careful elicitation of one's personal opinions about the error process is quite unlikely to lead to zero covariances between the errors, and inferences from observable data may be erroneous if there is no adjustment made for departures from sphericity. To give an example, a time-series data set may be enlarged by a factor of 12 merely by using monthly data instead of annual data. But because of the dependence in the monthly residuals, the number 12 greatly overstates the real gain in information. This has been accurately called "counting your wealth in small change."

The nonspherical regression process may be written as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u} \tag{8.1}$$

where $\mathbf{Y}$ is $T \times 1$, $\mathbf{X}$ is $T \times k$, $\boldsymbol{\beta}$ is $k \times 1$ and $\mathbf{u}$ is a $T \times 1$ normal error vector with mean vector zero and covariance matrix $\sigma^2\boldsymbol{\Sigma}$ and where $\boldsymbol{\Sigma}$ is a known matrix. The likelihood function may then be written as

$$f(\mathbf{Y}|\boldsymbol{\beta}, \sigma^2, \boldsymbol{\Sigma}) \propto |\sigma^2\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp \frac{-\frac{1}{2}(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})'\boldsymbol{\Sigma}^{-1}(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})}{\sigma^2}.$$

The exponent in this function may be decomposed as

$$(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})'\boldsymbol{\Sigma}^{-1}(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}) = (\mathbf{Y}-\mathbf{X}\mathbf{b})'\boldsymbol{\Sigma}^{-1}(\mathbf{Y}-\mathbf{X}\mathbf{b}) + (\boldsymbol{\beta}-\mathbf{b})'\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X}(\boldsymbol{\beta}-\mathbf{b})$$

where $\mathbf{b}$ is a solution to

$$\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X}\mathbf{b} = \mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{Y}.$$

When $\mathbf{b}$ is unique, that is, when $\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X}$ is invertible, $\mathbf{b}$ is called the generalized least-squares estimate of $\boldsymbol{\beta}$, since it is the value of $\boldsymbol{\beta}$ that minimizes $(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})'\boldsymbol{\Sigma}^{-1}(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})$.

By an inspection of these formulas, the consequences of $\boldsymbol{\Sigma} \neq \mathbf{I}$ are first, to alter the location of the likelihood ellipsoid from $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ to $(\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{Y}$, second, to alter the shape of the ellipsoids $(\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X}$

the error sum of squares is computed differently).[1] We know from the liscussion of sensitivity analysis in Chapter 5 that each of these may have ı significant effect on the posterior distribution. For example, although east squares and generalized least squares may be exactly the same, differences in the precision matrices $\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X}$ and $\mathbf{X}'\mathbf{X}$ may mean that the two posterior distributions, corresponding to these two error processes, are quite different.

A more common situation arises when $\boldsymbol{\Sigma}$ is not known with certainty. It is then necessary to select a personal prior distribution for $\boldsymbol{\Sigma}$, which may be done by first writing $\boldsymbol{\Sigma}$ as a deterministic function of some vector of parameters $\boldsymbol{\theta}$, and then by assigning to $\boldsymbol{\theta}$ some hopefully convenient prior. As far as I know, there is no convenient way to analyze such models. The posterior distribution of $\boldsymbol{\beta}$ may be written as

$$f(\boldsymbol{\beta}|\mathbf{Y},\mathbf{X}) \propto \int_{\boldsymbol{\theta}} f(\boldsymbol{\beta}|\mathbf{Y},\mathbf{X},\boldsymbol{\theta}) f(\boldsymbol{\theta}|\mathbf{Y},\mathbf{X}) d\boldsymbol{\theta} \tag{8.2}$$

where $f(\boldsymbol{\beta}|\mathbf{Y},\mathbf{X},\boldsymbol{\theta})$ is a tractable distribution of $\boldsymbol{\beta}$, given some covariance matrix $\boldsymbol{\Sigma}(\boldsymbol{\theta})$, and where $f(\boldsymbol{\theta}|\mathbf{Y},\mathbf{X})$ is proportional to the product of a (tractable) marginal likelihood function $f(\mathbf{Y}|\boldsymbol{\theta},\mathbf{X})$ and a prior for $\boldsymbol{\theta}$. Unfortunately, numerical methods are required to evaluate the integral (8.2)

The most commonly analyzed form of nonspherical disturbances is first-order autocorrelation with

$$\boldsymbol{\Sigma} = \sigma^2 \begin{bmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{T-1} \\ \rho & 1 & \rho & \cdots & \rho^{T-2} \\ \rho^2 & \rho & 1 & \cdots & \rho^{T-3} \\ \vdots & \vdots & \vdots & & \vdots \\ \rho^{T-1} & \rho^{T-2} & \rho^{T-3} & \cdots & 1 \end{bmatrix} \tag{8.3}$$

and

$$\boldsymbol{\Sigma}^{-1} = \sigma^{-2}(1-\rho^2)^{-1} \begin{bmatrix} 1 & -\rho & 0 & \cdots & 0 & 0 \\ -\rho & 1+\rho^2 & -\rho & \cdots & 0 & 0 \\ 0 & -\rho & 1+\rho^2 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1+\rho^2 & -\rho \\ 0 & 0 & 0 & \cdots & -\rho & 1 \end{bmatrix}.$$

ellipsoid a particular likelihood value.

The marginal likelihood $f(\mathbf{Y}|\rho)$ can be computed without difficulty, since $\boldsymbol{\Sigma}$ can be diagonalized by the transformation $\boldsymbol{\Sigma}^{-1} = \mathbf{C}'\mathbf{C}\sigma^{-2}$ where

$$\mathbf{C} = (1-\rho^2)^{-\frac{1}{2}} \begin{bmatrix} (1-\rho^2)^{\frac{1}{2}} & 0 & 0 & \cdots & 0 & 0 \\ -\rho & 1 & 0 & \cdots & 0 & 0 \\ 0 & -\rho & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & 0 \\ 0 & 0 & 0 & \cdots & -\rho & 1 \end{bmatrix}.$$

Thus the vector $\mathbf{Y}^* = \mathbf{CY}$ is distributed normally with mean $\mathbf{CX}\boldsymbol{\beta}$ and variance $\mathbf{C}\boldsymbol{\Sigma}\mathbf{C}' = [\mathbf{C}'^{-1}\boldsymbol{\Sigma}^{-1}\mathbf{C}^{-1}]^{-1} = \sigma^2[\mathbf{C}'^{-1}\mathbf{C}'\mathbf{C}\mathbf{C}^{-1}] = \sigma^2\mathbf{I}$. Then, for example, if the parameters $(\boldsymbol{\beta},\sigma^2)$ are assigned a conjugate prior, the marginal distribution of $\mathbf{Y}^* = \mathbf{CY}$ is given by Equation (4.13) in Chapter 4, since all the assumptions leading to that equation are satisfied.

If, in particular, we take the diffuse prior assumption with ν_1 and $\mathbf{N}^*$ in that equation set to zero, we obtain the marginal density of $\mathbf{Y}^*$ as

$$f(\mathbf{Y}^*|\rho) \propto |\mathbf{X}'\mathbf{C}'\mathbf{C}\mathbf{X}|^{-1/2}(ESS(\rho))^{-T/2}$$

where

$$ESS(\rho) = [\mathbf{CY} - \mathbf{CXb}(\rho)]'[\mathbf{CY} - \mathbf{CXb}(\rho)].$$

Of course, $\mathbf{Y}^*$ is not observed, and it is necessary to transform this density into a density of $\mathbf{Y} = \mathbf{C}^{-1}\mathbf{Y}^*$, a transformation with Jacobian $|\mathbf{C}| = (1-\rho^2)^{(1-T)/2}$:

$$f(\mathbf{Y}|\rho) \propto (1-\rho^2)^{(1-T)/2}|\mathbf{X}'\mathbf{C}'\mathbf{C}\mathbf{X}|^{-1/2}(ESS(\rho))^{-T/2}.$$

Under the same assumptions,[2] $\boldsymbol{\beta}$ given ρ has a Student distribution with parameters

$$E(\boldsymbol{\beta}|\rho) = \mathbf{b}(\rho) = (\mathbf{X}'\mathbf{C}'\mathbf{C}\mathbf{X})^{-1}\mathbf{X}'\mathbf{C}'\mathbf{CY}$$
$$\mathbf{H}^{**}(\rho) = \mathbf{X}'\mathbf{C}'\mathbf{C}\mathbf{X}/s^2$$
$$\nu = T$$
$$\nu s^2 = ESS(\rho).$$

As a result, the marginal posterior distribution of $\boldsymbol{\beta}$ is a mixture of Student distributions.[3]

[2]This follows directly from the material in Chapter 4. Incidentally, there are certain technical differences between this treatment of autocorrelation and Zellner's (1971, pp. 86–98), having to do with the distribution of the first observation and also the choice of diffuse prior.

[3]See Zellner (1971, pp. 86–98) for an explicit expression for the result of an essentially similar integral.

As an interesting alternative way of looking at these distributions, observe that $\boldsymbol{\Sigma}^{-1}$ can be decomposed as

$$\sigma^2(1-\rho^2)\boldsymbol{\Sigma}^{-1}=(1-2\rho+\rho^2)\mathbf{I}$$
$$+\rho\begin{bmatrix} 1 & -1 & 0 & \cdots & 0 & 0 \\ -1 & 2 & -1 & \cdots & 0 & 0 \\ 0 & -1 & 2 & \cdots & 0 & 0 \\ \vdots & & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 2 & -1 \\ 0 & 0 & 0 & \cdots & -1 & 1 \end{bmatrix}$$
$$+(\rho-\rho^2)\begin{bmatrix} 1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 0 & \cdots & 0 & 0 \\ \vdots & & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & & 0 & 0 \\ 0 & 0 & 0 & \cdots & 0 & 1 \end{bmatrix}$$
$$\equiv(1-2\rho+\rho^2)\mathbf{I}+\rho\mathbf{A}+\rho(1-\rho)\mathbf{B}.$$

In the event that prior information is relatively diffuse, $\mathbf{N}^*=\mathbf{0}$, the (generalized least-squares) posterior mean of $\boldsymbol{\beta}$ is

$$E(\boldsymbol{\beta}|\mathbf{Y},\rho)=(\theta_1\mathbf{X}'\mathbf{X}+\theta_2\mathbf{X}'\mathbf{A}\mathbf{X}+\theta_3\mathbf{X}'\mathbf{B}\mathbf{X})^{-1}(\theta_1\mathbf{X}'\mathbf{Y}+\theta_2\mathbf{X}'\mathbf{A}\mathbf{Y}+\theta_3\mathbf{X}'\mathbf{B}\mathbf{Y}),$$

where

$$\theta_1=(1-2\rho+\rho^2)$$
$$\theta_2=\rho$$
$$\theta_3=\rho(1-\rho).$$

Finally, since $\mathbf{B}$ is almost a zero matrix, we may write the conditional posterior mean approximately as

$$E(\boldsymbol{\beta}|\mathbf{Y},\rho)\doteq(\theta_1\mathbf{X}'\mathbf{X}+\theta_2\mathbf{X}'\mathbf{A}\mathbf{X})^{-1}(\theta_1\mathbf{X}'\mathbf{X}\mathbf{b}+\theta_2\mathbf{X}'\mathbf{A}\mathbf{X}\mathbf{b}^{\Delta})$$

where $\mathbf{b}$ is the usual least-squares estimate and $\mathbf{b}^{\Delta}$ is the "first difference" estimate, $\mathbf{b}^{\Delta}=(\mathbf{X}'\mathbf{A}\mathbf{X})^{-1}\mathbf{X}'\mathbf{A}\mathbf{Y}$ computed by least squares with data in first difference form

$$Y_t-Y_{t-1}=\sum_j \beta_j(x_{jt}-x_{jt-1}).$$

In words, the generalized least-squares estimate of $\boldsymbol{\beta}$ is approximately a matrix-weighted average of ordinary least squares and least squares with data in first differences. It is worthwhile recalling here the discussion of [illegible] in particular the fact that the

The first-order autocorrelation matrix (8.3) is one very special kind of interdependence. For other models the reader is referred to Granger and Newbold (1976) or to Box and Jenkins (1970).

8.2 Outliers and Nonnormal Errors

Researchers who share with computers the task of examining data points almost always want to discard extreme observations. In some cases the resultant apparent inferences are greatly affected by the choice of observations. Were the assumptions implicit in the least-squares estimate actually accepted, there would be no logic to such a procedure. But in the rejection of outliers, a researcher is implicitly rejecting the assumption of normality, in particular, he is opting for a distribution with fatter tails than a normal distribution. Just as it is undesirable to choose a prior or a model after having seen the data, it is also undesirable to choose an implicit data density in this way. First of all, the inferences that are thus gathered are not fully legitimate. Second, the rejection of outliers tends to ignore the fact that it may be quite important to know the probability laws according to which the outliers are generated. In this section we consider formal models that involve "outlier rejection." It must be said at the outset that these models are relatively intractable and that cheaper, "data analytic" methods of discarding outliers may be preferred.[4] In fact, the discussion in this section may be nothing more than an apology for reasonable procedures.

Consider first the following model, due to Box and Tiao (1968). Let the values $Y_1, Y_2, \ldots, Y_T$ be a random sample from a normal population with mean μ and variance σ^2. Assume that $Z_t=Y_t$ is usually observed but that occasionally an outlier occurs and instead $Z_{t'}=Y_{t'}+e_{t'}$ is observed, where $e_{t'}$ is distributed normally with mean 0 and variance $\sigma^2+\phi^2$. If σ^2 and ϕ^2 as well as the occurrence of the outliers are known, then the mode of the likelihood function of μ is located at the weighted mean of the observations with weight σ^{-2} on the regular observations and weight $(\sigma^2+\phi^2)^{-1}$ on the outliers. If I is the set of n regular observations and $\bar{I}$ is the set of $T-n$ outliers, the likelihood function is

$$f(Y_1, Y_2, \ldots, Y_T|\mu,\sigma^2,\phi^2)\propto(\sigma^2)^{-n/2}\exp\left[-\frac{1}{2\sigma^2}\sum_I(Y_t-\mu)^2\right]$$
$$(\sigma^2+\phi^2)^{-(T-n)/2}\exp\left[-\frac{1}{2(\sigma^2+\phi^2)}\sum_{\bar{I}}(Y_t-\mu)^2\right].$$

[4]By the term "data analytic" methods I refer to procedures such as discarding observations more than three standard deviations from the mean. These procedures do not refer explicitly

If the set $\bar{I}$ of outliers were known in advance, the analysis is relatively straightforward. The more interesting case with I uncertain involves great complications, since there are 2^T different subsets $\bar{I}$. As a result, there are 2^T different "models," each corresponding to a different allocation of outliers, and the posterior distribution is a mixture, $\Sigma_i w_i f_i(\mu)$, of 2^T distributions.

Note that the implicit data density in this example is a mixture of normals

$$f(Z_t|\mu,\sigma^2,\phi^2)=\pi f_N(Z_t|\mu,\sigma^2)+(1-\pi)f_N(Z_t|\mu,\sigma^2+\phi^2)$$

where $1-\pi$ is the probability of an outlier and $f_N(Z|\mu,\sigma^2)$ indicates a normal density with mean μ and variance σ^2. Another fat-tailed distribution that is a continuous mixture of normals is the Student function

$$f_S(Z|\mu,s^2,\nu)=\int f_N(Z|\mu,\sigma^2)f_\gamma(\sigma^{-2}|s^2,\nu)d\sigma^{-2} \tag{8.4}$$

An analysis of Student sampling has been done by Blattberg and Gonedes (1975).

Another class of nonnormal distributions has been analyzed extensively by Box and Tiao (1973, Chapter 3). They explore the class of exponential power distributions

$$f(Y|\mu,\phi,\beta)=k\phi^{-1}\exp\left(-\tfrac{1}{2}\left|\frac{Y-\mu}{\phi}\right|^{2/(1+\beta)}\right),\quad -\infty<Y<\infty, \tag{8.5}$$

where k is a normalizing constant depending on β and the parameters satisfy $\phi>0$, $-\infty<\mu<\infty$, $-1<\beta\leq 1$, with $\beta=0$ corresponding to the normal distribution.

8.3 Pooling Disparate Evidence

The inferential models so far discussed have been constructed to answer questions of the form: given that a coin lands heads up, what conclusions may be drawn about the probability of getting a head if the same coin is flipped again? We now turn to questions of the form: given that coin A lands heads up, what conclusions may be drawn about the probability of getting a head if coin B is flipped? No objective distinction should be made between these two inferential problems. *Any* inductive inference depends on the *subjective* link between observed and unobserved events. Although there may be general agreement that two flips of the same coin are more closely linked than two flips of two different coins, there can be no incontrovertible argument to that effect. Thus in this section we consider the kind of joint priors for p_A and p_B that would induce us to use the observed behavior of coin A to draw inferences about p_B, the probabil-

consider nontrivial joint priors for $(p_1,p_2,\ldots)$ where p_i is the probability of getting a head on the ith flip of some particular coin.

The regression model is written as

$$Y_{it}=\boldsymbol{\beta}_{it}'\mathbf{x}_{it}+\varepsilon_{it} \qquad \begin{matrix} i=1,\ldots,N \\ t=1,\ldots,T \end{matrix} \tag{8.6}$$

where $\boldsymbol{\beta}_{it}$ is a $(k\times 1)$ vector of coefficients and $\mathbf{x}_{it}$ is a $(k\times 1)$ vector of observable explanatory variables possibly including a constant. As an example of such a model, Y_{it} may be purchases of oranges by individual i in time period t, and x_{it} may be the income of individual i in period t. The errors ε_{it} and the parameters $\boldsymbol{\beta}_{it}$ are doubly subscripted to reflect the fact that the demand for oranges varies across individuals and across time (as social tastes for oranges change or as the individual grows older, for example).

The question of interest is whether observation of $(Y_{j\tau},\mathbf{x}_{j\tau}')$ yields information about $(\varepsilon_{it},\boldsymbol{\beta}_{it})$ for $t\neq\tau$. Any solution to this problem depends on the prior distribution for the $(k+1)NT$ unobservables. One extreme would have the vector $(\varepsilon_{it},\boldsymbol{\beta}_{it}')$ independent of $(\varepsilon_{j\tau},\boldsymbol{\beta}_{j\tau}')$ for all $(i,t)\neq(j,\tau)$, and it is obvious, in that case, that there is no information in the observation of the jth process, $Y_{j\tau}$, about the unobservables of the ith process $(\varepsilon_{it},\boldsymbol{\beta}_{it}')$. Other assumptions imply pooling of evidence in one way or another. Let us for now postpone the issues raised by the variability of parameters over time and impose the conditions

$$\boldsymbol{\beta}_{it}=\boldsymbol{\beta}_i \qquad \text{for all } t.$$

This implies a set of N equations

$$Y_{it}=\boldsymbol{\beta}_i'\mathbf{x}_{it}+\varepsilon_{it}, \qquad i=1,\ldots,N \tag{8.7}$$

each of which describes the generation of T observations. Vectors of observations and errors of the ith process are denoted by

$$\mathbf{Y}_i=\begin{bmatrix} Y_{i1} \\ Y_{i2} \\ \cdot \\ \cdot \\ \cdot \\ Y_{iT} \end{bmatrix}, \qquad \mathbf{X}_i=\begin{bmatrix} \mathbf{x}_{i1}' \\ \mathbf{x}_{i2}' \\ \cdot \\ \cdot \\ \cdot \\ \mathbf{x}_{iT}' \end{bmatrix}$$

and the whole set of observations by

$$\mathbf{Y}=\begin{bmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \\ \cdot \\ \cdot \\ \cdot \\ \mathbf{Y}_N \end{bmatrix}, \quad \mathbf{X}=\begin{bmatrix} \mathbf{X}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{X}_2 & \cdots & \mathbf{0} \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{X}_N \end{bmatrix}, \quad \boldsymbol{\beta}=\begin{bmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \\ \cdot \\ \cdot \\ \cdot \\ \boldsymbol{\beta}_N \end{bmatrix}, \quad \boldsymbol{\varepsilon}=\begin{bmatrix} \boldsymbol{\varepsilon}_1 \\ \boldsymbol{\varepsilon}_2 \\ \cdot \\ \cdot \\ \cdot \\ \boldsymbol{\varepsilon}_N \end{bmatrix}$$

This allows us to write the set of regression equations compactly as

$$\mathbf{Y}=\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon}. \tag{8.8}$$

The generalized least-squares estimate of $\boldsymbol{\beta}$ is thus

$$\mathbf{b}=(\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{Y}$$

where $\boldsymbol{\Sigma}$ is the $NT\times NT$ variance matrix of $\boldsymbol{\varepsilon}$. Furthermore, assuming a normal prior for $\boldsymbol{\beta}$ with mean $\mathbf{b}^*$ and variance $(\mathbf{H}^*)^{-1}$, the posterior moments of $\boldsymbol{\beta}$ are

$$E(\boldsymbol{\beta}|\mathbf{Y},\mathbf{X})=(\mathbf{H}^*+\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}(\mathbf{H}^*\mathbf{b}^*+\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X}\mathbf{b}) \tag{8.9}$$

$$V(\boldsymbol{\beta}|\mathbf{Y},\mathbf{X})=(\mathbf{H}^*+\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}. \tag{8.10}$$

Several special cases of these formulas are now discussed.

MULTIVARIATE REGRESSIONS: UNCORRELATED COEFFICIENTS

The first model of interest allows the errors ε_{it} to be correlated across equations but constrains to zero the prior covariance between coefficients of different equations. The initial reaction that this is unlikely to lead to pooling of evidence has led Zellner (1971) to call the multivariate regression model a system of "seemingly unrelated regressions."

Using the notation of equation (8.8) let the prior moments of $\boldsymbol{\beta}$ be

$$E(\boldsymbol{\beta})=\mathbf{b}^*,\ V(\boldsymbol{\beta})=\mathbf{V}^*=\mathbf{H}^{*-1}=\begin{bmatrix}\mathbf{H}_1^{*-1} & \mathbf{0} & \cdots & \mathbf{0}\\ \mathbf{0} & \mathbf{H}_2^{*-1} & \cdots & \mathbf{0}\\ \vdots & \vdots & & \vdots\\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{H}_N^{*-1}\end{bmatrix} \tag{8.11}$$

thereby indicating that there is no a priori correlation between the coefficients in the various processes. The processes are related in the sense that the residual terms are correlated:

$$E(\varepsilon_{it}\varepsilon_{j\tau})=\begin{cases}\sigma_{ij} & \text{if } t=\tau\\ 0 & \text{otherwise}\end{cases}.$$

Thus, letting $\boldsymbol{\Omega}$ be the $N\times N$ matrix of contemporaneous covariances $\boldsymbol{\Omega}=\{\sigma_{ij}\}$, the covariance matrix of $\boldsymbol{\varepsilon}$ becomes

$$\boldsymbol{\Sigma}=V(\boldsymbol{\varepsilon})=\boldsymbol{\Omega}\otimes\mathbf{I}_T \equiv\begin{bmatrix}\sigma_{11}\mathbf{I}_T & \sigma_{12}\mathbf{I}_T\cdots & \sigma_{1N}\mathbf{I}_T\\ \sigma_{21}\mathbf{I}_T & \sigma_{22}\mathbf{I}_T\cdots & \sigma_{2N}\mathbf{I}_T\\ \vdots & \vdots & \vdots\end{bmatrix}, \tag{8.12}$$

where $\mathbf{I}_T$ denotes a $T\times T$ identity and where $\otimes$ is the Kronecker product. With (8.11) and (8.12), the posterior moments (8.9) and (8.10) become

$$E(\boldsymbol{\beta}|\mathbf{Y},\mathbf{X},\boldsymbol{\Omega})=\left(\mathbf{H}^*+\mathbf{X}'[\boldsymbol{\Omega}\otimes\mathbf{I}_T]^{-1}\mathbf{X}\right)^{-1}\left(\mathbf{H}^*\mathbf{b}^*+\mathbf{X}'[\mathbf{I}_T\otimes\boldsymbol{\Omega}]^{-1}\mathbf{Y}\right) \tag{8.13}$$

$$V(\boldsymbol{\beta}|\mathbf{Y},\mathbf{X},\boldsymbol{\Omega})=\left(\mathbf{H}^*+\mathbf{X}'[\boldsymbol{\Omega}\otimes\mathbf{I}_T]^{-1}\mathbf{X}\right)^{-1} \tag{8.14}$$

with[5]

$$[\boldsymbol{\Omega}\otimes\mathbf{I}_T]^{-1}=\boldsymbol{\Omega}^{-1}\otimes\mathbf{I}_T\ .$$

The further assumption that the explanatory variables are the same in each equation $\mathbf{X}_i=\mathbf{X}_j$ involves no loss in generality, since the list of explanatory variables may be implicitly varied by using prior distributions that concentrate the probability of some parameters in the neighborhood of the origin. In Kronecker notation, let $\mathbf{X}=\mathbf{I}_N\otimes\mathbf{X}_i$, and the formula for the posterior mean becomes

$$E(\boldsymbol{\beta}|\mathbf{Y},\mathbf{X},\boldsymbol{\Omega})=\left(\mathbf{H}^*+\boldsymbol{\Omega}^{-1}\otimes\mathbf{X}_i'\mathbf{X}_i\right)^{-1}\left(\mathbf{H}^*\mathbf{b}^*+(\boldsymbol{\Omega}^{-1}\otimes\mathbf{X}_i')\mathbf{Y}\right)$$

since

$$(\mathbf{I}_N\otimes\mathbf{X}_i')(\boldsymbol{\Omega}^{-1}\otimes\mathbf{I}_T)(\mathbf{I}_N\otimes\mathbf{X}_i)=(\boldsymbol{\Omega}^{-1}\otimes\mathbf{X}_i')(\mathbf{I}_N\otimes\mathbf{X}_i)=\boldsymbol{\Omega}^{-1}\otimes\mathbf{X}_i'\mathbf{X}_i$$

and

$$(\mathbf{I}_N\otimes\mathbf{X}_i')(\boldsymbol{\Omega}^{-1}\otimes\mathbf{I}_T)\mathbf{Y}=(\boldsymbol{\Omega}^{-1}\otimes\mathbf{X}_i')\mathbf{Y}.$$

Two extreme cases imply no pooling of information across equations. If prior information is relatively diffuse, $\mathbf{H}^*=\mathbf{0}$, the posterior location becomes

$$\begin{aligned}E(\boldsymbol{\beta}|\mathbf{Y},\mathbf{X},\boldsymbol{\Omega})&=(\boldsymbol{\Omega}^{-1}\otimes\mathbf{X}_i'\mathbf{X}_i)^{-1}(\boldsymbol{\Omega}^{-1}\otimes\mathbf{X}_i')\mathbf{Y}\\&=\left[\boldsymbol{\Omega}\otimes(\mathbf{X}_i'\mathbf{X}_i)^{-1}\right]\left[\boldsymbol{\Omega}^{-1}\otimes\mathbf{X}_i'\right]\mathbf{Y}\\&=\left[\mathbf{I}_N\otimes(\mathbf{X}_i'\mathbf{X}_i)^{-1}\mathbf{X}_i\right]\mathbf{Y}\\&=\begin{bmatrix}(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{Y}_1\\ (\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2'\mathbf{Y}_2\\ \vdots\\ (\mathbf{X}_N'\mathbf{X}_N)^{-1}\mathbf{X}_N'\mathbf{Y}_N\end{bmatrix},\end{aligned} \tag{8.15}$$

which makes use of the assumptions $\mathbf{X}_i=\mathbf{X}_j$ for all i,j. In words, if prior information is relatively diffuse about the parameters $\boldsymbol{\beta}$, the posterior

[5]See Appendix 1 for the algebra of Kronecker products. Two properties are used here

location can be computed by performing least squares, equation by equation. Similarly, if Ω is a diagonal matrix, the formulas can be written as

$$E(\beta_i|\mathbf{Y},\mathbf{X},\Omega)=\left(\mathbf{H}^*_i+\sigma_{ii}^{-1}\mathbf{X}_i'\mathbf{X}_i\right)^{-1}\left(\mathbf{H}_i^*\mathbf{b}_i^*+\sigma_{ii}^{-1}\mathbf{X}_i'\mathbf{Y}_i\right), \quad i=1,\ldots,N,$$

and the N equations may be analyzed separately.

Thus when the coefficients in the various equations are a priori independent with covariance matrix (8.11), the desire to pool evidence springs from the coincidence of correlated errors and prior information about the coefficients. Since correlation of the error terms may in many circumstances be significant, the critical bottleneck to pooling across equations is the formation of legitimate prior information about the coefficients. It may seem strange that prior information about the coefficients in one equation induces this kind of pooling. More peculiar still is the fact that if β_i is the parameter of interest, until $\mathbf{Y}_i$ is observed, there is no informational value in observations of the other processes $\mathbf{Y}_j$, since the prior and posterior distributions of β_i coincide, $f(\beta_i)=f(\beta_i|\mathbf{Y}_j)$ for $j\neq i$.[6] Thus the only effect of $\mathbf{Y}_j$ is to alter the interpretation of $\mathbf{Y}_i$. This may all become clear if it is pointed out that the conditional mean of the residual in the ith equation depends on the true residual in the jth equation, $E(\boldsymbol{\epsilon}_i|\mathbf{Y}_j,\mathbf{X},\beta_j)=\sigma_{ij}\sigma_{jj}^{-1}(\mathbf{Y}_j-\mathbf{X}_j\beta_j)$. To adjust for this nonzero mean it would be appropriate to regress $\mathbf{Y}_i-\sigma_{ij}\sigma_{jj}^{-1}(\mathbf{Y}_j-\mathbf{X}_j\beta_j)$ on $\mathbf{X}_i$.[7] When the prior for β_j is diffuse, the best estimate of $\mathbf{Y}_j-\mathbf{X}_j\beta_j$ is just the vector of residuals in the jth equation. This vector is by construction orthogonal to $\mathbf{X}_j$ and hence to $\mathbf{X}_i$. This variable can, therefore, have no effect on the estimate of β_i when it is subtracted from $\mathbf{Y}_i$.

The point I have been leading up to may now be clear. The pooling phenomenon associated with multivariate regression is very subtly based on prior information. "Seemingly unrelated" regression estimates should be used with the same kind of care that we would apply to problems that call for more overt forms of prior information. Mechanical, thoughtless use of such routines is to be discouraged.

MULTIVARIATE REGRESSION: CORRELATED COEFFICIENTS

A more direct reason for pooling evidence across equations is a priori correlation of the coefficients. A time series of observations on many individuals is an example; the individuals are unlikely to be identical, but we do expect them to be "similar." For each individual we would specify a

[6]The reader is asked to verify this. Gary Chamberlain has pointed out to me a similar less confusing situation. Suppose in the usual normal regression model $\mathbf{Y}$ depends on two explanatory variables $\mathbf{Y}=\mathbf{x}_1\beta_1+\mathbf{x}_2\beta_2+\mathbf{u}$. The conditional distribution $f(\beta_1|\mathbf{x}_2)$ is independent of $\mathbf{x}_2$, yet $f(\beta_1|\mathbf{Y},\mathbf{x}_2)$ depends on $\mathbf{x}_2$.

[7]The resultant estimate of β_i is $\mathbf{b}_i-\sigma_{ij}\sigma_{jj}^{-1}(\mathbf{b}_j-\beta_j)$ where $\mathbf{b}_i$ is the usual least-squares

different regression equation, but we would have a prior distribution that summarizes the feeling that the coefficients are likely to be similar.

As an illustration of the intuitively compelling reasons for some kind of pooling of observations across processes, suppose we estimated for eight different individuals a linear consumption function with consumption expenditures as the dependent variable and income as the explanatory variable. The following table of least-squares results summarizes the data information about the slope coefficient (the marginal propensity to consume) in these equations:

individual	1	2	3	4	5	6	7	8
least-squares estimate	.81	.80	.82	.86	.85	.84	.86	.83
standard error	2.1	2.4	1.8	1.2	3.6	4.0	5.1	1.1

Note that the standard errors are very large, and a 95% posterior interval for individual one's coefficient, assuming a relatively diffuse prior, would be approximately $.81\pm4.2$. But notice also that the marginal propensities to consume are very close for all individuals. This fact intuitively makes us more confident about the number .81 than is suggested by this interval. Furthermore, we may want to adjust the number .81 upward to make it more representative of the class of estimates.

It goes without saying that it is not always desirable to pool evidence in this way. If equation 1 were a consumption function, equation 2 a production function, equation 3 an investment function, and so forth, the peculiar coincidence of coefficients would be regarded as a statistical artifact, and no pooling would be desirable. But for "similar" processes, pooling is intuitively sensible and, in fact, is necessary to avoid the following "clairvoyant" paradox: In a population of individuals that contains no clairvoyants, you will come to believe with essential certainty that someone is a clairvoyant.

As an example of the clairvoyant problem, suppose N different coins are flipped T times each, in an effort to find a coin that lands heads up with high probability. Let p_i be the probability of a head if coin i is flipped, and take as observations T flips of each of N different coins. Suppose that a prior distribution for these probabilities is selected that would not imply pooling of the evidence across different coins. In particular, let p_i, $i=1,\ldots,N$ be a set of N independent identically distributed random variables. If the number of coins is large enough, there will almost certainly be at least one coin that yielded all heads, even if the probability of a head is one-half for all coins. If, furthermore, T is large enough, the evidence of T heads in T flips will lead to the conclusion that this coin will almost certainly yield a head again. Thus you will conclude that there is a coin that usually lands heads up. This is a perfectly proper Bayesian procedure, and fault cannot be found with it on logical grounds. If you do not like its

mplications, it is only because you do not like the prior distribution. The ›rior implicitly says that the probabilities p_i are drawn independently from ın urn that does contain some values of p close to one. Given enough such elections, you should indeed obtain a p_i close to one. An alternative prior listribution might have the variance of this urn be uncertain. The probabilties p_i would then be a priori dependent, $f(p_1,\ldots,p_N)=\int_\theta[\Pi_i f(p_i|\theta)]d\theta$, where θ is a variance parameter. This prior would imply pooling of the evidence across different coins and would not necessarily lead you to conclude that there is a biased (clairvoyant) coin.

Three "everyday" examples may make the point most forcefully.

Example 1. Several days before a United States presidential election television newsmen find a town that has always voted for the winner of the past elections. A preference poll of the town's inhabitants is then used to predict the outcome of the election.

Example 2. One thousand individuals are sent a letter describing a revolutionary new investment advisory service. Half are told that stock A will rise in value, half are told that it will fall in value. If stock A rises in value those 500 who were so informed are sent another letter. Half are told that stock B will rise; half are told it will fall. By this process you will end up with approximately ten individuals who have been given seven accurate stock tips in a row. It is then time to begin charging for the investment advice.

Example 3. At the end of the first month of the baseball season, there are always some hitters with batting percentages above .400. But at the end of the season, it is very rare to have even one hitter with an average above .400. (For an analysis of batting averages, see Efron and Morris, 1975.)

Returning now to the regression problem, the two significant features of the pooling phenomenon above—shrinking estimates toward a common mean and reducing standard errors—can be effected by selecting a prior covariance matrix $(\mathbf{H}^*)^{-1}$ that does not have the block diagonal form. A convenient way to construct such a matrix is to assume that the vectors $\boldsymbol{\beta}_i$ are an exchangeable normal process, that is, to assume that your opinions about the vectors are normal and unaffected by their ordering. In that event you will act as if the coefficients were selected randomly from a fixed normal urn:[8]

$$\boldsymbol{\beta}_i=\boldsymbol{\beta}+\mathbf{u}_i$$

[8]The notion of exchangeability is due to deFinetti (1937) and is skillfully exploited by Lindley and Smith (1972), from which this section is derived.

with $\mathbf{u}_i$ distributed normally with mean vector zero and covariance matrix $V(\mathbf{u}_i)=\mathbf{V}$. A normal prior for the "hyperparameter" $\bar{\boldsymbol{\beta}}$ with mean $\boldsymbol{\xi}$ and variance $V(\bar{\boldsymbol{\beta}})$, implies that $\boldsymbol{\beta}_i$ is the sum of two normal vectors and is itself normal with moments

$$E\left[\begin{bmatrix}\boldsymbol{\beta}_1\\ \boldsymbol{\beta}_2\\ \vdots\\ \boldsymbol{\beta}_N\end{bmatrix}\right]=\begin{bmatrix}\boldsymbol{\xi}\\ \boldsymbol{\xi}\\ \vdots\\ \boldsymbol{\xi}\end{bmatrix}=\mathbf{1}_N\otimes\boldsymbol{\xi}$$

$$V\left[\begin{bmatrix}\boldsymbol{\beta}_1\\ \boldsymbol{\beta}_2\\ \vdots\\ \boldsymbol{\beta}_N\end{bmatrix}\right]=\begin{bmatrix}\mathbf{V}+V(\bar{\boldsymbol{\beta}}) & V(\bar{\boldsymbol{\beta}}) & \cdots & V(\bar{\boldsymbol{\beta}})\\ V(\bar{\boldsymbol{\beta}}) & \mathbf{V}+V(\bar{\boldsymbol{\beta}}) & & V(\bar{\boldsymbol{\beta}})\\ \vdots & \vdots & & \vdots\\ V(\bar{\boldsymbol{\beta}}) & V(\bar{\boldsymbol{\beta}}) & \cdots & \mathbf{V}+V(\bar{\boldsymbol{\beta}})\end{bmatrix}$$

$$=[\mathbf{I}_N\otimes\mathbf{V}]+[\mathbf{1}_N\otimes\mathbf{I}_k]V(\bar{\boldsymbol{\beta}})[\mathbf{1}_N\otimes\mathbf{I}_k]'.$$

The variance matrix may be inverted to obtain the precision matrix.[9]

$$\mathbf{H}^*=V^{-1}(\boldsymbol{\beta})=[\mathbf{I}_N\otimes\mathbf{V}]^{-1}-[\mathbf{I}_N\otimes\mathbf{V}]^{-1}[\mathbf{1}_N\otimes\mathbf{I}_k]$$
$$\left[(\mathbf{1}_N\otimes\mathbf{I}_k)'(\mathbf{I}_N\otimes\mathbf{V})^{-1}(\mathbf{1}_N\otimes\mathbf{I}_k)+V^{-1}(\bar{\boldsymbol{\beta}})\right]^{-1}$$
$$[\mathbf{1}_N\otimes\mathbf{I}_k]'[\mathbf{I}_N\otimes\mathbf{V}]^{-1}$$
$$=[\mathbf{I}_N\otimes\mathbf{V}^{-1}]-[\mathbf{1}_N\otimes\mathbf{V}^{-1}]\left[\mathbf{1}_N'\mathbf{1}_N\otimes\mathbf{V}^{-1}+V^{-1}(\bar{\boldsymbol{\beta}})\right]^{-1}[\mathbf{1}_N'\otimes\mathbf{V}^{-1}].$$

Finally, under the further natural assumption that information about the mean $\bar{\boldsymbol{\beta}}$ is relatively weak, $V^{-1}(\bar{\boldsymbol{\beta}})$ may be set to a zero matrix to obtain the (singular) precision matrix:

$$\begin{aligned}V^{-1}(\boldsymbol{\beta})&=[\mathbf{I}_N\otimes\mathbf{V}^{-1}]-[\mathbf{1}_N\otimes\mathbf{V}^{-1}][N\otimes\mathbf{V}^{-1}]^{-1}[\mathbf{1}_N'\otimes\mathbf{V}^{-1}]\\&=[\mathbf{I}_N\otimes\mathbf{V}^{-1}]-[\mathbf{1}_N N^{-1}\otimes\mathbf{I}_k][\mathbf{1}_N'\otimes\mathbf{V}^{-1}]\\&=\mathbf{I}_N\otimes\mathbf{V}^{-1}-\mathbf{1}_N N^{-1}\mathbf{1}_N'\otimes\mathbf{V}^{-1}\\&=(\mathbf{I}_N-\mathbf{1}_N N^{-1}\mathbf{1}_N')\otimes\mathbf{V}^{-1}.\end{aligned}$$

[9]Using the formula $(\mathbf{A}+\mathbf{BDB}')^{-1}=\mathbf{A}^{-1}-\mathbf{A}^{-1}\mathbf{B}(\mathbf{B}'\mathbf{A}^{-1}\mathbf{B}+\mathbf{D}^{-1})^{-1}\mathbf{B}'\mathbf{A}^{-1}$.

Note that this precision matrix times the prior location is zero, independent of the choice of ξ

$$\left[(\mathbf{I}_N - \mathbf{1}_N N^{-1}\mathbf{1}_N')\otimes\mathbf{V}^{-1}\right]\left[\mathbf{1}_N\otimes\xi\right]$$
$$=(\mathbf{1}_N - \mathbf{1}_N N^{-1}\mathbf{1}_N'\mathbf{1}_N)\otimes\mathbf{V}^{-1}\xi=\mathbf{1}_N 0\otimes\mathbf{V}^{-1}\xi=\mathbf{0}_{kN}$$

where $\mathbf{0}_{kN}$ is a zero vector of length kN.

The posterior moments may now be written as

$$\mathbf{b}^{**}=E(\beta|\mathbf{Y},\mathbf{X},\Omega)=\left[(\mathbf{I}_N-\mathbf{1}_N N^{-1}\mathbf{1}_N')\otimes\mathbf{V}^{-1}\right.$$
$$\left.+\mathbf{X}'(\Omega^{-1}\otimes\mathbf{I}_T)\mathbf{X}\right]^{-1}\left[\mathbf{X}'(\Omega^{-1}\otimes\mathbf{I}_T)\mathbf{X}\right]\mathbf{b} \qquad (8.16)$$

$$V(\beta|\mathbf{Y},\mathbf{X},\Omega)=\left[(\mathbf{I}_N-\mathbf{1}_N N^{-1}\mathbf{1}_N')\otimes\mathbf{V}^{-1}+\mathbf{X}'(\Omega^{-1}\otimes\mathbf{I}_T)\mathbf{X}\right]^{-1} \qquad (8.17)$$

where $\mathbf{b}$ is a solution to the normal equations

$$\mathbf{X}'(\Omega^{-1}\otimes\mathbf{I}_T)\mathbf{X}\mathbf{b}=\mathbf{X}'(\Omega^{-1}\otimes\mathbf{I}_T)\mathbf{Y}.$$

These formulas can be greatly simplified in the event that Ω is a diagonal matrix.

It is then convenient to write the posterior distribution of β and $\bar{\beta}$ as the product of the conditional distribution of β given $\bar{\beta}$ times a marginal on $\bar{\beta}$. Conditional on $\bar{\beta}$, β_i is independent of β_j, $i\neq j$, with moments

$$E(\beta_i|\bar{\beta},\mathbf{Y},\mathbf{X},\Omega)=(\mathbf{V}^{-1}+\sigma_{ii}^{-1}\mathbf{X}_i'\mathbf{X}_i)^{-1}(\mathbf{V}^{-1}\bar{\beta}+\sigma_{ii}^{-1}\mathbf{X}_i'\mathbf{X}_i\mathbf{b}_i) \qquad (8.18)$$

$$V(\beta_i|\bar{\beta},\mathbf{Y},\mathbf{X},\Omega)=(\mathbf{V}^{-1}+\sigma_{ii}^{-1}\mathbf{X}_i'\mathbf{X}_i)^{-1}. \qquad (8.19)$$

These are the usual formulas if the prior were normal with moments $\bar{\beta}$ and $\mathbf{V}$. To compute the distribution of $\bar{\beta}$ given $\mathbf{Y}$ it is necessary to write the distribution of $\mathbf{Y}$ given $\bar{\beta}$ as

$$\mathbf{Y}_i=\mathbf{X}_i\beta_i+\varepsilon_i=\mathbf{X}_i\bar{\beta}+\mathbf{X}_i\mathbf{u}_i+\varepsilon_i,$$

which is the usual regression process with mean $\mathbf{X}_i\bar{\beta}$ and variance $\mathbf{X}_i\mathbf{V}\mathbf{X}_i'+\sigma_{ii}\mathbf{I}_T$. With $\bar{\beta}$ a priori diffuse and given all the vectors $\mathbf{Y}_i$, we have straightforwardly

$$E(\bar{\beta}|\mathbf{Y},\Omega,\mathbf{X})=\left[\sum_i \mathbf{X}_i'(\mathbf{X}_i\mathbf{V}\mathbf{X}_i'+\sigma_{ii}\mathbf{I}_T)^{-1}\mathbf{X}_i\right]^{-1}\left[\sum_i \mathbf{X}_i'(\mathbf{X}_i\mathbf{V}\mathbf{X}_i'+\sigma_{ii}\mathbf{I}_T)^{-1}\mathbf{Y}_i\right],$$

which is the generalized least-squares estimate of $\bar{\beta}$ and

$$V(\bar{\beta}|\mathbf{Y},\Omega,\mathbf{X})=\left[\sum_i \mathbf{X}_i'(\mathbf{X}_i\mathbf{V}\mathbf{X}_i'+\sigma_{ii}\mathbf{I}_T)^{-1}\mathbf{X}_i\right]^{-1}. \qquad (8.20)$$

These last two formulas can be further simplified by observing that

$$\mathbf{X}_i'(\mathbf{X}_i\mathbf{V}\mathbf{X}_i'+\sigma_{ii}\mathbf{I}_T)^{-1}=\sigma_{ii}^{-1}\mathbf{X}_i'-\sigma_{ii}^{-1}\mathbf{X}_i'\mathbf{X}_i(\sigma_{ii}^{-1}\mathbf{X}_i'\mathbf{X}_i+\mathbf{V}^{-1})^{-1}\mathbf{X}_i'\sigma_{ii}^{-1}$$
$$=\sigma_{ii}^{-1}\mathbf{X}_i'\mathbf{X}_i\left[(\mathbf{X}_i'\mathbf{X}_i)^{-1}-\sigma_{ii}^{-1}(\sigma_{ii}^{-1}\mathbf{X}_i'\mathbf{X}_i+\mathbf{V}^{-1})^{-1}\right]\mathbf{X}_i'$$
$$=\sigma_i^{-1}\mathbf{X}_i'\mathbf{X}_i\left[\sigma_{ii}^{-1}\mathbf{I}_k+(\mathbf{X}_i'\mathbf{X}_i)^{-1}\mathbf{V}^{-1}-\sigma_{ii}^{-1}\mathbf{I}_k\right]$$
$$(\sigma_{ii}^{-1}\mathbf{X}_i'\mathbf{X}_i+\mathbf{V}^{-1})^{-1}\mathbf{X}_i'$$
$$=\sigma_{ii}^{-1}\mathbf{V}^{-1}(\sigma_{ii}^{-1}\mathbf{X}_i'\mathbf{X}_i+\mathbf{V}^{-1})^{-1}\mathbf{X}_i'.$$

Thus we may write

$$E(\bar{\beta}|\mathbf{Y},\Omega,\mathbf{X})=\left[\sum_i(\sigma_{ii}^{-1}\mathbf{X}_i'\mathbf{X}_i+\mathbf{V}^{-1})^{-1}\mathbf{X}_i'\mathbf{X}_i\sigma_{ii}^{-1}\right]^{-1}$$
$$\times\left[\sum_i(\sigma_{ii}^{-1}\mathbf{X}_i'\mathbf{X}_i+\mathbf{V}^{-1})^{-1}\mathbf{X}_i'\mathbf{X}_i\mathbf{b}_i\sigma_{ii}^{-1}\right] \qquad (8.21)$$

where $\mathbf{b}_i=(\mathbf{X}_i'\mathbf{X}_i)^{-1}\mathbf{X}_i'\mathbf{Y}_i$.

The pooling of information across processes with this kind of correlation structure and with the processes otherwise independent (Ω diagonal) is thus summarized by two equations. Equation (8.18) describes the posterior location of β_i as a compromise between the least-squares estimate $\mathbf{b}_i$ and the grand mean $\bar{\beta}$, which is itself a matrix-weighted average (8.21) of each of the least-squares points.

ERROR-COMPONENTS MODEL

What is known as the error-components model, introduced into the econometric literature by Balestra and Nerlove (1966), is a special case of the multivariate model discussed in the previous subsection. The model assumes that the slope vectors in the various processes are identical and also constrains the contemporaneous precision matrix Ω^{-1} to be proportional to a special matrix. The model is written as

$$y_{it}=\sum_{j=1}^{k-1}x_{ijt}\bar{\beta}_j+\bar{\beta}_0+\alpha_i+\gamma_t+\varepsilon_{it} \qquad (8.22)$$

to indicate that there are $k-1$ slope parameters ($\bar{\beta}_j, j=1,\ldots,k-1$) common to every process and that the process level or constant includes four additive variables: a constant $\bar{\beta}_0$; a component α_i, common to all observations of the ith process; a component γ_t, common to all observations in the tth period; and an independent normal error ε_{it}, assumed to have mean zero and variance σ_ε^2. Furthermore, α_i and γ_t are assumed to be indepen-

Note that this precision matrix times the prior location is zero, independent of the choice of ξ

$$\left[(\mathbf{I}_N - \mathbf{1}_N N^{-1}\mathbf{1}_N')\otimes\mathbf{V}^{-1}\right]\left[\mathbf{1}_N\otimes\xi\right]$$
$$=(\mathbf{1}_N - \mathbf{1}_N N^{-1}\mathbf{1}_N'\mathbf{1}_N)\otimes\mathbf{V}^{-1}\xi=\mathbf{1}_N 0\otimes\mathbf{V}^{-1}\xi=\mathbf{0}_{kN}$$

where $\mathbf{0}_{kN}$ is a zero vector of length kN.

The posterior moments may now be written as

$$\mathbf{b}^{**}=E(\beta|\mathbf{Y},\mathbf{X},\Omega)=\left[(\mathbf{I}_N-\mathbf{1}_N N^{-1}\mathbf{1}_N')\otimes\mathbf{V}^{-1}\right.$$
$$\left.+\mathbf{X}'(\Omega^{-1}\otimes\mathbf{I}_T)\mathbf{X}\right]^{-1}\left[\mathbf{X}'(\Omega^{-1}\otimes\mathbf{I}_T)\mathbf{X}\right]\mathbf{b} \qquad (8.16)$$

$$V(\beta|\mathbf{Y},\mathbf{X},\Omega)=\left[(\mathbf{I}_N-\mathbf{1}_N N^{-1}\mathbf{1}_N')\otimes\mathbf{V}^{-1}+\mathbf{X}'(\Omega^{-1}\otimes\mathbf{I}_T)\mathbf{X}\right]^{-1} \qquad (8.17)$$

where $\mathbf{b}$ is a solution to the normal equations

$$\mathbf{X}'(\Omega^{-1}\otimes\mathbf{I}_T)\mathbf{X}\mathbf{b}=\mathbf{X}'(\Omega^{-1}\otimes\mathbf{I}_T)\mathbf{Y}.$$

These formulas can be greatly simplified in the event that Ω is a diagonal matrix.

It is then convenient to write the posterior distribution of β and $\bar{\beta}$ as the product of the conditional distribution of β given $\bar{\beta}$ times a marginal on $\bar{\beta}$. Conditional on $\bar{\beta}$, β_i is independent of β_j, $i\neq j$, with moments

$$E(\beta_i|\bar{\beta},\mathbf{Y},\mathbf{X},\Omega)=(\mathbf{V}^{-1}+\sigma_{ii}^{-1}\mathbf{X}_i'\mathbf{X}_i)^{-1}(\mathbf{V}^{-1}\bar{\beta}+\sigma_{ii}^{-1}\mathbf{X}_i'\mathbf{X}_i\mathbf{b}_i) \qquad (8.18)$$

$$V(\beta_i|\bar{\beta},\mathbf{Y},\mathbf{X},\Omega)=(\mathbf{V}^{-1}+\sigma_{ii}^{-1}\mathbf{X}_i'\mathbf{X}_i)^{-1}. \qquad (8.19)$$

These are the usual formulas if the prior were normal with moments $\bar{\beta}$ and $\mathbf{V}$. To compute the distribution of $\bar{\beta}$ given $\mathbf{Y}$ it is necessary to write the distribution of $\mathbf{Y}$ given $\bar{\beta}$ as

$$\mathbf{Y}_i=\mathbf{X}_i\beta_i+\varepsilon_i=\mathbf{X}_i\bar{\beta}+\mathbf{X}_i\mathbf{u}_i+\varepsilon_i,$$

which is the usual regression process with mean $\mathbf{X}_i\bar{\beta}$ and variance $\mathbf{X}_i\mathbf{V}\mathbf{X}_i'+\sigma_{ii}\mathbf{I}_T$. With $\bar{\beta}$ a priori diffuse and given all the vectors $\mathbf{Y}_i$, we have straightforwardly

$$E(\bar{\beta}|\mathbf{Y},\Omega,\mathbf{X})=\left[\sum_i \mathbf{X}_i'(\mathbf{X}_i\mathbf{V}\mathbf{X}_i'+\sigma_{ii}\mathbf{I}_T)^{-1}\mathbf{X}_i\right]^{-1}\left[\sum_i \mathbf{X}_i'(\mathbf{X}_i\mathbf{V}\mathbf{X}_i'+\sigma_{ii}\mathbf{I}_T)^{-1}\mathbf{Y}_i\right],$$

which is the generalized least-squares estimate of $\bar{\beta}$ and

$$V(\bar{\beta}|\mathbf{Y},\Omega,\mathbf{X})=\left[\sum_i \mathbf{X}_i'(\mathbf{X}_i\mathbf{V}\mathbf{X}_i'+\sigma_{ii}\mathbf{I}_T)^{-1}\mathbf{X}_i\right]^{-1}. \qquad (8.20)$$

These last two formulas can be further simplified by observing that

$$\mathbf{X}_i'(\mathbf{X}_i\mathbf{V}\mathbf{X}_i'+\sigma_{ii}\mathbf{I}_T)^{-1}=\sigma_{ii}^{-1}\mathbf{X}_i'-\sigma_{ii}^{-1}\mathbf{X}_i'\mathbf{X}_i(\sigma_{ii}^{-1}\mathbf{X}_i'\mathbf{X}_i+\mathbf{V}^{-1})^{-1}\mathbf{X}_i'\sigma_{ii}^{-1}$$
$$=\sigma_{ii}^{-1}\mathbf{X}_i'\mathbf{X}_i\left[(\mathbf{X}_i'\mathbf{X}_i)^{-1}-\sigma_{ii}^{-1}(\sigma_{ii}^{-1}\mathbf{X}_i'\mathbf{X}_i+\mathbf{V}^{-1})^{-1}\right]\mathbf{X}_i'$$
$$=\sigma_i^{-1}\mathbf{X}_i'\mathbf{X}_i\left[\sigma_{ii}^{-1}\mathbf{I}_k+(\mathbf{X}_i'\mathbf{X}_i)^{-1}\mathbf{V}^{-1}-\sigma_{ii}^{-1}\mathbf{I}_k\right]$$
$$(\sigma_{ii}^{-1}\mathbf{X}_i'\mathbf{X}_i+\mathbf{V}^{-1})^{-1}\mathbf{X}_i'$$
$$=\sigma_{ii}^{-1}\mathbf{V}^{-1}(\sigma_{ii}^{-1}\mathbf{X}_i'\mathbf{X}_i+\mathbf{V}^{-1})^{-1}\mathbf{X}_i'.$$

Thus we may write

$$E(\bar{\beta}|\mathbf{Y},\Omega,\mathbf{X})=\left[\sum_i(\sigma_{ii}^{-1}\mathbf{X}_i'\mathbf{X}_i+\mathbf{V}^{-1})^{-1}\mathbf{X}_i'\mathbf{X}_i\sigma_{ii}^{-1}\right]^{-1}$$
$$\times\left[\sum_i(\sigma_{ii}^{-1}\mathbf{X}_i'\mathbf{X}_i+\mathbf{V}^{-1})^{-1}\mathbf{X}_i'\mathbf{X}_i\mathbf{b}_i\sigma_{ii}^{-1}\right] \qquad (8.21)$$

where $\mathbf{b}_i=(\mathbf{X}_i'\mathbf{X}_i)^{-1}\mathbf{X}_i'\mathbf{Y}_i$.

The pooling of information across processes with this kind of correlation structure and with the processes otherwise independent (Ω diagonal) is thus summarized by two equations. Equation (8.18) describes the posterior location of β_i as a compromise between the least-squares estimate $\mathbf{b}_i$ and the grand mean $\bar{\beta}$, which is itself a matrix-weighted average (8.21) of each of the least-squares points.

ERROR-COMPONENTS MODEL

What is known as the error-components model, introduced into the econometric literature by Balestra and Nerlove (1966), is a special case of the multivariate model discussed in the previous subsection. The model assumes that the slope vectors in the various processes are identical and also constrains the contemporaneous precision matrix Ω^{-1} to be proportional to a special matrix. The model is written as

$$y_{it}=\sum_{j=1}^{k-1}x_{ijt}\bar{\beta}_j+\bar{\beta}_0+\alpha_i+\gamma_t+\varepsilon_{it} \qquad (8.22)$$

to indicate that there are $k-1$ slope parameters ($\bar{\beta}_j, j=1,\ldots,k-1$) common to every process and that the process level or constant includes four additive variables: a constant $\bar{\beta}_0$; a component α_i, common to all observations of the ith process; a component γ_t, common to all observations in the tth period; and an independent normal error ε_{it}, assumed to have mean zero and variance σ_ε^2. Furthermore, α_i and γ_t are assumed to be indepen-

8.4 Time-Varying Parameters

The usual analysis of a regression process implicitly or explicitly rests on the assumption that the parameters that govern the generation of the data are "more-or-less" the same for all data points. Formal analysis requires the much stricter assumption of perfect constancy, although in practice this, like all other assumptions, is thought to hold in some approximate sense. The parameters are thought to be "sufficiently" constant to allow a fruitful analysis based on the constancy assumption. If the parameters do vary, then estimators describe "some kind of weighted" average of the parameter in question. The vagueness in this informal relaxation of the constancy assumption clearly leaves much to be desired, particularly in poorly specified models, in which "parameters" are functions of the time-varying correlations between the included and excluded variables.

An interesting example of conflicting behavior occurs when a data set is arbitrarily selected. Data sets are often truncated because of the possibility of structural shifts, yet the resulting data subset is then analyzed as if structural changes were impossible. For example, pre-1953 data may be excluded from the analysis on the basis of structural changes. Paradoxically, the same researcher who discards pre-1953 data on the basis of structural change proceeds to analyze the remaining data with simple regression methods. It is, of course, most unlikely that the economic world would undergo an important and fundamental change in 1953 and thereafter remain relatively stagnant. In fact, when we decide to ignore pre-1953 data we are likely to feel that the 1954 data point is only marginally relevant as well.

The heteroscedastic model with declining variances is often suggested in such circumstances, since it can be used to discount the importance of the earlier data points. Although this discounting is intellectually appealing, it is based on an unacceptable assumption about the behavior of the error term; specifically, data points are weighted by the precision of the error term which is assumed to be small for the older observations. However, one's desire to discount the older observations is not related to the precision of the process. Rather, as one gathers older and older data, he begins to question the appropriateness of the constancy assumption. He is likely to be interested in the most recent structure, and the more distant the data point, the less related the structure, and the more meaningless the information obtained. Although the heteroscedastic discount is appealing, there is no assurance that it accurately reflects the decay in the informational value associated with the changing structure, since it is based on the

decay in informational value associated with a decreasing process precision.[11]

From a Bayesian point of view, this problem is straightforward. Every data point may be assumed to be generated by a unique regression process; one observation is made from each process, and the pooling of evidence across processes as described in Section 8.3 applies. Of course, the prior distribution reflects the fact that the regression parameters are thought to be roughly constant over time. This statistical model is the natural extreme of the Bayesian view of inference. Inasmuch as no two data points are related objectively in any way at all, it is impossible to make objective inferences about the nature of the world. Inferences are possible only if subjective prior information is available, that is, only if you (irrationally?) believe the world is orderly.

The model with time-varying parameters has three important implications. First, the discounting of the evidence in earlier observations is based on structural change. Second, the diffuseness of predictions increases naturally as we attempt to project the current structure farther and farther into the future. That is, the value of sample information decays with time, paralleling the decaying relationship between the sample process and the future process. It is intuitively clear that our ability to predict and/or control economic systems decays with time. Stochastic control systems built around a constant parameter assumption result in solutions that rest on greatly overestimated knowledge of the system's future. This tends to result in reckless current decisions, which ignore important elements of uncertainty in the future. The third implication of this model is that "outliers" are legitimately discarded when they suggest structural change. Extreme data points require a suitable adjustment of the regression coefficients applying to the outlier period, and regression coefficients applying to other periods may be insensitive to the presence of the outlier.

The model being discussed is

$$y_t = \mathbf{x}_t'\boldsymbol{\beta}_t + u_t, \qquad t = 1, \ldots, T, \tag{8.23}$$

[11] It should be pointed out that it is possible to build a formal model of time-varying parameters that does imply heteroscedasticity. Cooley and Prescott (1973a,b) write a model as $y_t = \beta x_t + \alpha_t$, where α_t is the time-varying parameter. The stochastic model they suggest can be described by the equations $\alpha_t = u_t + \varepsilon_t, u_t = u_{t-1} + v_t$, with v_t and ε_t being independent spherical normal random variables. Conditional on u_T, say, the variance of α_t is an increasing function of $|T-t|$, which is a heteroscedastic feature. But the model also implies a special kind of correlation between the residuals.

where β_t is a k-dimensional vector of parameters applying in the tth period, $\mathbf{x}_t$ is a $(k\times 1)$ vector of explanatory variables, and u_t is an independent normal random error with mean zero and variance σ^2. This can be written in the form of a multivariate regression as

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_T \end{bmatrix} = \begin{bmatrix} \mathbf{x}_1' & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{x}_2' & \cdots & \mathbf{0} \\ \vdots & \vdots & & \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{x}_T' \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_T \end{bmatrix} + \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_T \end{bmatrix}.$$

The prior distribution that is commonly used for time-varying parameters is normal with mean $E\beta_t = \mathbf{b}^*$ and variance matrix

$$V\begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \vdots \\ \beta_T \end{bmatrix} = \begin{bmatrix} \mathbf{V} & \phi\mathbf{V} & \phi^2\mathbf{V} & \cdots & \phi^{T-1}\mathbf{V} \\ \mathbf{V}\phi' & \mathbf{V} & \phi\mathbf{V} & \cdots & \phi^{T-2}\mathbf{V} \\ \mathbf{V}\phi'^2 & \mathbf{V}\phi' & \mathbf{V} & \cdots & \phi^{T-3}\mathbf{V} \\ \vdots & \vdots & \vdots & & \vdots \\ \mathbf{V}\phi'^{T-1} & \mathbf{V}\phi'^{T-2} & \mathbf{V}\phi'^{T-3} & \cdots & \mathbf{V} \end{bmatrix}.$$

The reader may verify that the conditional moments of such a normal process are

$$E(\beta_t|\beta_{t-1},\beta_{t-2},\ldots) = \mathbf{b}^* + \phi(\beta_{t-1} - \mathbf{b}^*) \tag{8.24}$$

$$V(\beta_t|\beta_{t-1},\beta_{t-2},\ldots) = \mathbf{V} - \phi\mathbf{V}^{-1}\phi. \tag{8.25}$$

The important feature of these moments is that they depend only on the most recent value of the parameter vector. This allows us to write the prior as $f(\beta_T|\beta_{T-1})f(\beta_{T-1}|\beta_{T-2})\cdots f(\beta_1)$.

Suppose first that we observe y_1 only. The distribution conditional on y_1 only is proportional to

$$\begin{aligned} f(\beta|y_1) &\propto f(\beta_T|\beta_{T-1})f(\beta_{T-1}|\beta_{T-2})\cdots f(\beta_1)f(y_1|\beta_1) \\ &= f(\beta_T|\beta_{T-1})f(\beta_{T-1}|\beta_{T-2})\cdots f(\beta_2|\beta_1)f(\beta_1|y_1). \end{aligned}$$

In words, the observation of y_1 affects only the marginal distribution of β_1 and not the conditional distributions. In the usual way the moments are

$$E(\beta_1|y_1) = (\sigma^{-2}\mathbf{x}_1\mathbf{x}_1' + \mathbf{V}^{-1})^{-1}(\sigma^{-2}\mathbf{x}_1 y_1 + \mathbf{V}^{-1}\mathbf{b}^*)$$

$$V(\beta_1|y_1) = (\sigma^{-2}\mathbf{x}_1\mathbf{x}_1' + \mathbf{V}^{-1})^{-1}.$$

The moments of β_2 given y_1 can be computed by integrating out β_1 from the joint distribution $f(\beta_2|\beta_1)f(\beta_1|y_1)$ which we can do simply by using the moments of β_1 and the formulas (8.24) and (8.25):

$$E(\beta_2|y_1) = \mathbf{b}^* + \phi[E(\beta_1|y_1) - \mathbf{b}^*] \tag{8.26}$$

$$V(\beta_2|y_1) = \mathbf{V} - \phi\mathbf{V}^{-1}\phi' + \phi\mathbf{V}(\beta_1|y_1)\phi'. \tag{8.27}$$

Next, suppose y_2 is observed. If interest centers on β_2, the moments just reported can be used as if they were prior moments, since we can write

$$\begin{aligned} f(\beta_2|y_1,y_2) &\propto [\textstyle\int_{\beta_1} f(\beta_2|\beta_1)f(\beta_1)f(y_1|\beta_1)\,d\beta_1]f(y_2|\beta_2) \\ &= f(\beta_2|y_1)f(y_2|\beta_2) \end{aligned}$$

where $f(\beta_2|y_1)$ has the moments (8.26) and (8.27). Thus, as usual, we have the moments of β_2 as

$$\begin{aligned} E(\beta_2|y_1,y_2) &= (V^{-1}(\beta_2|y_1) + \sigma^{-2}\mathbf{x}_2\mathbf{x}_2')^{-1} \\ &\quad \times (V^{-1}(\beta_2|y_1)E(\beta_2|y_1) + \sigma^{-2}\mathbf{x}_2 y_2) \\ V(\beta_2|y_1,y_2) &= (V^{-1}(\beta_2|y_1) + \sigma^{-2}\mathbf{x}_2'\mathbf{x}_2)^{-1}. \end{aligned}$$

Repeated application of this logic leads to the recursive relationships due to Kalman (1960)

$$E(\beta_t|y^{t-1}) = (\mathbf{I}-\phi)\mathbf{b}^* + \phi E(\beta_{t-1}|\mathbf{y}^{t-1})$$

$$E(\beta_t|\mathbf{y}^t) = V(\beta_t|\mathbf{y}^t)(V^{-1}(\beta_t|\mathbf{y}^{t-1})E(\beta_t|\mathbf{y}^{t-1}) + \sigma^{-2}\mathbf{x}_t y_t)$$

where $\mathbf{y}^t = (y_t, y_{t-1}, \ldots, y_1)$, and

$$V^{-1}(\beta_t|\mathbf{y}^t) = V^{-1}(\beta_t|\mathbf{y}^{t-1}) + \sigma^{-2}\mathbf{x}_t\mathbf{x}_t'$$

$$V(\beta_t|\mathbf{y}^{t-1}) = \phi V(\beta_{t-1}|\mathbf{y}^{t-1})\phi' + \mathbf{V} - \phi\mathbf{V}^{-1}\phi'.$$

8.5 Inferences about the Hyperparameters

The reader should have objected before reaching this point that a large number of parameters or hyperparameters whose values are likely to be relatively uncertain have been treated as if they were known. Conceptually it is straightforward to assign a probability distribution to any unknown parameters and to proceed directly to Bayes' rule—probably by integrating out the parameters of little interest. In most cases this is a most unpleasant task. The purpose of this chapter is not to solve real inference problems but only to illustrate how data sets ought to be massaged in a number of

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).