CHAPTER 3

THE LINEAR-REGRESSION MODEL

This chapter reviews theories of inference with the linear-regression model. The first section summarizes the standard theorems of classical inference. The second section deals with pooling information from two samples. The third section describes Bayesian pooling of information from one sample with prior information. And the fourth section describes Bayesian inference about the mean vector and variance-covariance matrix of a multivariate normal distribution.

For more complete introductory material the reader is referred to classical treatments by Johnston (1973), Theil (1971), and Rao (1965), and to Bayesian treatments by Lindley (1965), Zellner (1971), and Box and Tiao (1973).

3.1 Classical Inference with the Linear-Regression Model: A Review

Theoretical descriptions often amount to nothing more than the statement that some hypothetical variable η might depend on some vector of hypothetical variables $\boldsymbol{\chi}$. Empirical workers daringly translate this into a statement about observable variables such as

$$Y_t = \beta_1 X_{1t} + \beta_2 X_{2t} + \cdots + \beta_k X_{kt} + u_t$$

where Y_t is the observable counterpart of η, $\mathbf{X}_t' = (X_{1t}, X_{2t}, \ldots, X_{kt})$ is the observable counterpart of $\boldsymbol{\chi}$, t indexes a set of T observations on each variable $t = 1, \ldots, T$, and where the unobservables $\beta_1, \beta_2, \ldots, \beta_k$ and u_t represent features of the precise relationship between η and $\boldsymbol{\chi}$ that are not disclosed by the theory. The functional dependence of Y_t on $\boldsymbol{\chi}_t$ is thereby assumed to be linear with an unknown parameter vector $\boldsymbol{\beta}' = (\beta_1, \beta_2, \ldots, \beta_k)$, and the theory is asserted to be incomplete, $u_t \neq 0$, or at least the linear approximation to it is incomplete.

The model may be written in matrix form as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u} \tag{3.1}$$

where $\mathbf{Y}$ is a $T \times 1$ vector of observables, $\mathbf{X}$ is a $T \times k$ matrix of observables, $\boldsymbol{\beta}$ is a $k \times 1$ vector of unobservables, and $\mathbf{u}$ is a $T \times 1$ vector of unobservables. Traditionally, $\boldsymbol{\beta}$ is called a parameter vector and $\mathbf{u}$ a disturbance or error vector. By definition, $\mathbf{u}$ is all of those things that determine $\mathbf{Y}$, excluding $\mathbf{X}\boldsymbol{\beta}$, and (3.1) is merely a tautological definition of $\mathbf{u}$, $\mathbf{u} \equiv \mathbf{Y} - \mathbf{X}\boldsymbol{\beta}$. To be more explicit, write $\mathbf{Y}$ as

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

where $\mathbf{Z}\boldsymbol{\gamma}$ is the part of $\mathbf{Y}$ left out of Equation (3.1), with $\mathbf{Z}$ a $(T \times m)$ matrix of variables and $\boldsymbol{\gamma}$ an $(m \times 1)$ vector of fixed effects, and where m need not be finite.

Substantive content is introduced into the analysis by assigning to $\mathbf{u} \equiv \mathbf{Z}\boldsymbol{\gamma}$ a frequency distribution, say, multivariate normal with mean zero and variance matrix $\boldsymbol{\Sigma}$. It is thereby asserted that if the matrix $\mathbf{X}$ were held fixed and the matrix $\mathbf{Z}$ allowed to vary within the confines of some more-or-less well-defined experimental conditions, the vector $\mathbf{Z}\boldsymbol{\gamma}$ would appear to have been drawn from a particular normal distribution. More importantly, it is assumed that there is no *tendency for any of the T residual effects to exceed or fall short of zero on the average*, $E(\mathbf{Z}\boldsymbol{\gamma}|\mathbf{X}) = \mathbf{0}$. Or in the more familiar parlance, the left-out effects are assumed to be uncorrelated with the included effects.

The requirement $E(\mathbf{Z}\boldsymbol{\gamma}|\mathbf{X}) = \mathbf{0}$ is a crucial and quite unlikely assumption. To give a trivial example, suppose in a sample of individuals the dependent variable Y measures sunburn susceptibility and the explanatory variable $\mathbf{X}$ measures hair color. The correlation between these two variables may lead us erroneously to conclude that red hair causes sunburn, when in fact, genetic inheritance (Z) determines both sunburn susceptibility and hair color.

The most foolish errors of inference derive from this kind of model misspecification; yet books about statistical inference hardly mention it. Of course, as discussed in Chapter 1, the choice of specification in this sense is not within the purview of the theory of statistical inference, which almost by definition takes the model as well defined. This enormously restricts the usefulness of statistical theory in nonexperimental research. It is sensible

sometimes during a data analysis to take the model as well specified; but it is senseless always to do so. Any inference will be thoroughly discredited by the identification of particular left-out variables. In devoting a whole chapter to post-data model construction, this book places relatively great emphasis on this topic. Even here, the heavy allocation of space to problems that presume the definition of the underlying causal model reflects what *can* be said about the various topics, not what *should* be said. More is said of this in Chapter 9, but until then the specification assumption $E(\mathbf{u}|\mathbf{X})=\mathbf{0}$ should be taken as given.

FINDING A COMPLETE THEORY

An initial question that may be asked of the data is whether the theory is, in fact, complete, $\mathbf{u}=\mathbf{0}$. In particular, is there a value of $\boldsymbol{\beta}$, say, $\mathbf{b}$, that makes the theory appear to be complete, $\mathbf{Y}-\mathbf{X}\mathbf{b}=\mathbf{0}$? Whenever the number of observations T exceeds the number of parameters k, such a value of $\mathbf{b}$ is quite unlikely to exist. We can, however, find a value that makes the theory appear as complete as possible by making the residual vector $\mathbf{e}=\mathbf{Y}-\mathbf{X}\mathbf{b}$ as small as possible. One (arbitrary) measure of smallness is the square of the length of $\mathbf{e}, \mathbf{e}'\mathbf{e}$. Minimizing $\mathbf{e}'\mathbf{e}$ by setting its derivatives to zero implies the equations

$$\mathbf{0}=\frac{\partial}{\partial\boldsymbol{\beta}}(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})=-2\mathbf{X}'\mathbf{Y}+2\mathbf{X}'\mathbf{X}\boldsymbol{\beta}.$$

Solving this with the assumption that $\mathbf{X}'\mathbf{X}$ is invertible yields the least squares value of $\boldsymbol{\beta}$

$$\mathbf{b}=(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}. \tag{3.2}$$

THE MULTIPLE CORRELATION COEFFICIENT

Although $\mathbf{e}'\mathbf{e}$ is one measure of the completeness of the theory, it is more common to use an R^2 defined by

$$R^2=\left[\sum_t\left(Y_t-\bar{Y}\right)^2-\mathbf{e}'\mathbf{e}\right]\Big/\sum_t\left(Y_t-\bar{Y}\right)^2, \tag{3.3}$$

where $\bar{Y}=\Sigma Y_t/T$. The R^2 can be shown to lie between zero and one. A model with only a constant term $Y_t=\beta_1+u_t$ can be written in vector notation as $\mathbf{Y}=\mathbf{1}_T\beta_1+\mathbf{u}$, where $\mathbf{1}_T$ is a $T\times 1$ vector of ones. The least-squares estimate of β_1 is $\hat{\beta}_1=(\mathbf{1}_T'\mathbf{1}_T)^{-1}\mathbf{1}_T'\mathbf{Y}=\bar{Y}$, where $\bar{Y}$ is the mean of the observations Y_t, $t=1,\ldots,T$. The corresponding residual sum of squares is $(\mathbf{Y}-\mathbf{1}_T\hat{\beta}_1)'(\mathbf{Y}-\mathbf{1}_T\hat{\beta}_1)=\Sigma(Y_t-\bar{Y})^2$. *Assuming that the first column of* $\mathbf{X}$ *is a vector of ones,* it is always possible to make $\mathbf{e}'\mathbf{e}$ equal to $\Sigma(Y_t-\bar{Y})^2$ by setting $\boldsymbol{\beta}'=(\bar{Y},0,0,\ldots,0)$. Since we are choosing $\mathbf{b}$ to make $\mathbf{e}'\mathbf{e}$ as small as possible, it must be true that $0\leqslant\mathbf{e}'\mathbf{e}\leqslant\Sigma(Y_t-\bar{Y})^2$ and thus $0\leqslant R^2\leqslant 1$. An

R^2 of one means that the model is possibly complete, $\mathbf{e}'\mathbf{e}=0$, whereas an R^2 of zero means that the model is totally incapable of explaining the variability in the observed data.

ESTIMATING $\boldsymbol{\beta}$

The least squares vector $\mathbf{b}$ has properties other than making the theory appear as complete as possible. A frequency distribution for $\mathbf{u}$ with $E\mathbf{u}=\mathbf{0}_T$, and $V\mathbf{u}=\sigma^2\mathbf{I}_T$ implies a frequency distribution for $\mathbf{b}$. ($\mathbf{I}_T$ is the $T\times T$ identity matrix. $\mathbf{0}_T$ is a $T\times 1$ vector of zeros.) Since $\mathbf{b}$ is a linear function of $\mathbf{u}$,

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

the moments of $\mathbf{b}$ are straightforwardly calculated as

$$E\mathbf{b}=\boldsymbol{\beta}+(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E(\mathbf{u})=\boldsymbol{\beta},$$

$$V\mathbf{b}=(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'V(\mathbf{u})\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}=\sigma^2(\mathbf{X}'\mathbf{X})^{-1}.$$

It is important to understand that these moments describe the behavior of $\mathbf{b}$ in repeated samples. This is, if the "experiment" is repeated over and over with fixed $\mathbf{X}$ matrix, the average value of $\mathbf{b}$ would be $\boldsymbol{\beta}$. In that sense only, a particular value of $\mathbf{b}$ is taken as indicating where $\boldsymbol{\beta}$ lies. These results are summarized as follows:

THEOREM 3.1 (MOMENTS OF LEAST-SQUARES ESTIMATOR). *Assuming* $\mathbf{X}'\mathbf{X}$ *is invertible,* $E\mathbf{u}=\mathbf{0}_T$ *and* $V(\mathbf{u})=\sigma^2\mathbf{I}_T$, *the least-squares estimator of* $\boldsymbol{\beta}$ *is* $\mathbf{b}=(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$, *with mean* $\boldsymbol{\beta}$ *and variance* $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$.

A desirable feature of the estimator $\mathbf{b}$ from the standpoint of calculation is that it is a linear function of $\mathbf{Y}$. It is also an unbiased estimator of $\boldsymbol{\beta}$, $E(\mathbf{b})=\boldsymbol{\beta}$. It is of interest to deterrmine if some other unbiased linear estimator might be better. Suppose we wished to estimate the scalar parameter $\theta=\boldsymbol{\psi}'\boldsymbol{\beta}$ where $\boldsymbol{\psi}$ is a k-dimensional vector of constants. The least-squares estimator of θ is taken to be $\boldsymbol{\psi}'\mathbf{b}$. An alternative linear estimator is

$$\hat{\theta}=\mathbf{A}'\mathbf{Y}+a$$

where $\mathbf{A}$ is a vector of constants, and a is a scalar constant. The estimator is unbiased only if

$$\boldsymbol{\psi}'\boldsymbol{\beta}=E\hat{\theta}, \qquad \text{for all} \quad \boldsymbol{\beta},$$

that is, only if

$$\boldsymbol{\psi}'\boldsymbol{\beta}=A'E\mathbf{Y}+a=\mathbf{A}'\mathbf{X}\boldsymbol{\beta}+a.$$

For this to be true for all $\boldsymbol{\beta}$ we must have

$$a=0, \qquad \psi'=\mathbf{A}'\mathbf{X}.$$

The variance of the estimator $\hat{\boldsymbol{\theta}}$ is

$$V\hat{\boldsymbol{\theta}}=\mathbf{A}'V(\mathbf{Y})\mathbf{A}=\sigma^2\mathbf{A}'\mathbf{A}.$$

Minimization of the variance of $\hat{\boldsymbol{\theta}}$ subject to the unbiasedness restriction is a simple constrained maximization problem. With $\boldsymbol{\lambda}$ as the vector of k Lagrange multipliers, this requires the derivatives of the function $\sigma^2\mathbf{A}'\mathbf{A}+2\boldsymbol{\lambda}'(\psi-\mathbf{X}'\mathbf{A})$ to be set to zero; that is,

$$\begin{aligned}\mathbf{0}&=\frac{\partial f}{\partial \mathbf{A}}=2\sigma^2\mathbf{A}'-2\boldsymbol{\lambda}'\mathbf{X}'\\ \mathbf{0}&=\frac{\partial f}{\partial \boldsymbol{\lambda}}=\psi'-\mathbf{A}'\mathbf{X}\end{aligned} \tag{3.4}$$

Postmultiplying the first equation by $\mathbf{X}$ we obtain

$$\sigma^2\mathbf{A}'\mathbf{X}-\boldsymbol{\lambda}'\mathbf{X}'\mathbf{X}=\mathbf{0}$$

which implies

$$\boldsymbol{\lambda}'=\sigma^2\mathbf{A}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}=\sigma^2\psi'(\mathbf{X}'\mathbf{X})^{-1}.$$

Substituting this last expression into (3.4) yields

$$\mathbf{0}=2\sigma^2\mathbf{A}'-2\sigma^2\psi'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$

or

$$\mathbf{A}'=\psi'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'.$$

Thus the minimum variance linear unbiased estimator of $\psi'\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\theta}}=\mathbf{A}'\mathbf{Y}=\psi'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y},$$

which is, of course, just the least-squares estimator, and we have established the following result.

THEOREM 3.2 (GAUSS–MARKOV). *Assuming* $\mathbf{X}'\mathbf{X}$ *is invertible,* $E(\mathbf{u})=\mathbf{0}$ *and* $V(\mathbf{u})=\sigma^2\mathbf{I}$ *with* σ^2 *finite, the least-squares estimator* $\psi'\mathbf{b}$ *of the linear combination of coefficients* $\psi'\boldsymbol{\beta}$ *has minimum variance among the class of unbiased linear estimators.*

If $\mathbf{u}$ has a normal distribution, the least-squares estimator $\mathbf{b}$ which is a linear function of $\mathbf{u}$ is also normally distributed. Furthermore, $\mathbf{b}$ is the maximum likelihood estimator of $\boldsymbol{\beta}$. The distribution of $\mathbf{Y}$ is

$$f(\mathbf{Y}|\mathbf{X},\boldsymbol{\beta},\sigma^2)=(2\pi\sigma^2)^{-T/2}\exp\left[-\frac{1}{2\sigma^2}(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})\right]$$

Using the fact that

$$\begin{aligned}(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})&=(\mathbf{Y}-\mathbf{X}\mathbf{b}+\mathbf{X}\mathbf{b}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{Y}-\mathbf{X}\mathbf{b}+\mathbf{X}\mathbf{b}-\mathbf{X}\boldsymbol{\beta})\\&=(\mathbf{Y}-\mathbf{X}\mathbf{b})'(\mathbf{Y}-\mathbf{X}\mathbf{b})+(\boldsymbol{\beta}-\mathbf{b})'\mathbf{X}'\mathbf{X}(\boldsymbol{\beta}-\mathbf{b})\end{aligned} \tag{3.5}$$

since $\mathbf{X}'(\mathbf{Y}-\mathbf{X}\mathbf{b})=\mathbf{0}$, we may write the likelihood function as

$$L(\boldsymbol{\beta},\sigma^2;\mathbf{Y},\mathbf{X})$$

$$\propto(\sigma^2)^{-T/2}\exp\left[-\frac{1}{2\sigma^2}(\mathbf{Y}-\mathbf{X}\mathbf{b})'(\mathbf{Y}-\mathbf{X}\mathbf{b})\right]\exp\left[-\frac{1}{2\sigma^2}(\boldsymbol{\beta}-\mathbf{b})'\mathbf{X}'\mathbf{X}(\boldsymbol{\beta}-\mathbf{b})\right] \tag{3.6}$$

which attains its maximum at

$$\boldsymbol{\beta}=\mathbf{b}, \qquad \sigma^2=\frac{(\mathbf{Y}-\mathbf{X}\mathbf{b})'(\mathbf{Y}-\mathbf{X}\mathbf{b})}{T}=\hat{\sigma}^2.$$

Summarizing these results in a theorem:

THEOREM 3.3 (MAXIMUM LIKELIHOOD ESTIMATOR). *Assuming that* $\mathbf{X}'\mathbf{X}$ *is invertible and that* $\mathbf{u}$ *is normally distributed with mean vector* $\mathbf{0}_T$ *and variance matrix* $\sigma^2\mathbf{I}_T$, *the least-squares estimator* $\mathbf{b}$ *is the maximum likelihood estimator and is normally distributed with mean* $\boldsymbol{\beta}$ *and variance* $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$.

ESTIMATING σ^2

The maximum likelihood estimator of σ^2 is $\hat{\sigma}^2=(\mathbf{Y}-\mathbf{X}\mathbf{b})'(\mathbf{Y}-\mathbf{X}\mathbf{b})/T$. With $\mathbf{M}_x=\mathbf{I}-\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ this estimator can be written as $\hat{\sigma}^2=\mathbf{Y}'\mathbf{M}_x\mathbf{Y}/T=\mathbf{u}\mathbf{M}_x\mathbf{u}/T$, since $\mathbf{M}_x\mathbf{X}=\mathbf{0}$. The expected value of $\hat{\sigma}^2$ is $E(\hat{\sigma}^2)=E(\mathbf{u}'\mathbf{M}_x\mathbf{u}/T)=E(tr[\mathbf{u}'\mathbf{M}_x\mathbf{u}]/T)=E(tr[\mathbf{M}_x\mathbf{u}\mathbf{u}']/T)=\sigma^2 tr\mathbf{M}_x/T=\sigma^2(T-tr[\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'])/T=\sigma^2(T-tr[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}])/T=\sigma^2(T-k)/T$. Therefore $\hat{\sigma}^2$ is biased estimator, $E(\hat{\sigma}^2)\neq\sigma^2$. The unbiased estimator is

$$s^2\equiv\frac{(\mathbf{Y}-\mathbf{X}\mathbf{b})'(\mathbf{Y}-\mathbf{X}\mathbf{b})}{T-k}.$$

CONFIDENCE REGIONS AND HYPOTHESIS TESTS

The construction of confidence intervals and hypothesis tests makes use of the following result.

LEMMA. *If the* $p\times 1$ *vector* $\mathbf{z}$ *is normally distributed with mean* $E(\mathbf{z})$ *and variance* $V(\mathbf{z})$, *then the quantity* $[\mathbf{z}-E\mathbf{z}]'[V(\mathbf{z})]^{-1}[\mathbf{z}-E\mathbf{z}]$ *is the sum of squares of* p *independent, standard, normal, random variables, which by definition has a chi-square distribution on* p *degrees of freedom.*

The proof of this lemma follows straightforwardly from the observation that there exists a matrix $\mathbf{C}$ such that $\mathbf{C}V(\mathbf{z})\mathbf{C}'$ is the identity matrix, and therefore, $\mathbf{Cz}-\mathbf{C}E(\mathbf{z})$ has mean vector zero and an identity variance matrix. Thus $[\mathbf{Cz}-\mathbf{C}E(\mathbf{z})]'[\mathbf{Cz}-\mathbf{C}E(\mathbf{z})]=[\mathbf{z}-E(\mathbf{z})]'\mathbf{C}'\mathbf{C}[\mathbf{z}-E(\mathbf{z})]=[\mathbf{z}-E(\mathbf{z})][V^{-1}(\mathbf{z})][\mathbf{z}-E(\mathbf{z})]$ is the sum of squares of p independent standard normal random variables.

Partitioning the matrices conformably,

$$\mathbf{b}'=(\mathbf{b}_I',\mathbf{b}_J')$$

$$\boldsymbol{\beta}'=(\boldsymbol{\beta}_I',\boldsymbol{\beta}_J')$$

$$(\mathbf{X}'\mathbf{X})^{-1}=\begin{bmatrix}\left[(\mathbf{X}'\mathbf{X})^{-1}\right]_{II} & \left[(\mathbf{X}'\mathbf{X})^{-1}\right]_{IJ}\\ \left[(\mathbf{X}'\mathbf{X})^{-1}\right]_{JI} & \left[(\mathbf{X}'\mathbf{X})^{-1}\right]_{JJ}\end{bmatrix},$$

and making use of the fact that marginal distributions of multivariate normal random variables are themselves normal, we have the following consequence of this lemma.

THEOREM 3.4 (THE CHI-SQUARE TEST STATISTIC). *The scalar random variable*

$$[\mathbf{b}_I-\boldsymbol{\beta}_I]'\left[\sigma^2(\mathbf{X}'\mathbf{X})^{-1}\right]_{II}^{-1}[\mathbf{b}_I-\boldsymbol{\beta}_I]=\chi^2 \tag{3.6}$$

has a chi-square distribution with p degrees of freedom, where p is the dimension of $\mathbf{b}_I$.

Thus if $\chi_\alpha^2(p)$ is the upper α percentage point of the chi-square distribution with p degrees of freedom, then the ellipsoid [1]

$$[\mathbf{b}_I-\boldsymbol{\beta}_I]'\left[\sigma^2(\mathbf{X}'\mathbf{X})^{-1}\right]_{II}^{-1}[\mathbf{b}_I-\boldsymbol{\beta}_I]\leqslant\chi_\alpha^2(p) \tag{3.7}$$

is a $1-\alpha$% confidence region for $\boldsymbol{\beta}$. Similarly, an α-level test of the point null hypothesis $\boldsymbol{\beta}_I=\boldsymbol{\beta}_I^*$ versus the alternative $\boldsymbol{\beta}_I\neq\boldsymbol{\beta}_I^*$, would reject the null hypothesis if $\boldsymbol{\beta}_I=\boldsymbol{\beta}_I^*$ were outside this region.

If σ^2 is unknown, these statements, although still true, lose their usefulness, since the regions described are functions of σ^2. It seems natural to use some estimate in place of the unknown σ^2. Remarkably enough, for the normal linear regression model this is approximately correct. To show that, we need the following result.

[1] Actually, this is an ellipsoidal cylinder with $\boldsymbol{\beta}_J$ unconstrained.

THEOREM 3.5. *The quantity $(\mathbf{Y}-\mathbf{Xb})'(\mathbf{Y}-\mathbf{Xb})/\sigma^2=(T-k)s^2/\sigma^2$ has a chi-square distribution with $T-k$ degrees of freedom and is independent of $\boldsymbol{\beta}-\mathbf{b}$.*

The residual sum of squares $(\mathbf{Y}-\mathbf{Xb})'(\mathbf{Y}-\mathbf{Xb})$ can be written as $\mathbf{Y}'\mathbf{M}_x\mathbf{Y}$ where $\mathbf{M}_x=\mathbf{I}_T-\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$. Making use of $\mathbf{M}_x\mathbf{X}=\mathbf{0}$, the residual sum of squares can also be written

$$(\mathbf{Y}-\mathbf{Xb})'(\mathbf{Y}-\mathbf{Xb})=\mathbf{u}'\mathbf{M}_x\mathbf{u}.$$

The idempotent matrix $\mathbf{M}_x$ is shown in Appendix 1 to have k characteristic values equal to zero and the remaining $T-k$ equal to one. Thus $\mathbf{u}'\mathbf{M}_x\mathbf{u}$ can be written as $\mathbf{u}'\mathbf{C}'\boldsymbol{\Lambda}\mathbf{C}\mathbf{u}=\mathbf{u}^{*\prime}\boldsymbol{\Lambda}\mathbf{u}^*$ where $\mathbf{u}^*$ is normally distributed with mean zero and variance matrix $\sigma^2\mathbf{I}_T$, and where $\boldsymbol{\Lambda}$ is a diagonal matrix with k zero diagonal elements and $T-k$ elements equal to one. We have thus written $(\mathbf{Y}-\mathbf{Xb})'(\mathbf{Y}-\mathbf{Xb})/\sigma^2$ as the sum of squares of $T-k$ independent, standard, normal, random variables, which by definition is distributed chi-square with $T-k$ degrees of freedom.

The independence of the two random variables in the theorem follows from the independence of $\mathbf{Y}-\mathbf{Xb}=\mathbf{M}_x\mathbf{u}$ and $(\boldsymbol{\beta}-\mathbf{b})=(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}$. Given the normality assumption, these quantities may be shown to be independent by computing their covariance $E[\mathbf{M}_x\mathbf{u}][(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}]'=\sigma^2\mathbf{M}_x\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}=\mathbf{0}$.

Definition. If the random variables χ_1^2 and χ_2^2 are independent and have chi-square distributions with p_1 and p_2 degrees of freedom, respectively, then the ratio

$$F=\frac{(\chi_1^2/p_1)}{(\chi_2^2/p_2)}$$

has an F-distribution with degrees of freedom parameters p_1 and p_2.

THEOREM 3.6 (THE F-TEST STATISTIC). *Assuming that $\mathbf{X}'\mathbf{X}$ is invertible and that $\mathbf{u}$ is normally distributed with mean vector $\mathbf{0}_T$ and variance matrix $\sigma^2\mathbf{I}_T$, the quantity*

$$\frac{[\mathbf{b}_I-\boldsymbol{\beta}_I]'\left[(\mathbf{X}'\mathbf{X})^{-1}\right]_{II}^{-1}[\mathbf{b}_I-\boldsymbol{\beta}_I]}{ps^2}=F \tag{3.8}$$

has an F distribution with p and $T-k$ degrees of freedom, where p is the dimension of $\mathbf{b}_I$.

This is a direct consequence of Theorems 3.4 and 3.5 and should be compared with Theorem 3.4, which applies when σ^2 is known. The statement implies $1-\alpha\%$ confidence ellipsoids of the form

$$[\mathbf{b}_I-\boldsymbol{\beta}_I]'\left[(\mathbf{X}'\mathbf{X})^{-1}\right]_{II}^{-1}[\mathbf{b}_I-\boldsymbol{\beta}_I]\leqslant ps^2F_\alpha(p,T-k)$$

where p is the dimension of $\mathbf{b}_I$; $p\leqslant k$, and $F_\alpha(p,T-k)$ is the upper α percentage point of an F distribution with p and $T-k$ degrees of freedom. In the special case when $p=1$, $F^{1/2}$ is said to have a Student's t distribution, and this ellipsoid becomes a confidence interval for a particular coefficient

$$|b_i-\beta_i|\leqslant s\left[(\mathbf{X}'\mathbf{X})^{-1}\right]_{ii}^{-1}t_\alpha(T-k).$$

Furthermore, an α-level test of the hypothesis $\boldsymbol{\beta}_I=\boldsymbol{\beta}_I^*$ versus the alternative $\boldsymbol{\beta}_I\neq\boldsymbol{\beta}_I^*$ would reject the hypothesis if $\boldsymbol{\beta}_I^*$ were not within the confidence region.

The quadratic form (3.3) can be written in another informative way. We may partition $\mathbf{X}'\mathbf{X}$ as above,

$$\mathbf{X}'\mathbf{X}=\begin{bmatrix}\mathbf{X}_I'\mathbf{X}_I & \mathbf{X}_I'\mathbf{X}_J\\ \mathbf{X}_J'\mathbf{X}_I & \mathbf{X}_J'\mathbf{X}_J\end{bmatrix}$$

and, by the partitioned inverse rule, we have

$$\left[(\mathbf{X}'\mathbf{X})^{-1}\right]_{II}^{-1}=\mathbf{X}_I'\mathbf{X}_I-\mathbf{X}_I'\mathbf{X}_J(\mathbf{X}_J'\mathbf{X}_J)^{-1}\mathbf{X}_J'\mathbf{X}_I.$$

Let the minimum error sum of squares be

$$ESS=(\mathbf{Y}-\mathbf{X}\mathbf{b})'(\mathbf{Y}-\mathbf{X}\mathbf{b}),$$

and the constrained minimum error sum of squares with constraint $\boldsymbol{\beta}_I=\boldsymbol{\beta}_I^*$ be

$$ESS_0=(\mathbf{Y}-\mathbf{X}_I\boldsymbol{\beta}_I^*-\mathbf{X}_J\mathbf{b}_J^*)'(\mathbf{Y}-\mathbf{X}_I\boldsymbol{\beta}_I^*-\mathbf{X}_J\mathbf{b}_J^*)$$

with $\mathbf{b}_J^*=(\mathbf{X}_J'\mathbf{X}_J)^{-1}\mathbf{X}_J'(\mathbf{Y}-\mathbf{X}_I\boldsymbol{\beta}_I^*)$. By some tedious manipulations we may show that

$$(\boldsymbol{\beta}_I^*-\mathbf{b}_I)'\left[(\mathbf{X}'\mathbf{X})^{-1}\right]_{II}^{-1}(\boldsymbol{\beta}_I^*-\mathbf{b}_I)=ESS_0-ESS \tag{3.9}$$

and therefore the numerator of the F statistic is just the increase in the error sum of squares implied by the restriction $\boldsymbol{\beta}_I=\boldsymbol{\beta}_I^*$

$$F=\frac{ESS_0-ESS}{ps^2} \tag{3.10}$$

Furthermore, using the definition of R^2 given by Equation (3.3) and letting R^2 correspond to the unconstrained regression and R_0^2 correspond to the constrained regression, the F statistic may also be written as

$$F=\left(\frac{R^2-R_0^2}{1-R^2}\right)\left(\frac{T-k}{p}\right) \tag{3.11}$$

CONSTRAINED ESTIMATION

The consequences of constraints other than $\boldsymbol{\beta}_I=\boldsymbol{\beta}_I^*$ are easy to compute. Minimization of the sum of squares $(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})$ subject to restriction $\mathbf{R}\boldsymbol{\beta}=\mathbf{r}$ is a Lagrangian problem requiring the derivatives of the following expression to be set to zero.

$$(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})-2\boldsymbol{\lambda}'(\mathbf{R}\boldsymbol{\beta}-\mathbf{r}).$$

That is

$$\mathbf{R}\boldsymbol{\beta}-\mathbf{r}=\mathbf{0} \tag{3.12}$$

$$\mathbf{X}'(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})-\mathbf{R}'\boldsymbol{\lambda}=\mathbf{0}. \tag{3.13}$$

Premultiplying (3.13) by $\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}$ yields

$$\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}-\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{X})\boldsymbol{\beta}=\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'\boldsymbol{\lambda}$$

and using (3.12)

$$\boldsymbol{\lambda}=\left(\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'\right)^{-1}(\mathbf{R}\mathbf{b}-\mathbf{r}),$$

which can be reinserted into (3.13) to yield, finally, the constrained least-squares value

$$\begin{aligned}\hat{\boldsymbol{\beta}}(\mathbf{R},\mathbf{r})&=(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}-(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'\left(\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'\right)^{-1}(\mathbf{R}\mathbf{b}-\mathbf{r})\\&=\mathbf{b}-(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'\left(\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'\right)^{-1}(\mathbf{R}\mathbf{b}-\mathbf{r}).\end{aligned} \tag{3.14}$$

Using this value of $\hat{\boldsymbol{\beta}}$ in the sum of squares expression we have

$$(\mathbf{Y}-\mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{Y}-\mathbf{X}\hat{\boldsymbol{\beta}})=(\mathbf{Y}-\mathbf{X}\mathbf{b})'(\mathbf{Y}-\mathbf{X}\mathbf{b})+(\mathbf{R}\mathbf{b}-\mathbf{r})'\left(\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'\right)^{-1}(\mathbf{R}\mathbf{b}-\mathbf{r}) \tag{3.15}$$

where the last term in the last line is, therefore, the increase in the error sum of squares implied by the restriction $\mathbf{R}\boldsymbol{\beta}=\mathbf{r}$. By inspection, this establishes the validity of Equation (3.9) with $\mathbf{r}=\boldsymbol{\beta}_I^*$ and $\mathbf{R}=(\mathbf{I}_p\mathbf{0})$. Computation of the mean and variance of $\hat{\boldsymbol{\beta}}$ is left as an exercise.

TREATMENT OF THE CONSTANT TERM

A regression function ordinarily has a constant as one of the "variables." Equivalently, one of the columns of the $\mathbf{X}$ matrix is a vector of ones. Although everything that has been said still applies, it is informative to treat the constant somewhat differently.

Let us remove from the $\mathbf{X}$ matrix the vector of ones and write the gression as

$$\mathbf{Y}=\mathbf{X}\boldsymbol{\beta}+\mathbf{1}_T\alpha+\mathbf{u}$$

here $\mathbf{1}_T$ is a vector of ones and α is the scalar constant. The error sum of quares may be minimized with respect to α as

$$\begin{aligned}0&=\frac{\partial}{\partial\alpha}(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{1}_T\alpha)'(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{1}_T\alpha)\\&=-2\mathbf{1}_T'(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})-2\mathbf{1}_T'\mathbf{1}_T\alpha\end{aligned}$$

hus as a function of $\boldsymbol{\beta}$, the least-squares estimate of α is

$$a(\boldsymbol{\beta})=\mathbf{1}_T'(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})(\mathbf{1}_T'\mathbf{1}_T)^{-1}.$$

nserting this value into the error sum of squares yields

$$\begin{aligned}&\left(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{1}_T\mathbf{1}_T'[\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}]T^{-1}\right)'\left(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{1}_T\mathbf{1}_T'[\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}]T^{-1}\right)\\&\quad=(\mathbf{MY}-\mathbf{MX}\boldsymbol{\beta})'(\mathbf{MY}-\mathbf{MX}\boldsymbol{\beta})\end{aligned}$$

where $\mathbf{M}=\mathbf{I}_T-\mathbf{1}_T(\mathbf{1}_T'\mathbf{1}_T)^{-1}\mathbf{1}_T'$. The least-squares estimate of $\boldsymbol{\beta}$ can, therefore, be computed by transforming the observations $(\mathbf{Y},\mathbf{X})$ by the matrix $\mathbf{M}$ and minimizing the transformed error sum of squares. It is straightforward to show that this yields

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

But now notice the effect of the matrix $\mathbf{M}$ on a vector $\mathbf{Y}$:

$$\mathbf{MY}=\mathbf{Y}-\mathbf{1}_T(\mathbf{1}_T'\mathbf{Y})T^{-1}=\mathbf{Y}-\mathbf{1}_T\bar{Y}$$

where $\bar{Y}$ is the average value of the observed variables. In words, the matrix $\mathbf{M}$ subtracts out the mean of a variable, and the least-squares estimate of the slope coefficients can be computed by first subtracting the means from all the variables and then computing the least-squares value in the usual way.

Parenthetically, if the regression process is written $\mathbf{Y}=\mathbf{X}\boldsymbol{\beta}+\mathbf{Z}\boldsymbol{\gamma}+\mathbf{u}$ where $\mathbf{X}$ and $\mathbf{Z}$ are observable matrices, the least-squares estimate of $\boldsymbol{\beta}$ is $\mathbf{b}=(\mathbf{X}'\mathbf{M}_z'\mathbf{M}_z\mathbf{X})^{-1}\mathbf{X}'\mathbf{M}_z\mathbf{M}_z\mathbf{Y}=(\mathbf{X}'\mathbf{M}_z\mathbf{X})^{-1}\mathbf{X}'\mathbf{M}_z\mathbf{Y}$ where $\mathbf{M}_z=\mathbf{I}_T-\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'$. The residual vector formed from the regression of $\mathbf{Y}$ on $\mathbf{Z}$ is $\mathbf{Y}-\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{Y}=\mathbf{M}_z\mathbf{Y}$ and the residual matrix formed from the regression of each of the columns of $\mathbf{X}$ on $\mathbf{Z}$ is $\mathbf{X}-\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}=\mathbf{M}_z\mathbf{X}$. The least-squares estimate $\mathbf{b}$ is then just the regression of the residual vector $\mathbf{M}_z\mathbf{Y}$ on the residual matrix $\mathbf{M}_z\mathbf{X}$.

THE ADJUSTED R^2 AND REGRESSION SELECTION STRATEGIES

The R^2 given by (3.3) has been suggested as a measure of the "completeness" of a theory. For that function it has the defect that if there are as many independent variables as there are observations, the R^2 takes on the value one. Moreover, as variables are added to the equation, the R^2 necessarily increases. This creates an unwarranted preference for models with many explanatory variables, and some adjustment for the number of explanatory variables seems desirable. An analogous problem occurs when estimating σ^2. The maximum likelihood estimator of σ^2 is biased upward, and the unbiased estimator, $s^2=ESS/(T-k)$, is usually preferred.[2] If we think of $1-R^2$ as estimating σ^2/σ_y^2 it is by the same logic natural to define the adjusted R^2 by the equation

$$1-\bar{R}^2\equiv\frac{s^2}{s_y^2}=\frac{ESS/(T-k)}{\Sigma(y_t-\bar{y})^2/T-1}. \tag{3.16}$$

The relationship between R^2 and $\bar{R}^2$ is then

$$1-\bar{R}^2=\frac{(T-1)}{(T-k)}(1-R^2). \tag{3.17}$$

No book on specification searches would be complete without the following two results. The first, due to Theil (1971, p. 543), has been used to justify search strategies that maximize the $\bar{R}^2$. The second defines a simple algorithm for increasing the $\bar{R}^2$.

THEOREM 3.7 (EXPECTED $\bar{R}^2$). *Given two normal regression models, one of which is assumed to be true, the expected value of s^2 for the true model is less than or equal to the expected value of s^2 for the other model.*

To prove this, suppose that $\mathbf{Y}=\mathbf{X}\boldsymbol{\beta}+\mathbf{u}$ is the true model and $\mathbf{Y}=\mathbf{Z}\boldsymbol{\gamma}+\mathbf{u}$ is the alternative, where $\mathbf{X}$ and $\mathbf{Z}$ are $T\times k_x$ and $T\times k_z$ matrices. Then

$$s_x^2=(T-k_x)^{-1}\mathbf{Y}'\mathbf{M}_x\mathbf{Y},\qquad s_z^2=(T-k_z)^{-1}\mathbf{Y}'\mathbf{M}_z\mathbf{Y}$$

where

$$\mathbf{M}_x=\mathbf{I}-\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\quad\text{and}\quad\mathbf{M}_z=\mathbf{I}-\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'.$$

The expected error sum of squares for the false model is

$$\begin{aligned}E(\mathbf{Y}'\mathbf{M}_z\mathbf{Y}|\boldsymbol{\beta})&=E((\mathbf{X}\boldsymbol{\beta}+\mathbf{u})'\mathbf{M}_z(\mathbf{X}\boldsymbol{\beta}+\mathbf{u})|\boldsymbol{\beta})\\&=\boldsymbol{\beta}'\mathbf{X}'\mathbf{M}_z\mathbf{X}\boldsymbol{\beta}+E(\mathbf{u}'\mathbf{M}_z\mathbf{u}|\boldsymbol{\beta})=\boldsymbol{\beta}'\mathbf{X}'\mathbf{M}_z\mathbf{X}\boldsymbol{\beta}+(T-k_z)\sigma^2\\&\geqslant(T-k_z)\sigma^2.\end{aligned}$$

Thus $E(\mathbf{Y}'\mathbf{M}_z\mathbf{Y}|\boldsymbol{\beta})/(T-k_z)\geqslant\sigma^2=E(\mathbf{Y}'\mathbf{M}_x\mathbf{Y}|\boldsymbol{\beta})/(T-k_x)$.

Proof of the next result is left to the reader.

THEOREM 3.8 (INCREASING THE $\bar{R}^2$). *Omitting a variable from an equation will increase the $\bar{R}^2$ if and only if the square of the t-statistic (the F) for that coefficient is less than one.*

[2]It is also preferred by this Bayesian, but not of course because of the bias. See section 7.1.

3.2 Pooling Two Samples

ayesian inference is different from classical inference in that it makes use f information that is not contained in the sample under study. Bayesian ieory is concerned with the optimal pooling of sample information with onsample information. Classical theory can also be used to pool informaon from more than one source, provided the information is generated by xperiments that admit a frequency interpretation.

Suppose, in particular, that a given data set was arbitrarily split into two arts. Then the regression equation could be written as

$$\begin{bmatrix} \mathbf{Y}^* \\ \mathbf{Y} \end{bmatrix} = \begin{bmatrix} \mathbf{X}^* \\ \mathbf{X} \end{bmatrix} \boldsymbol{\beta} + \begin{bmatrix} \mathbf{u}^* \\ \mathbf{u} \end{bmatrix}$$

here the * indicates the first subset of the data. The usual least-squares stimate of $\boldsymbol{\beta}$ would be

$$\mathbf{b}^{**} = \left(\begin{bmatrix} \mathbf{X}^* \\ \mathbf{X} \end{bmatrix}' \begin{bmatrix} \mathbf{X}^* \\ \mathbf{X} \end{bmatrix} \right)^{-1} \begin{bmatrix} \mathbf{X}^* \\ \mathbf{X} \end{bmatrix}' \begin{bmatrix} \mathbf{Y}^* \\ \mathbf{Y} \end{bmatrix} = (\mathbf{X}^{*\prime}\mathbf{X}^* + \mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}^{*\prime}\mathbf{Y}^* + \mathbf{X}'\mathbf{Y})$$

$$= (\mathbf{X}^{*\prime}\mathbf{X}^* + \mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}^{*\prime}\mathbf{X}^*\mathbf{b}^* + \mathbf{X}'\mathbf{X}\mathbf{b}) \tag{3.18}$$

here $\mathbf{b}^* = (\mathbf{X}^{*\prime}\mathbf{X}^*)^{-1}\mathbf{X}^*\mathbf{Y}^*$ and $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$. In words, the least-squares stimate of $\boldsymbol{\beta}$ is a matrix-weighted average of the pair of estimates omputed from each of two subsets of the data, $\mathbf{b}^*$ and $\mathbf{b}$.

Another possibility is that the variance of the residuals in the first part f the data set is different from the variance in the second. Indicating these ariances by σ^{*2} and σ^2, respectively, the model may be transformed to ake the residual covariance matrix equal the identity matrix by dividing ' and $\mathbf{X}$ by σ and $\mathbf{Y}^*$ and $\mathbf{X}^*$ by σ^*. The resulting estimate of $\boldsymbol{\beta}$ is

$$\mathbf{b}^{**} = (\sigma^{*-2}\mathbf{X}^{*\prime}\mathbf{X}^* + \sigma^{-2}\mathbf{X}'\mathbf{X})^{-1}(\sigma^{*-2}\mathbf{X}^{*\prime}\mathbf{X}^*\mathbf{b}^* + \sigma^{-2}\mathbf{X}'\mathbf{X}\mathbf{b}). \tag{3.19}$$

The notation just used may seem confusing, but it is chosen to anticipate he Bayesian analysis of the next section, in which personal prior opinions nd up being equivalent to the preliminary observations $\mathbf{Y}^*, \mathbf{X}^*$. The prior nformation is pooled with the current information exactly in accordance vith formula (3.19).[3]

The variances σ^2 and σ^{*2} may also be estimated. The likelihood function mplied by the model is

$$L(\boldsymbol{\beta}, \sigma^2, \sigma^{*2}; \mathbf{Y}, \mathbf{Y}^*, \mathbf{X}, \mathbf{X}^*) \propto (\sigma^2)^{-T/2} \exp\left[-\frac{1}{2\sigma^2} (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) \right]$$

$$(\sigma^{*2})^{-T^*/2} \exp\left[-\frac{1}{2\sigma^2} (\mathbf{Y}^* - \mathbf{X}^*\boldsymbol{\beta})'(\mathbf{Y}^* - \mathbf{X}^*\boldsymbol{\beta}) \right].$$

[3] A schizophrenic attitude toward probabilities allows Theil and Goldberger (1961) to ntroduce prior information into classical inference in this way. An earlier paper of the same enre is Durbin (1953).

The reader may verify that maximizing this likelihood function implies choosing $\boldsymbol{\beta}$ in accord with Equation (3.19) and choosing the variances as

$$\hat{\sigma}^2 = \frac{(\mathbf{Y} - \mathbf{X}\mathbf{b}^{**})'(\mathbf{Y} - \mathbf{X}\mathbf{b}^{**})}{T} \tag{3.20}$$

$$\hat{\sigma}^{*2} = \frac{(\mathbf{Y}^* - \mathbf{X}^*\mathbf{b}^{**})'(\mathbf{Y}^* - \mathbf{X}^*\mathbf{b}^{**})}{T^*}. \tag{3.21}$$

3.3 Bayesian Inference with the Linear-Regression Model

Bayesian inference with the linear-regression model is now introduced and shown to be formally equivalent to the pooling problem just discussed. Bayes' rule applied to the uncertain parameters $(\boldsymbol{\beta}, \sigma^2)$ is

$$f(\boldsymbol{\beta}, \sigma^2 | \mathbf{Y}, \mathbf{X}) = \frac{f(\mathbf{Y}, \mathbf{X} | \boldsymbol{\beta}, \sigma^2) f(\boldsymbol{\beta}, \sigma^2)}{f(\mathbf{Y}, \mathbf{X})},$$

where $\mathbf{Y}$ and $\mathbf{X}$ are observable data. The linear-regression model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$, where $\mathbf{u}$ is normally distributed with mean $\mathbf{0}$ and variance $\sigma^2\mathbf{I}$, implies a conditional distribution for $\mathbf{Y}$: $f_N(\mathbf{Y} | \mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$, indicating that $\mathbf{Y}$ is normally distributed with mean $\mathbf{X}\boldsymbol{\beta}$ and variance $\sigma^2\mathbf{I}$. If, furthermore, $\mathbf{X}$ is distributed independently of $\boldsymbol{\beta}$ and σ^2, the posterior distribution can be written as

$$f(\boldsymbol{\beta}, \sigma^2 | \mathbf{Y}, \mathbf{X}) = \frac{f_N(\mathbf{Y} | \mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}) f(\mathbf{X}) f(\boldsymbol{\beta}, \sigma^2)}{f(\mathbf{Y} | \mathbf{X}) f(\mathbf{X})}$$

$$= \frac{f_N(\mathbf{Y} | \mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}) f(\boldsymbol{\beta}, \sigma^2)}{f(\mathbf{Y} | \mathbf{X})}.$$

Before discussing the choice of prior, $f(\boldsymbol{\beta}, \sigma^2)$, a few things may be said about the immediately preceding formula. It implies that randomness in the matrix $\mathbf{X}$ is irrelevant for inference if $\mathbf{X}$ is distributed independently of $\boldsymbol{\beta}$ and σ^2. Classical inference, in contrast, cannot easily ignore the randomness in $\mathbf{X}$, since sampling properties are necessarily affected. Another important point is that both the vectors of unobservables, $\boldsymbol{\beta}$ and $\mathbf{u}$, are assigned personal probability distributions. Classically, the residual vector $\mathbf{u}$ is thought to have a frequency distribution and the parameter vector $\boldsymbol{\beta}$ is thought to be a fixed vector of constants. But from the Bayesian point of view, your personal opinion about both $\boldsymbol{\beta}$ and $\mathbf{u}$ is described in probabilistic terms, and the extreme distinction between $\boldsymbol{\beta}$ and $\mathbf{u}$ is regarded as unwarranted.

Having assumed that the residual vector is normally distributed, only the choice of a prior distribution for $(\boldsymbol{\beta}, \sigma^2)$ remains. Two alternatives are suggested here. The resulting posterior distributions are described in the following pair of theorems, and further discussion follows. Both results

ake use of the assumption that **Y** given **X**, $\boldsymbol{\beta}$ and σ^2 is normal with mean $\boldsymbol{\beta}$ and variance $\sigma^2\mathbf{I}$.

ONJUGATE PRIOR

THEOREM 3.9. *Given the normal linear-regression model and a normal-gamma prior distribution for* $(\boldsymbol{\beta},\sigma^{-2})$:

$$f(\boldsymbol{\beta},\sigma^{-2})=f_N\left(\boldsymbol{\beta}|\mathbf{b}^*,\sigma^2(\mathbf{N}^*)^{-1}\right)f_\gamma(\sigma^{-2}|s^{*2},\nu^*),$$

which indicates that $\boldsymbol{\beta}$ *given* σ^2 *is distributed normally with mean* $\mathbf{b}^*$ *and variance* $\sigma^2(\mathbf{N}^*)^{-1}$ *and that* σ^{-2} *is distributed gamma with parameters* s^{*2} *and* ν^*, *then the posterior distribution of* $(\boldsymbol{\beta},\sigma^{-2})$ *is in the normal-gamma family with parameters*

$$\mathbf{b}^{**}=(\mathbf{N}^*+\mathbf{X}'\mathbf{X})^{-1}(\mathbf{N}^*\mathbf{b}^*+\mathbf{X}'\mathbf{X}\mathbf{b}) \tag{3.22}$$

$$\mathbf{N}^{**}=\mathbf{N}^*+\mathbf{X}'\mathbf{X} \tag{3.23}$$

$$\nu^{**}=\nu^*+T \tag{3.24}$$

$$(s^{**})^2=\left[\nu^{**}\right]^{-1}\left[\nu^*s^{*2}+ESS+(\mathbf{b}-\mathbf{b}^*)'\mathbf{N}^*(\mathbf{X}'\mathbf{X}+\mathbf{N}^*)^{-1}\mathbf{X}'\mathbf{X}(\mathbf{b}-\mathbf{b}^*)\right]. \tag{3.25}$$

The marginal posterior distribution of $\boldsymbol{\beta}$ *is multivariate Student with parameters* $\mathbf{b}^{**}$, $(s^{**})^2(\mathbf{N}^{**})^{-1}$, *and* ν^{**}.

roof: Let $h=\sigma^{-2}$ and use the likelihood function (3.6) together with the ior to obtain

$$f(\boldsymbol{\beta},h|\mathbf{Y},\mathbf{X})\propto h^{T/2}\exp\left[-\tfrac{1}{2}hESS\right]\exp\left[-\tfrac{1}{2}h(\boldsymbol{\beta}-\mathbf{b})'\mathbf{X}'\mathbf{X}(\boldsymbol{\beta}-\mathbf{b})\right]$$
$$h^{k/2}\exp\left[-\tfrac{1}{2}h(\boldsymbol{\beta}-\mathbf{b}^*)\mathbf{N}^*(\boldsymbol{\beta}-\mathbf{b}^*)\right]h^{\frac{\nu^*}{2}-1}\exp\left[-\tfrac{1}{2}\nu^*s^{*2}h\right].$$

sing result (T10) in Appendix 1, this can be rewritten as

$$f(\boldsymbol{\beta},h|\mathbf{Y},\mathbf{X})\propto h^{k/2}\exp\left[-\tfrac{1}{2}h(\boldsymbol{\beta}-\mathbf{b}^{**})'\mathbf{N}^{**}(\boldsymbol{\beta}-\mathbf{b}^{**})\right]$$
$$h^{\frac{\nu^{**}}{2}-1}\exp\left[-\tfrac{1}{2}\nu^{**}s^{**2}h\right],$$

hich by inspection is the normal-gamma distribution described in the eorem above.

The normal-gamma prior is Raiffa and Schlaifer's (1961) "natural con-gate" prior. The member of this class of distributions that is sometimes ed to represent prior ignorance is

$$f(\boldsymbol{\beta},\sigma^{-2})\propto(\sigma^{-2})^{k/2}$$

ith $\mathbf{N}^*=\mathbf{0}$ and $\nu^*=0$. Given these prior parameters, the posterior parame-

ters become $\mathbf{b}$, $\mathbf{X}'\mathbf{X}$, T, and ESS/T. The marginal posterior distribution of $\boldsymbol{\beta}$, therefore, reproduces the classical result, except that the degrees of freedom parameter is T instead of $T-k$, a point further discussed below.

Observe also that the posterior mean (3.22) is formally the same as Equation (3.18), which describes the pooling of two samples of the same regression process. Thus the normal gamma prior is equivalent to a previous sample of the same process with estimate equal to $\mathbf{b}^*$, "$\mathbf{X}'\mathbf{X}$" matrix equal to $\mathbf{N}^*$, degrees of freedom equal to ν^*, and error sum of squares equal to ν^*s^{*2}.

The posterior distribution for $\boldsymbol{\beta}$ associated with this normal-gamma prior has the dual features that it is unimodal and located at a fixed, weighted average of the sample location $\mathbf{b}$ and the prior location $\mathbf{b}^*$. In that sense it never distinguishes sample information from prior information, no matter how strong their apparent conflict. This is so because a conjugate prior treats prior information as if it were a previous sample of the same process. It may be argued that most prior information is distinctly different from sample information, and when they are apparently in conflict, the posterior distribution ought to be multimodal with modes at both the sample location and the prior location. A distribution that has this feature is described in the following theorem, which uses a prior suggested by Dickey (1975).

STUDENT PRIOR

THEOREM 3.10 *If* $\boldsymbol{\beta}$ *has a multivariate Student prior distribution independent of* σ^2, *and if* σ^{-2} *has a gamma prior distribution*

$$f(\boldsymbol{\beta},\sigma^{-2})=f_S^k\left(\boldsymbol{\beta}|\mathbf{b}^*,\mathbf{H}^{*-1},\nu_\beta^*\right)f_\gamma\left(\sigma^{-2}|s^{*2},\nu_\sigma^*\right),$$

then the marginal posterior distribution for $\boldsymbol{\beta}$ *is proportional to the product of two Student functions*[4]

$$f(\boldsymbol{\beta}|\mathbf{Y})\propto f_S^k\left(\boldsymbol{\beta}|\mathbf{b},\mathbf{H}^{-1},\nu_\sigma^{**}\right)f_S^k\left(\boldsymbol{\beta}|\mathbf{b},\mathbf{H}^{*-1},\nu_\beta^*\right)$$

where $\mathbf{b}$ *is the least-squares estimate, and*

$$\mathbf{H}=(s^{**})^{-2}\mathbf{X}'\mathbf{X}$$
$$\nu_\sigma^{**}=\nu_\sigma^*+T-k$$
$$(s^{**})^2=(\nu_\sigma^{**})^{-1}\left(\nu_\sigma^*s^{*2}+ESS\right).$$

[4]The product of two Student distributions can be written as a one-dimensional mixture of Student distributions (Dickey, 1975).

of: Integrate σ^2 from the product of the likelihood function and prior:

$$\begin{aligned}\boldsymbol{\beta}|\mathbf{Y} &\propto \int_{\sigma^{-2}} f(\mathbf{Y}|\boldsymbol{\beta},\sigma^2) f(\boldsymbol{\beta},\sigma^{-2}) d\sigma^{-2} \\ &= \int_{\sigma^{-2}} f_N(\mathbf{Y}|\mathbf{X}\boldsymbol{\beta},\sigma^2 I) f_\gamma(\sigma^{-2}|s^{*2},\nu_\sigma^*) f_S(\boldsymbol{\beta}|\mathbf{b}^*,\mathbf{H}^{*-1},\nu_\beta^*) d\sigma^{-2} \\ &= f_S^T(\mathbf{Y}|\mathbf{X}\boldsymbol{\beta},s^{*2}\mathbf{I},\nu_\sigma^*) f_S(\boldsymbol{\beta}|\mathbf{b}^*,\mathbf{H}^{*-1},\nu_\beta^*).\end{aligned}$$

e Student distribution in $\mathbf{Y}$ can be written as a Student function in $\boldsymbol{\beta}$ as

$$\begin{aligned}f_S^T(\mathbf{Y}|\mathbf{X}\boldsymbol{\beta},s^{*2}\mathbf{I},\nu_\sigma^*) &\propto [\nu_\sigma^* s^{*2} + (\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})]^{-(\nu_\sigma^*+T)/2} \\ &\propto [\nu_\sigma^* s^{*2} + ESS + (\boldsymbol{\beta}-\mathbf{b})'\mathbf{X}'\mathbf{X}(\boldsymbol{\beta}-\mathbf{b})]^{-(\nu^{**}+k)/2} \\ &\propto f_S^k(\boldsymbol{\beta}|\mathbf{b},s^{**2}(\mathbf{X}'\mathbf{X})^{-1},\nu_\sigma^{**})\end{aligned}$$

h ν^{**} and s^{**}defined as before.

The difference between the prior of Theorem (3.9) and the prior of eorem (3.10) is in the latter case $\boldsymbol{\beta}$ and σ^2 are independent. The former ijugate prior implies that if information is obtained about σ^2, opinions out $\boldsymbol{\beta}$ change. Either $\boldsymbol{\beta}$ becomes more uncertain or more certain. Of irse, your priors are your own business, but I can say that I tend to fer to have $\boldsymbol{\beta}$ and σ^2 independent. A counterargument is that if you cover that the process is noisy (σ^2 large), you may come to doubt the idity of your prior information.[5]

As pointed out earlier, the conjugate prior cannot reproduce classical st squares because it does not allow loss of degrees of freedom. The ident-gamma prior with $\mathbf{H}^*=0$ and $\nu_\sigma^*=0$ does exactly reproduce classi- least squares with

$$t=\frac{\beta_i-b_i}{s\left[(\mathbf{X}'\mathbf{X})^{-1}\right]_{ii}}$$

en $\mathbf{Y}$, having a t distribution with $T-k$ degrees of freedom.

The natural conjugate prior was shown to be equivalent to a previous nple of the same process. The Student prior for $\boldsymbol{\beta}$, together with the fuse prior for $\sigma^2(\nu_\sigma^*=0)$, is equivalent to a previous sample of a regres- n process with a different variance. Bayesian pooling of two such nples without any other information would lead to the "double-Student" sterior of Theorem 3.10. Also, the modes of the posterior distribution

[5]To anticipate Chapter 9, it may make sense to write the regression process as $\mathbf{Y}=\mathbf{X}(\boldsymbol{\beta}+$ $)+\mathbf{u}$ where $\boldsymbol{\beta}^c$ reflects the specification error. Although your prior for the true parameter $\boldsymbol{\beta}$ y be independent of σ^2, your prior for $\boldsymbol{\beta}^c$ may not be.

may be found by setting its logarithmic derivatives to zero

$$\mathbf{0}=\partial \ln f(\boldsymbol{\beta}|\mathbf{Y})/\partial\boldsymbol{\beta}=\lambda\mathbf{X}'\mathbf{X}(\boldsymbol{\beta}-\mathbf{b})+\lambda^*\mathbf{H}^*(\boldsymbol{\beta}-\mathbf{b}^*)$$

where

$$\lambda=\frac{\nu_\sigma^*+T}{\nu_\sigma^* s^{*2}+ESS+(\boldsymbol{\beta}-\mathbf{b})'\mathbf{X}'\mathbf{X}(\boldsymbol{\beta}-\mathbf{b})} \tag{3.26}$$

$$\lambda^*=\frac{\nu_\beta^*+k}{\nu_\beta^*+(\boldsymbol{\beta}-\mathbf{b}^*)'\mathbf{H}^*(\boldsymbol{\beta}-\mathbf{b}^*)}. \tag{3.27}$$

With a suitable choice of ν_σ^*, ν_β^*, and $\mathbf{H}^*$, these equations are equivalent to (3.19), (3.20), and (3.21) which describe pooling information from two different processes.

A GRAPHICAL PRESENTATION

A graphical analysis of inference with the linear regression model is instructive. The graphs now to be discussed apply to the results just reported, but they also apply to a wider class of distributions. We need only assume that $\mathbf{A}\boldsymbol{\beta}$ and $\mathbf{u}$ have independent, spherically symmetric distributions.[6]

A random variable $\mathbf{z}$ is said to have a spherically symmetric distribution if its density function depends only on the length of $\mathbf{z}$: $f(\mathbf{z})=cg(\mathbf{z}'\mathbf{z})$. Thus the assumption of spherical symmetry implies that the densities of $\mathbf{Y}$ and $\boldsymbol{\beta}$ may be written as

$$f(\mathbf{Y}|\boldsymbol{\beta}) \propto g_u[(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})]$$
$$f(\boldsymbol{\beta}) \propto g_\beta[(\boldsymbol{\beta}-\mathbf{b}^*)'\mathbf{A}'\mathbf{A}(\boldsymbol{\beta}-\mathbf{b}^*)].$$

Using (3.5) and setting $\mathbf{A}'\mathbf{A}=\mathbf{N}^*$, the posterior distribution can be written as

$$\begin{aligned}f(\boldsymbol{\beta}|\mathbf{Y}) &\propto f(\mathbf{Y}|\boldsymbol{\beta}) f(\boldsymbol{\beta}) \\ &\propto g_u[ESS+(\boldsymbol{\beta}-\mathbf{b})'\mathbf{X}'\mathbf{X}(\boldsymbol{\beta}-\mathbf{b})]\, g_\beta[(\boldsymbol{\beta}-\mathbf{b}^*)'\mathbf{N}^*(\boldsymbol{\beta}-\mathbf{b}^*)].\end{aligned}$$

The full posterior distribution, of course, depends on the functions g_u and g_β. The mode, however, necessarily lies on a curve. To find the equation for this curve, we may differentiate the logarithm of the posterior

$$\frac{\partial \ln f(\boldsymbol{\beta}|\mathbf{Y})}{\partial\boldsymbol{\beta}}=g_u^{-1}g_u' 2\mathbf{N}(\boldsymbol{\beta}-\mathbf{b})+g_\beta^{-1}g_\beta'(2\mathbf{N}^*(\boldsymbol{\beta}-\mathbf{b}^*)). \quad (\mathbf{N}=\mathbf{X}'\mathbf{X})$$

[6]Hill (1969).

Setting this equal to zero and solving for $\boldsymbol{\beta}$ yields

$$\boldsymbol{\beta}=(\mathbf{N}+\lambda\mathbf{N}^*)^{-1}(\mathbf{N}\mathbf{b}+\lambda\mathbf{N}^*\mathbf{b}^*) \tag{3.28}$$

where

$$\lambda=\frac{g_\beta^{-1}g_\beta'}{g_u^{-1}g_u'}. \tag{3.29}$$

As λ varies from zero to infinity, Equation (3.28) sweeps out a curve in k-space affectionately called by Dickey (1975) the *curve décolletage*. It is anchored by the least-squares point **b** at one end and by the "prior point" $\mathbf{b}^*$ at the other. Choice of a point along the curve depends on the functions g_u and g_β, but spherical symmetry is sufficient to imply the curve.

Geometrically, the curve décolletage is the locus of tangencies between the sample family of ellipsoids, $(\boldsymbol{\beta}-\mathbf{b})'\mathbf{N}(\boldsymbol{\beta}-\mathbf{b})=c$, and the prior family $(\boldsymbol{\beta}-\mathbf{b}^*)'\mathbf{N}^*(\boldsymbol{\beta}-\mathbf{b}^*)=c^*$. To find this locus of tangencies, set the derivatives of $(\boldsymbol{\beta}-\mathbf{b})'\mathbf{N}(\boldsymbol{\beta}-\mathbf{b})+\lambda(\boldsymbol{\beta}-\mathbf{b}^*)'\mathbf{N}^*(\boldsymbol{\beta}-\mathbf{b}^*)$ to zero: $2\mathbf{N}(\boldsymbol{\beta}-\mathbf{b})+2\lambda\mathbf{N}^*(\boldsymbol{\beta}-\mathbf{b}^*)=\mathbf{0}$, which is just Equation (3.28).

A curve décolletage is illustrated in Figure 3.1. Ellipses around the least-squares point **b** are isolikelihood ellipses. The data prefer the point **b**, and the data are indifferent between any points on a given likelihood ellipse. The most preferred point from the standpoint of the prior is $\mathbf{b}^*$, and an ellipse around the prior point is a prior isodensity ellipse. The curve décolletage contains all points jointly preferred by data and prior. Given any point not on the curve, there is a better point on the curve, in the sense that neither the likelihood value nor the prior density is less, and at least one is greater.[7]

The location of modes on the curve décolletage depends on the functions g_u and g_β. One possibility is the exponential family indexed by the precision parameter h:

$$g(z^2|h)=e^{-\frac{1}{2}hz^2}$$

with

$$g'g^{-1}=-\frac{h}{2}.$$

Thus letting the data and prior densities have different precision parameters, $g_u(z^2)=g(z^2|h)$ and $g_\beta(z^2)=g(z^2|h^*)$, we would have, from (3.29), $\lambda=h^*/h$; the posterior mode, since λ is independent of $\boldsymbol{\beta}$, is just (3.28), using this value of λ.

[7]Economists are reminded of the analogous Edgeworth–Bowley diagram. This analogy is pursued in Chapter 5, and the curve décolletage is there referred to as the information contract curve.

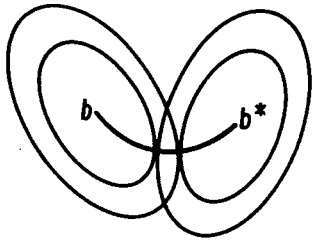

Fig. 3.1 Curve décolletage.

Describing this more traditionally, we have just assumed that the data and prior densities are normal $\mathbf{u}\sim N(0,h^{-1}(\mathbf{B}'\mathbf{B})^{-1})$, $\boldsymbol{\beta}\sim N(\mathbf{b}^*,h^{*-1}\mathbf{N}^{*-1})$, with h/h^* known. This same result has been described in Theorem 3.9.

A somewhat less restrictive class of functions for labeling the information indifference curves is the Student family

$$g_S(z^2|a,\nu)=(a+z^2)^{-\nu/2} \qquad a,\nu>0$$

with

$$g_S'g_S^{-1}=-\frac{\nu}{2(a+z^2)}.$$

Letting $g_u(z^2)=g_S(z^2|a,\nu)$ and $g_\beta(z^2)=g_S(z^2|a^*,\nu^*)$ we have from (3.29)

$$\lambda=\frac{\nu^*/[a^*+(\boldsymbol{\beta}-\mathbf{b}^*)'\mathbf{N}(\boldsymbol{\beta}-\mathbf{b}^*)]}{\nu/[a+ESS+(\boldsymbol{\beta}-\mathbf{b})'\mathbf{N}(\boldsymbol{\beta}-\mathbf{b})]}.$$

This family of labeling distributions has the property that the logarithmic derivative $g'g^{-1}$ decreases with z^2. It is thus relatively steep around the origin and relatively flat elsewhere as indicated in Figure 3.2. This means

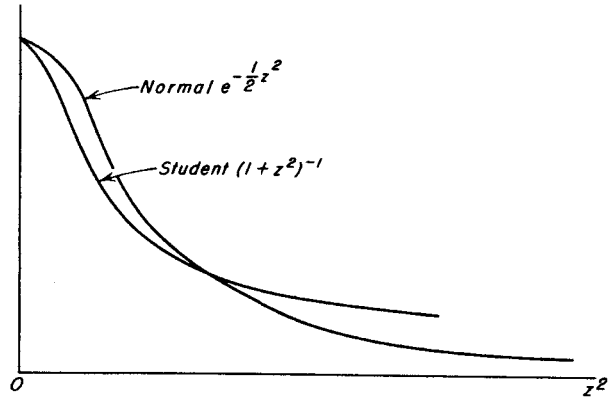

Fig. 3.2 Normal and Student labeling functions.

hat an information source will be relatively resistant to moving the stimate of β away from its most preferred point in the neighborhood of hat point and relatively indifferent to adjustments that occur farther from hat point. Note also that the logarithmic derivative of the likelihood unction depends not only on a and ν but on ESS and β as well. The ocation of modes on the curve décolletage is, therefore, data dependent.

These assumptions are described more traditionally in Theorem 3.10.

SWEEPING OUT THE MEANS

Ordinarily, there is a constant in a regression function about which we know very little relative to the prior information on the slopes. We show here that the effect of this structure of information is merely to sweep out the means of the observed variables.

Writing the constant explicitly, the regression process becomes

$$\mathbf{Y}=\mathbf{X}\beta+\mathbf{1}_T\alpha+\mathbf{u}$$

with definitions as before, but with $\mathbf{1}_T$ a vector of ones and α the scalar "process level." Let $\mathbf{u}$ be normal with mean zero and covariance matrix $\sigma^2\mathbf{I}_T$, and let α be normal with mean zero and variance v. We may then write

$$\mathbf{Y}=\mathbf{X}\beta+\varepsilon$$

where ε is normal with mean zero and covariance matrix $\sigma^2\mathbf{I}_T+\mathbf{1}_T v\mathbf{1}_T'$. The precision matrix for ε is

$$(\sigma^2\mathbf{I}_T+\mathbf{1}_T v\mathbf{1}_T')^{-1}=\sigma^{-2}\left(\mathbf{I}_T-\mathbf{1}_T\left[\mathbf{1}_T'\mathbf{1}_T+\frac{\sigma^2}{v}\right]^{-1}\mathbf{1}_T'\right).$$

In the limit as v increases the conditional distribution of $\mathbf{Y}$ given β becomes improper with singular precision matrix

$$\sigma^{-2}\mathbf{M}=\sigma^{-2}\left(\mathbf{I}_T-\mathbf{1}_T(\mathbf{1}_T'\mathbf{1}_T)^{-1}\mathbf{1}_T'\right).$$

The determinant of the variance matrix of $\mathbf{Y}$, using (T18) in Appendix 1, is $(\sigma^2)^T(1+Tv\sigma^{-2})$, which for large v behaves like $(\sigma^2)^{T-1}v$. Thus the limiting likelihood function is

$$f(\mathbf{Y}|\beta)\propto(\sigma^2)^{-(T-1)/2}\exp\left[-\frac{1}{2\sigma^2}(\mathbf{Y}-\mathbf{X}\beta)'\mathbf{M}(\mathbf{Y}-\mathbf{X}\beta)\right].$$

But $\mathbf{M}$ is the idempotent matrix that was shown in Section 3.1 to sweep out the means of the variables. That is, given that the constant α is diffuse and independent of the other random variables, inferences about the slope β may proceed by first subtracting out the means of the observables $\mathbf{Y}$ and $\mathbf{X}$ and then proceeding as above.

3.4 Multivariate Normal Sampling

In this section, Bayesian inference about the parameters of a multivariate normal distribution is discussed. (Several results here are referred to occasionally, but it is not necessary for the reader at this point to master this material.) The $(k\times 1)$ vector $\mathbf{z}_t$ is to be normally distributed with mean vector μ and variance matrix Σ. A sample of T independently distributed vectors $\mathbf{z}_t$ implies the likelihood function

$$L(\mu,\Sigma;\mathbf{z}_1,\mathbf{z}_2,\ldots,\mathbf{z}_T)\propto|\Sigma|^{-T/2}\exp\left[-\frac{1}{2}\sum_t(\mathbf{z}_t-\mu)'\Sigma^{-1}(\mathbf{z}_t-\mu)\right].$$

Letting $\bar{\mathbf{z}}=\Sigma_t\mathbf{z}_t/T$

$$\mathbf{S}=\sum_t(\mathbf{z}_t-\bar{\mathbf{z}})(\mathbf{z}_t-\bar{\mathbf{z}})',$$

the exponent in the likelihood function can be written as

$$\sum_t(\mathbf{z}_t-\mu)'\Sigma^{-1}(\mathbf{z}_t-\mu)=\sum_t\left[(\mathbf{z}_t-\bar{\mathbf{z}})'\Sigma^{-1}(\mathbf{z}_t-\bar{\mathbf{z}})+(\bar{\mathbf{z}}-\mu)'\Sigma^{-1}(\bar{\mathbf{z}}-\mu)\right]$$

$$=\sum_t\operatorname{tr}\left[\Sigma^{-1}(\mathbf{z}_t-\bar{\mathbf{z}})(\mathbf{z}_t-\bar{\mathbf{z}})'\right]+T(\bar{\mathbf{z}}-\mu)'\Sigma^{-1}(\bar{\mathbf{z}}-\mu)$$

$$=\operatorname{tr}(\Sigma^{-1}\mathbf{S})+T(\bar{\mathbf{z}}-\mu)'\Sigma^{-1}(\bar{\mathbf{z}}-\mu).$$

A normal-Wishart distribution for (μ,Σ^{-1}) is a convenient prior:

$$f_{NW}(\mu,\Sigma^{-1}|\bar{\mathbf{z}}^*,\mathbf{S}^*,T^*,\nu^*)\propto|\Sigma^{-1}|^{1/2}\exp\left[-\frac{T^*}{2}(\mu-\bar{\mathbf{z}}^*)'\Sigma^{-1}(\mu-\bar{\mathbf{z}}^*)\right]$$

$$|\Sigma^{-1}|^{(\nu^*-k-1)/2}\exp\left[-\frac{1}{2}\operatorname{tr}\Sigma^{-1}\mathbf{S}^*\right].$$

Combining this prior with the likelihood function and using (T10) in Appendix 1 the posterior distribution is seen to be in the normal-Wishart family with parameters

$$\bar{\mathbf{z}}^{**}=(T+T^*)^{-1}(T\bar{\mathbf{z}}+T^*\bar{\mathbf{z}}^*)$$

$$\mathbf{S}^{**}=\mathbf{S}+\mathbf{S}^*+(\bar{\mathbf{z}}-\bar{\mathbf{z}}^*)'\Sigma^{-1}(\bar{\mathbf{z}}-\bar{\mathbf{z}}^*)TT^*(T+T^*)^{-1}$$

$$T^{**}=T+T^*$$

$$\nu^{**}=T+\nu^*.$$

The predictive distribution of the next sample vector, $\mathbf{z}_F$, is used in Chapter 6. Conditional on μ and Σ, $\mathbf{z}_F$ is normal with mean vector μ and covariance matrix Σ. Conditional on Σ and the data $(\bar{\mathbf{z}},\mathbf{S},\mathbf{T})$, μ is normal with mean $\bar{\mathbf{z}}^{**}$ and covariance matrix Σ/T^{**}. Thus integrating with

espect to μ, the vector $\mathbf{z}_F$ is normal with mean $\bar{\mathbf{z}}^{**}$ and covariance matrix $\Sigma(1+(T^{**})^{-1})$. Integrating with respect to Σ^{-1} as in Appendix 2, the listribution for $\mathbf{z}_F$ is Student

$$f(\mathbf{z}_F|\bar{\mathbf{z}},\mathbf{S},T)=f_S^k\left(\mathbf{z}_F|\bar{\mathbf{z}}^{**},\mathbf{S}^{**}\left(1+(T^{**})^{-1}\right)/\nu^{**},\nu^{**}-k+1\right).$$

The results needed for Chapter 6 are the conditional moments of $\mathbf{z}_F^I$ given $\mathbf{z}_F^J$ where the partition of $\mathbf{z}_F$ is $\mathbf{z}_F'=(\mathbf{z}_F^I,\mathbf{z}_F^J)$. Using the diffuse prior assumption that $\mathbf{S}^*=0$, $T^*=0$, $\nu^*=0$, the conditional moments are

$$E\left(\mathbf{z}_F^I|\mathbf{z}_F^J\right)=\bar{\mathbf{z}}^I+\mathbf{S}_{IJ}\mathbf{S}_{JJ}^{-1}\left(\mathbf{z}_F^J-\bar{\mathbf{z}}^J\right)$$

$$V\left(\mathbf{z}_F^I|\mathbf{z}_F^J\right)=\left(\mathbf{S}_{II}-\mathbf{S}_{IJ}\mathbf{S}_{JJ}^{-1}\mathbf{S}_{JI}\right)(1+T^{-1})(T-k+1+k_J)^{-1}$$

where k_J is the number of elements in $\mathbf{z}_F^J$.

CHAPTER **4**

HYPOTHESES-TESTING SEARCHES

The first variety of specification search that we discuss corresponds to the familiar hypothesis-testing problem. We assume the existence of a set of M "models" or hypotheses of the form

$$H_i:\mathbf{Y}=\mathbf{X}_i\boldsymbol{\beta}_i+\mathbf{u}_i,\qquad i=1,\ldots,M,\qquad(4.1)$$

where $\mathbf{Y}$ is a $(T\times 1)$ vector of observable variables, $\mathbf{X}_i$ is a $(T\times k_i)$ matrix of observable explanatory variables, $\boldsymbol{\beta}_i$ is a $(k_i\times 1)$ vector of parameters, and $\mathbf{u}_i$ is a $(T\times 1)$ vector of unobservable disturbances assumed to be normally distributed with mean zero and variance-covariance matrix $\sigma_i^2\mathbf{I}_T$. The statistical problem is to determine which of these M models did, in fact, generate the data and at the same time to make inferences about the coefficient vectors $\boldsymbol{\beta}_i$.

The formal classical theory of hypothesis testing describes the decision problem of selecting an action from among a set of feasible actions. An action is either right or wrong, depending on the "true state of nature," and the statistician is interested in being wrong as infrequently as possible. The problem being considered in this chapter involves a set of M actions of the form "act as if hypothesis H_i were true" and a set of M states of nature of the form "hypothesis H_i is true." An error occurs when action i is chosen but hypothesis j $(i\neq j)$ is true.