ON THE USE OF INCOMPLETE PRIOR INFORMATION IN REGRESSION ANALYSIS

H. THEIL*
Netherlands School of Economics

This article deals with the use of prior beliefs in the estimation of regression coefficients; in particular, it considers the problems that arise when the residual variance of the regression equation is unknown and it offers a large-sample solution. Additional contributions deal with testing the hypothesis that prior and sample information are compatible with each other; and with a scalar measure for the shares of these two kinds of information in the posterior precision.

1. INTRODUCTION

IN A recent article [7] the present author and A. S. Goldberger developed the method of "mixed" estimation, which is an effort to incorporate prior knowledge of coefficients in regression analysis and other linear statistical models. This prior knowledge was formulated in terms of prior estimates of the parameters which are assumed to be unbiased and to have a given moment matrix. At about the same time, but independently, Raiffa and Schlaifer published their monograph on Bayesian applications [4], which contains a chapter on the use of prior distributions in regression analysis. Under the usual assumption of quadratic loss functions the two approaches lead to identical coefficient estimates as long as it assumed that the disturbances of the regression model have the same *known* variance of σ^2. But in most applications this assumption is not realistic. To handle this complication, an iterative procedure was suggested in reference [7], whereas reference [4] proceeds on the basis of a prior distribution of σ^2. The latter approach has certainly some intuitive appeal, but its application is hampered by two difficulties. One is that it will be extremely difficult in many econometric applications to make sensible a priori statements about the magnitude of σ^2; the other is that the second-order moments of the prior distribution of the coefficients are supposed not to be measured in the "natural" units but in terms of the true σ^2. For example, if we have a consumption function of the classical form $C=\alpha+\beta Y+u$,[1] then it is not sufficient to say that we have a prior distribution for β which is normal (say) with mean 0.8 and standard deviation 0.03. Instead, we should formulate the prior distribution in the following terms: the mean is 0.8 and the standard deviation is $k\sigma$, where σ is the standard deviation of the disturbances while k has to be specified numerically (e.g., 0.02 or 345, the dimension being reciprocal to money units per time unit). Clearly, it is far from easy to formulate a satisfactory k-specification.

The present paper is more classical than Bayesian, although its subject is essentially the same: the combination of prior and sample information, both being stochastic but independent of each other. Its purpose is threefold. First, we shall present (in Section 2) a method which handles the complication of an unknown σ^2, at least asymptotically. No prior distribution for this parameter

* The author is indebted to Mr. Thomas Rothenberg of the Econometric Institute for his comments on an earlier draft of this paper as well as to the referees for various clarifications on Bayesian estimation.

[1] C = consumption, Y = income, u = disturbance, α and β are parameters to be estimated.

will be used, the argument being that there is a considerable class of cases in which nothing is known about σ^2 except that it is positive. In fact, we shall take account of the possibility that there are also other parameters about which nothing is known before the data are examined; this will, e.g., frequently apply to the constant term of the regression. The second purpose of the paper is to establish a procedure for testing the compatibility of sample and prior information; this is carried out in Section 3. A numerical example is presented in Section 4. The third purpose, finally, is to propose a measure for the relative contributions of sample and prior information to our posterior knowledge, which will be done in Section 5.

2. MIXED REGRESSION ESTIMATION FOR UNKNOWN σ^2

2.1. The Case of Known σ^2

Our sample information will be supposed to consist of the $T\times(\Lambda+1)$ matrix $[y\ X]$, where y is interpreted as a vector of values taken by a dependent variable and X as a matrix of values taken by Λ independent variables. They are connected by

$$y = X\beta + u, \tag{2.1}$$

where β is a Λ-element parameter vector and u a disturbance vector. The objective of the analysis is to estimate β. The following assumptions will be made throughout this paper: (i) The matrix X has rank Λ and consists of nonstochastic elements. (ii) The disturbances have zero mean and constant (finite) variance and are uncorrelated:

$$Eu = 0; \qquad E(uu') = \sigma^2 I, \tag{2.2}$$

where I is the unit matrix of order T.

In addition to sample information we have prior information. It is assumed that it consists of a k-element vector r which estimates $R\beta$, the latter vector consisting of k linear combinations of the β-components. Hence we can write

$$r = R\beta + v, \tag{2.3}$$

where v is the error of the prior information. For example, if $R=[I\ O]$ where I is the $k\times k$ unit matrix and 0 the $k\times(\Lambda-k)$ zero matrix, then we have prior estimates of the first k elements of β. We shall take β as fixed, which implies that r must be random: it is a vector of estimates, either from previous samples or from introspection. This is to be contrasted with the Bayesian approach which would regard (2.3) as inducing a prior distribution on β. We shall assume throughout this paper: (iii) The $k\times\Lambda$ matrix R has rank k and consists of (known) nonstochastic elements. (iv) The error vector v is distributed independently of u and has the following (known) nonsingular matrix of second-order moments:

$$E(vv') = \Psi. \tag{2.4}$$

To estimate β we can use the sample information alone, in which case the ordinary least-squares estimator,

$$b = (X'X)^{-1}X'y, \tag{2.5}$$

is best linear unbiased. But we wish to take account of the prior information as well, so that we combine (2.1) and (2.3) as follows:

$$\begin{bmatrix} y \\ r \end{bmatrix} = \begin{bmatrix} X \\ R \end{bmatrix} \beta + \begin{bmatrix} u \\ v \end{bmatrix}. \tag{2.6}$$

It follows from Assumptions (ii) and (iv) that the matrix of the second-order moments of the combined error vector is

$$E\left(\begin{bmatrix} u \\ v \end{bmatrix} [u' \quad v']\right) = \begin{bmatrix} \sigma^2 I & 0 \\ 0 & \Psi \end{bmatrix}. \tag{2.7}$$

Then, by applying generalized least-squares to (2.6) we obtain the following estimation equations:[2]

$$[X' \quad R'] \begin{bmatrix} \sigma^2 I & 0 \\ 0 & \Psi \end{bmatrix}^{-1} \begin{bmatrix} y \\ r \end{bmatrix} = [X' \quad R'] \begin{bmatrix} \sigma^2 I & 0 \\ 0 & \Psi \end{bmatrix}^{-1} \begin{bmatrix} X \\ R \end{bmatrix} \hat{\beta}^*,$$

where $\hat{\beta}^*$ is the resulting estimator. This result can be simplified to

$$\varphi X'y + R'\Psi^{-1}r = (\varphi X'X + R'\Psi^{-1}R)\hat{\beta}^*, \tag{2.8}$$

where φ stands for the "precision" of the regression process:

$$\varphi = \frac{1}{\sigma^2} \cdot \tag{2.9}$$

By inverting $\varphi X'X + R'\Psi^{-1}R$ we can solve (2.8) for $\hat{\beta}^*$. It can be easily shown that this inverse is the matrix of second-order sampling moments of $\hat{\beta}^*$:

$$V(\hat{\beta}^*) = E[(\hat{\beta}^* - \beta)(\hat{\beta}^* - \beta)'] = (\varphi X'X + R'\Psi^{-1}R)^{-1}. \tag{2.10}$$

Furthermore, on the condition that $Ev = 0$—which will be indicated in the sequel as *Assumption* (v)—one can easily prove that $\hat{\beta}^*$ is unbiased. It is then best linear unbiased with respect to sample and prior information simultaneously.

2.2. The f-Class

The difficulty is that σ^2 and hence φ are unknown. Let us therefore suppose that the investigator decides to replace φ by a known positive number f. In the first instance it will be assumed that f is nonstochastic, which will be indicated as *Assumption* (vi). Instead of (2.8) we then have the following estimation equations:

$$fX'y + R'\Psi^{-1}r = (fX'X + R'\Psi^{-1}R)\hat{\beta}_f, \tag{2.11}$$

which defines a class $\hat{\beta}_f$ of estimators, to be called the f-class.[3] On solving for $\hat{\beta}_f$ and applying (2.1) and (2.3) we obtain

[2] Generalized least-squares as developed by Aitken [1] amounts to the following. Write (2.6) in the form $z = W\beta + s$ and (2.7) as $E(ss') = \Omega$. Then the estimation equations are $W'\Omega_2^{-1} = W'\Omega^{-1}W\hat{\beta}^*$ and the inverse of $W'\Omega^{-1}W$ is the matrix of second-order sampling moments of $\hat{\beta}^*$.

[3] The f-class is similar to the k-class of simultaneous-equation estimators in the sense that in both cases the estimators are generated by linear estimation equations whose left-hand vectors and right-hand square matrices are linear functions of f or k. For the k-class see H. Theil [6, Section 6.2].

$$\hat{\beta}_f = \beta + (fX'X + R'\Psi^{-1}R)^{-1}(fX'u + R'\Psi^{-1}v), \tag{2.12}$$

whence it follows that $\hat{\beta}_f$ is unbiased for whatever f when Assumptions (v) and (vi) are true.[4] Furthermore, it follows from (2.12) that under these assumptions the covariance matrix of $\hat{\beta}_f$ is

$$V(\hat{\beta}_f) = (fX'X + R'\Psi^{-1}R)^{-1}\left(\frac{f^2}{\varphi}X'X + R'\Psi^{-1}R\right)(fX'X + R'\Psi^{-1}R)^{-1}. \tag{2.13}$$

If Assumption (vi) is true but (v) is not, then $V(\hat{\beta}_f)$ is not the covariance matrix but the matrix of second-order sampling moments. And if $f=\varphi$, (2.13) reduces to (2.10), which is as it should be.

Let us now write

$$f = \varphi(1 + \epsilon), \tag{2.14}$$

so that ϵ stands for the relative specification error of f with respect to φ. Then $f^2/\varphi = f(1+\epsilon)$, so that (2.13) can be written in the form

$$V(\hat{\beta}_f) = (fX'X + R'\Psi^{-1}R)^{-1}[I + \epsilon fX'X(fX'X + R'\Psi^{-1}R)^{-1}]. \tag{2.15}$$

It follows directly that if we write

$$V(\hat{\beta}_f) \approx (fX'X + R'\Psi^{-1}R)^{-1}, \tag{2.16}$$

the right-hand side involves a relative approximation error which is of the same order of magnitude as ϵ, i.e., of the same order of magnitude as the relative specification error in f.

Summarizing, we can conclude that as far as bias is concerned, any f-class estimator has the same properties as $\hat{\beta}^*$ [i.e., unbiased if Assumption (v) is true, biased otherwise]; and that as far as the second-order sampling moments are concerned, the use of the approximation formula (2.16), which is completely analogous to (2.10), involves a relative approximation error which is of the same order of magnitude as that of f. The proviso is, of course, that f should be nonstochastic, see Assumption (vi).

2.3. f-Estimates Based on the Sample

One can use the f-class in an experimental manner, viz., by plotting the Λ components of $\hat{\beta}_f$ as a function of f. It is easily seen from (2.11) that if f increases indefinitely, the estimator converges to the ordinary least-squares estimator (2.5) which is based on the sample only. On the other hand, if f approaches zero we attach more and more weight to the prior information. In the special case when R is square, so that the number of linear combinations of β-components about which prior information is available is equal to the number of these components, we have $\hat{\beta}_f = (R'\Psi^{-1}R)^{-1}R'\Psi^{-1}r = R^{-1}r$ in the limit for $f=0$.

To obtain unique estimates of the β-components we shall use an estimate of φ based on the sample. The argument is that if we know nothing a priori about σ^2 and hence about φ, we have at least a sample estimate which is consistent,

[4] Assumptions (i) through (iv), which are imposed throughout this paper, will no longer be mentioned explicitly.

so that it is worthwhile to analyze the implications of its use in f-class β-estimation. An obvious choice is

$$\frac{1}{s^2} = \frac{T - \Lambda}{y'y - y'X(X'X)^{-1}X'y}, \tag{2.17}$$

s^2 being the ordinary (unbiased) least-squares σ^2-estimator based on the sample. We shall write $\hat{\beta}$ for the f-class estimator whose f is the reciprocal of this s^2:

$$\hat{\beta} = (s^{-2}X'X + R'\Psi^{-1}R)^{-1}(s^{-2}X'y + R'\Psi^{-1}r). \tag{2.18}$$

Specifically, our *Assumption* (vii) will be that f is a random variable and that the difference $f-\varphi$ is $O(T^{-1/2})$ in probability, T being the number of observations in the sample information. This assumption is satisfied for $\hat{\beta}$ in the classical normal case; for if the disturbances are normally distributed, the sampling error of s^2 is $O(T^{-1/2})$ is probability, hence the same must apply to that of $1/s^2$. From now on the analysis will be of the asymptotic variety, and we should be careful in this respect in view of the two different sources of information which we are using. For example, in (2.10) we found that the moment matrix of $\hat{\beta}^*$ is the inverse of $\varphi X'X+R'\Psi^{-1}R$. Now $\varphi X'X$ is a matrix whose elements are $O(T)$, so that $\varphi X'X$ and $R'\Psi^{-1}R$ are of the same order of magnitude only if the elements of $R'\Psi^{-1}R$ are also $O(T)$. Obviously, the latter matrix has nothing to do with the size of the sample. But there is nevertheless a good reason for proceeding on the basis that both matrices are of the same order of magnitude; for if this would not be the case, either the sample or the prior information would dominate the other so that there is then little reason to use the information which is dominated.

On expanding according to powers of T and applying Assumption (vii) we find

$$\begin{aligned}(fX'X + R'\Psi^{-1}R)^{-1} &= (\varphi X'X + R'\Psi^{-1}R)^{-1} \\ &\quad - (f - \varphi)(\varphi X'X + R'\Psi^{-1}R)^{-1}X'X(\varphi X'X + R'\Psi^{-1}R)^{-1} + O(T^{-1}),\end{aligned}$$

so that (after some rearrangements)

$$\hat{\beta}_f = \hat{\beta}^* + (f - \varphi)(\varphi X'X + R'\Psi^{-1}R)^{-1}X'[u - X(\hat{\beta}^* - \beta)] + O(T^{-1}).$$

Now in the second term on the right, $\hat{\beta}^*-\beta$ is $O(T^{-1/2})$ in probability in view of (2.10). This matrix is premultiplied by $X'X$ whose elements are $O(T)$, hence $X'X(\hat{\beta}^*-\beta)$ is $O(T^{1/2})$ in probability. This vector is subtracted from $X'u$, which is also $O(T^{1/2})$ in probability, because its mean is zero and its covariance matrix is $\sigma^2X'X=O(T)$. Hence, $X'[u-X(\hat{\beta}^*-\beta)]$ is $O(T^{1/2})$ in probability, at least not of a higher order. This vector is premultiplied by $(\varphi X'X+R'\Psi^{-1}R)^{-1}$ which is $O(T^{-1})$, and this in turn by $f-\varphi$ which is $O(T^{-1/2})$ according to Assumption (vii). It follows that the second term as a whole is $O(T^{-1})$, hence

$$\hat{\beta}_f = \hat{\beta}^* + O(T^{-1}) \tag{2.19}$$

if Assumption (vii) is indeed true. It follows immediately that if Assumption (v) is also true, so that $\hat{\beta}^*$ is unbiased, $\hat{\beta}_f$ is asymptotically unbiased and that its bias is $O(T^{-1})$. Regarding the second-order sampling moments, we see from

(2.19) that the difference between $\hat{\beta}_f$ and $\hat{\beta}^*$ is $O(T^{-1})$ in probability and from (2.10) that the difference between $\hat{\beta}^*$ and $\hat{\beta}$ is $O(T^{-1/2})$ in probability. This implies that $\hat{\beta}_f$ has the same asymptotic moment matrix as $\hat{\beta}^*$, which explains the first equality sign of

$$\mathrm{V}(\hat{\beta}_f) = (\varphi X'X + R'\Psi^{-1}R)^{-1} + o(T^{-1}) = (fX'X + R'\Psi^{-1}R)^{-1} + o(T^{-1}), \quad (2.20)$$

where $o(T^{-1})$ stands for a matrix whose elements are of a higher order of smallness than T^{-1}. The second equality sign is explained by the fact that if Assumption (vii) is true, replacing φ by f leads to a relative error which is to be neglected to our order of approximation.[5]

2.4. Further Iterations

The mere fact that $\hat{\beta}_f$ has the same asymptotic moment matrix as $\hat{\beta}^*$ [if $f-\varphi=O(T^{-1/2})$ in probability] shows that it is impossible to gain in asymptotic efficiency as far as the estimation of β is concerned. It is, therefore, not true that we obtain asymptotically more efficient estimates by computing first $\hat{\beta}$ as defined in (2.18), then $\hat{\sigma}^2$ which is the σ^2-estimator based on the sum of squares of the $\hat{\beta}$-estimated disturbances (the elements of $y-X\hat{\beta}$), then $\hat{\beta}_f$ with f defined as the reciprocal of $\hat{\sigma}^2$, and so on. One might think that $\hat{\sigma}^2$ is asymptotically a more efficient estimator of σ^2 than s^2 (because s^2 is not based on prior information), but this is not the case either. The reason is the following. The disturbance vector as estimated by ordinary least-squares is $y-Xb=u-X(b-\beta)$. The elements of u are $O(1)$ in probability and dominate those of $X(b-\beta)$, which are $O(T^{-1/2})$. Now if we replace b by $\hat{\beta}$ we improve only on the dominated term. This is well illustrated by the fact that if the disturbances are normally distributed, the sampling variance of s^2 is $2\sigma^4/(T-\Lambda)$. The T in the denominator refers to the dominating u-vector, the Λ to $X(b-\beta)$.

If one wishes nevertheless to use an estimator of σ^2 which takes the prior information into account, it is recommended to use

$$\hat{\sigma}^2 = \frac{(y - X\hat{\beta})'(y - X\hat{\beta})}{T - \mathrm{tr}\, s^{-2}X'X(s^{-2}X'X + R'\Psi^{-1}R)^{-1}} \quad (2.21)$$

(where tr A, the trace of a square matrix A, stands for the sum of its diagonal elements). It is shown in the Appendix that the bias of $\hat{\sigma}^2$ is of a higher order of smallness than $1/T$.

3. TESTING THE COMPATIBILITY OF PRIOR AND SAMPLE INFORMATION

Any applied research worker will certainly wish to take account of the possibility that his prior and sample information may be in conflict with each other. The following procedure (which is of the classical type) is proposed to test the null-hypothesis that they are in agreement. Under this null-hypothesis we have two independent estimators of the k-element vector $R\beta$, viz., the prior estimator

[5] It is not difficult to see that this result can be generalized for the case when the prior information is derived from a previous sample of size T' and when Ψ is not known but estimated as $s'^2\Psi_0$ such that $s'-\sigma'=0(1/\sqrt{T'})$ and $\Psi=\sigma'^2\Psi_0$. The obvious way to estimate is to apply (2.18) with Ψ replaced by $s'^2\Psi_0$ and (2.19) then holds except that the error term is now $0(1/T'')$ with $T''=\mathrm{Min}\,(T, T')$. The same applies with the same modification to (2.20).

r and the least squares sample estimator $R(X'X)^{-1}X'y$. Their difference is $v-R(X'X)^{-1}X'u$ and the matrix of second moments of this difference is

$$E[\{v - R(X'X)^{-1}X'u\}\{v - R(X'X)^{-1}X'u\}'] = \sigma^2 R(X'X)^{-1}R' + \Psi. \qquad (3.1)$$

This suggests the use of the following test statistic:

$$\gamma = (r - Rb)[\sigma^2 R(X'X)^{-1} R' + \Psi]^{-1}(r - Rb), \qquad (3.2)$$

and we shall derive its distribution under normality conditions, more precisely —*Assumption* (viii): The elements of the vector u are normally distributed and those of v are normally distributed with zero mean. [The zero-mean condition on u has already been imposed in Assumption (ii).] Under this assumption the difference $d=r-Rb$ is normally distributed with zero mean and (3.1) as covariance matrix. Since the latter matrix is positive-definite, its inverse can be written as $D'D$ where D is a $k\times k$ matrix of full rank. The test statistic γ takes then the form $d'D'Dd=(Dd)'Dd$, i.e., it is the sum of squares of the k elements of the vector Dd. Each of these elements is normally distributed with zero mean; in addition to this, they are uncorrelated and have unit variance as follows from

$$E(Ddd'D') = DE(dd')D' = D[\sigma^2 R(X'X)^{-1}R' + \Psi]D' = D(D'D)^{-1}D' = I.$$

It follows that γ is distributed according to χ^2 with k degrees of freedom. For practical purposes one will have to work with

$$\hat{\gamma} = (r - Rb)'[s^2 R(X'X)^{-1}R' + \Psi]^{-1}(r - Rb), \qquad (3.3)$$

which will be called the *compatibility statistic* and which has asymptotically the same distribution. If desired, one can replace s^2 by $\hat{\sigma}^2$ as defined in (2.21).

4. A NUMERICAL EXAMPLE[6]

Our example deals with the demand for textile in The Netherlands in 1923–1938. The regression equation has the form

$$\log c_t = \alpha + \beta_1 \log p_t + \beta_2 \log M_t + u_t, \qquad (4.1)$$

where c_t stands for per capita textile consumption in year t, p for the deflated price index of textiles, and M for real per capita income. Hence $T=17$, $\Lambda=3$. We have no prior beliefs regarding α and σ^2. For the price elasticity we take a prior estimate of -0.7, for the income elasticity of 1, in both cases with a standard error of 0.15. In addition to this we shall assume that if our prior beliefs overestimate the income sensitivity, the same should be true for the price sensitivity and vice versa, at least on the average. Given the negative sign of the price elasticity this means that the prior cross-moment is negative. In numerical terms our prior specification is

$$r = \begin{bmatrix} -0.7 \\ 1 \end{bmatrix}; \qquad R = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}; \qquad \Psi = \begin{bmatrix} 0.0225 & -0.01 \\ -0.01 & 0.0225 \end{bmatrix}. \qquad (4.2)$$

[6] The author is indebted to Mr. J. I. Vorst for his computational assistance.

The sample is given here in terms of sums of squares and products:[7]

$$\begin{bmatrix} X'X & X'y \\ y'X & y'y \end{bmatrix} = \begin{bmatrix} 17 & 31.83389 & 34.20783 & 36.07631 \\ & 59.75950 & 64.06457 & 67.44192 \\ & & 68.84179 & 72.59642 \\ & & & 76.65898 \end{bmatrix}. \tag{4.3}$$

The first step consists of computations on the sample: the least-squares coefficients, their estimated covariance matrix, the estimated variance of the disturbances (s^2), and its reciprocal:

$$b = \begin{bmatrix} 1.374 \\ -0.829 \\ 1.143 \end{bmatrix}; \qquad s^2(X'X)^{-1} = \begin{bmatrix} 0.09404 & 0.00008 & -0.04680 \\ & 0.00131 & -0.00126 \\ & & 0.02443 \end{bmatrix}. \tag{4.4}$$

$$s^2 = 0.0001836; \qquad s^{-2} = 5445. \tag{4.5}$$

Hence the sample produces an estimate of the price elasticity of -0.83 with a standard error of 0.04 and an estimate of the income elasticity of 1.14 with a standard error of 0.16.

The second step is the testing of the compatibility of prior and sample information. This requires the matrix $s^2R(X'X)^{-1}R'$, which in this case is nothing else than the lower right-hand 2×2 submatrix of $s^2(X'X)^{-1}$. After adding Ψ and inverting we obtain the matrix of the quadratic form $\hat{\gamma}$, see (3.3). Its vector $r-Rb$ consists of two components, viz., 0.129 and -0.143 [see (4.2) and (4.4)]. The result is

$$\hat{\gamma} = 0.86. \tag{4.6}$$

This is far from significant, the 10 per cent significance point of the χ^2-distribution with $k=2$ degrees of freedom being about $4\frac{1}{2}$.[8]

The third step consists of the computation of $\hat{\beta}$ and its estimated asymptotic moment matrix. This matrix is the inverse of $s^{-2}X'X+R'\Psi^{-1}R$, where $R'\Psi^{-1}R$ is in this case simply Ψ^{-1} bordered by a zero column on the left and a zero row on top. The inverse is

$$(s^{-2}X'X + R'\Psi^{-1}R)^{-1} = \begin{bmatrix} 0.04143 & -0.00064 & -0.01999 \\ & 0.00122 & -0.00082 \\ & & 0.01070 \end{bmatrix} \tag{4.7}$$

and after postmultiplication by $s^{-2}X'y+R'\Psi^{-1}r$ we obtain

$$\hat{\beta} = \begin{bmatrix} 1.467 \\ -0.821 \\ 1.089 \end{bmatrix}. \tag{4.8}$$

The result is therefore that the "new" estimated price and income elasticities—based both on data and on prior beliefs—are -0.82 and 1.09, respectively,

[7] The underlying data have been published in Theil and Nagar [8], Appendix.

[8] If s^2 in $\hat{\gamma}$ is replaced by $\hat{\sigma}^2$ as derived in (4.11) below, the result is only slightly different.

and that the (asymptotic) standard errors are reduced to 0.03 and 0.10, respectively. The estimate of the constant term is raised from 1.37 to 1.47, its standard error is reduced from 0.31 to 0.20.

The fourth step (optional!) is the computation of $\hat{\sigma}^2$. For this purpose we need

$$\operatorname{tr} s^{-2}X'X(s^{-2}X'X + R'\Psi^{-1}R)^{-1} = 2.380, \tag{4.9}$$

which is computed by multiplying the λth row of $s^{-2}X'X$ into the λth column of $(s^{-2}X'X+R'\Psi^{-1}R)^{-1}$ and summing over λ. We also need the sum of squares of the $\hat{\beta}$-estimated disturbances:

$$y'y - 2(X'y)'\hat{\beta} + \hat{\beta}'X'X\hat{\beta} = 0.002597 \tag{4.10}$$

and the result is

$$\hat{\sigma}^2 = 0.00178, \tag{4.11}$$

which is very close to the sample estimate (4.5).

5. MEASURING THE SHARES OF PRIOR AND SAMPLE INFORMATION IN THE POSTERIOR PRECISION[9]

We can say that in the example just given, our prior knowledge consists of elasticity estimates of -0.7 and 1 with standard errors of 0.15 (plus a prior covariance of 0.01); that our sample knowledge consists of (a) a constant term of 1.37 and elasticity estimates of -0.83 and 1.14 with standard errors of 0.31, 0.04 and 0.16, respectively (plus three covariance estimates), and (b) of a φ-estimate of 5445 which differs from φ itself to $O(T^{1/2})$; and that our posterior knowledge consists of a constant term of 1.47 and elasticity estimates of -0.82 and 1.09 with asymptotic standard errors of 0.20, 0.03 and 0.10, respectively. Given that prior and sample information are assumed to be independent, it is a reasonable question to ask how we can measure the share of these two ingredients in our posterior knowledge.

The essential point, of course, is that our posterior knowledge is measured by the moment matrix of the posterior estimator:

$$V(\hat{\beta}) \approx (\varphi X'X + R'\Psi^{-1}R)^{-1}, \tag{5.1}$$

because this matrix measures the precision of the estimator. We shall confine ourselves to the leading term of the moment matrix [i.e., to the inverse after $\approx$ sign in (5.1)]. It is seen that $\varphi X'X$ refers to sample information and $R\Psi^{-1}R$ to prior information; and hence, writing $A=\varphi X'X$ and $B=R'\Psi^{-1}R$ for notational convenience, our problem can be formulated as follows: how can we measure the share of two (symmetric and positive-definite or positive semi-definite) matrices A and B in $(A+B)^{-1}$? In other words, how do we specify a (scalar) function $g(A, B)$ measuring the share of A in $(A+B)^{-1}$ [and similarly $g(B, A)$ for measuring the share of B in $(A+B)^{-1}$]?

If we pose the problem in this manner, there is of course no unique solution. We shall therefore impose certain "reasonable" requirements which g has to satisfy. The first is the *adding-up criterion*, which specifies that the two g's

[9] The subject of this section was originally approached by the author in collaboration with the late Mr. P. J. M. van den Bogaard.

always add up to 1:

$$g(A, B) + g(B, A) \equiv 1, \tag{5.2}$$

for whatever (symmetric and positive-definite or positive semi-definite) A and B. This ensures that the g's really measure "shares." The second is the *zero-unit criterion,* which implies

$$g(O, B) = 0; \qquad g(A, O) = 1, \tag{5.3}$$

i.e., the limits of the interval (0, 1) are attained if one of the two matrices is a zero matrix. The third is the *invariance criterion for nonsingular linear transformations,* which can be explained as follows. It would evidently be unsatisfactory if a mere change in units of some explanatory variable would lead to a change in the way in which the shares of prior and sample information are measured. Now if we would multiply the units of the first explanatory variable (say) by k, this means that the first coefficient is multiplied by $1/k$ and hence that the first row and the first column of $A = \varphi X'X$ and $B = R'\Psi^{-1}R$ are multiplied by $(1/k)^{-1} = k$. We should require that g be invariant against such changes. More generally, if we change the units of all explanatory variables, this amounts to pre- and post-multiplying A and B by the same diagonal matrix. We shall go even farther by requiring that g be invariant when A and B are both post-multiplied by K and premultiplied by the transpose of K, for whatever square and nonsingular K; this amounts to invariance against any (nonsingular) linear combination of the explanatory variables. Hence:

$$g(K'AK, K'BK) \equiv g(A, B) \tag{5.4}$$

for all square and nonsingular K-matrices (and for all A, B).

The three criteria (5.2) – (5.4) do not determine g uniquely, which is the reason why we impose the following *linearity criterion.* Suppose we have two pairs of symmetric and positive-definite or semi-definite matrices, A_1, B_1 and A_2, B_2, all of the same order and also with the same sum matrix:

$$A_1 + B_1 = A_2 + B_2. \tag{5.5}$$

In our terminology this means that in both cases we have the same posterior moment matrix $(\varphi X'X + R'\Psi^{-1}R)^{-1}$, but $\varphi X'X$ and $R'\Psi^{-1}R$ are different. Next, we consider a third pair, $pA_1 + qA_2$ and $pB_1 + qB_2$, were p and q are any non-negative scalars with sum 1. It is clear that (5.5) is also the sum of these two matrices, hence the same posterior moment matrix applies to this case. The linearity criterion requires that the share of $pA_1 + qA_2$ is then p times the share of A_1 plus q times the share of A_2:

$$g(pA_1 + qA_2, pB_1 + qB_2) \equiv pg(A_1, B_1) + qg(A_2, B_2) \tag{5.6}$$

where $p + q = 1$ (p, $q \geq 0$), and whenever (5.5) is satisfied.

We shall now show that the only g which satisfies these four criteria is

$$g(A, B) = \frac{1}{\Lambda} \operatorname{tr} A(A + B)^{-1}. \tag{5.7}$$

In the scalar case of one explanatory variable ($\Lambda = 1$), (5.7) reduces to $A/(A + B)$. On substituting $A = \varphi X'X$ and $B = R'\Psi^{-1}R$ we find that this is the ratio of the reciprocal of the prior variance to the reciprocal of the posterior variance, which is a measure proposed earlier by Schlaifer [5, pp. 440 ff.]. It is also seen that, when A and B differ only in a multiplicative scalar, i.e., $A=p(A+B)$ and $B=q(A+B)$, we have $g(A, B)=p$; which seems reasonable.

The proof is as follows. Since $A+B$ is a symmetric and positive-definite matrix, there exists a square and nonsingular matrix Q such that

$$(A + B)^{-1} = QQ'. \tag{5.8}$$

Let us define

$$A_* = Q'AQ \quad \text{and} \quad B_* = Q'BQ, \tag{5.9}$$

which should have the same g according to the invariance criterion (5.4). It is easily seen that the sum of these matrices is the unit matrix:

$$A_* + B_* = Q'(A + B)Q = Q'(QQ')^{-1}Q = Q'Q'^{-1}Q^{-1}Q = I. \tag{5.10}$$

Since A_* is a real symmetric matrix, it can be written as

$$A_* = ZMZ^{-1} = ZMZ', \tag{5.11}$$

where M is a diagonal matrix whose elements are the characteristic roots of A_* and where Z is an orthogonal matrix. From (5.10) and (5.11) it is easily seen that

$$B_* = Z(I - M)Z', \tag{5.12}$$

where the diagonal elements of $I-M$ are the roots of B_*. Since both A_* and B_* are positive semi-definite, we can conclude that all diagonal elements of M are ≥ 0 and ≤ 1.

On combining (5.9), (5.11) and (5.12) we conclude

$$A = (Z'Q^{-1})'M(Z'Q^{-1}) \quad \text{and} \quad B = (Z'Q^{-1})'(I - M)(Z'Q^{-1}), \tag{5.13}$$

whence it follows immediately that the invariance criterion (5.4) prescribes that the shares (the g's) of A and B should be the same as those of M and $I-M$; in other words, that g should be uniquely determined by the latent roots ofA_*. Furthermore, it is easily seen that g should be a symmetric function of these roots; for if we premultiply M by a permutation matrix and postmultiply by the transpose of that matrix [which should leave g unchanged according to (5.4)], the result is that certain diagonal elements of M are interchanged in position.

The linearity criterion (5.6) can be used to show that g should be a linear function of these roots. We may assume without loss of generality that $A_1+B_1=A_2+B_2=I$; viz., by applying the Q-transformation of (5.8) and (5.9). We then have

$$\begin{aligned} A_1 &= Z_1M_1Z_1'; & A_2 &= Z_2M_2Z_2'; \\ B_1 &= Z_1(I - M_1)Z_1'; & B_2 &= Z_2(I - M_2)Z_2', \end{aligned} \tag{5.14}$$

where M_1 and M_2 are diagonal matrices whose elements are the roots of A_1 and A_2, respectively, while Z_1 and Z_2 are orthogonal matrices. We shall now distinguish between the case in which Z_1 and Z_2 are equal and the case in which they are different; and we shall show (i) that in the case of equality g should necessarily be linear in the latent roots and (ii) that the trace definition (5.7) satisfies the present linearity criterion even if the Z's are different. The latter proposition is verified in a straightforward manner. As to the former, if $Z_1 = Z_2 = Z$ (say), we have

$$\begin{aligned} pA_1 + qA_2 &= Z(pM_1 + qM_2)Z' \\ pB_1 + qB_2 &= Z[I - (pM_1 + qM_2)]Z'. \end{aligned} \tag{5.15}$$

On combining (5.4) and (5.6) we find

$$\begin{aligned} g[pM_1 + qM_2,\ p(I - M_1) + q(I - M_2)] &\equiv \\ \equiv pg(M_1, I - M_1) + qg(M_2, I - M_2); \end{aligned} \tag{5.16}$$

or, if we write $g(M, I-M)$ in the simpler form $h(\mu)$ where μ is the vector of latent roots (the diagonal of M):

$$h(p\mu^1 + q\mu^2) \equiv ph(\mu^1) + qh(\mu^2), \tag{5.17}$$

where μ^1 corresponds with M_1 and μ^2 with M_2. It follows that h should be linear in the latent roots. This linear function should be symmetric (see the end of the preceding paragraph), hence it should be of the form $c + k\sum \mu_\lambda$, where c and k are scalars which are still to be determined. It is easily seen that $c = 0$ in view of the zero-unit criterion (5.3), and $k = 1/\Lambda$ in view of the adding-up criterion (5.2). It follows that the only g which satisfies all four criteria is the average value of the latent roots of A_* which is the same as $1/\Lambda$ times the trace of A_*. Since we have

$$\operatorname{tr} A_* = \operatorname{tr} Q'AQ = \operatorname{tr} AQQ' = \operatorname{tr} A(A + B)^{-1}, \tag{5.18}$$

we have proved that (5.7) is the measure which we want.[10]

We now return to the original interpretation $A = \varphi X'X$ and $B = R'\Psi^{-1}R$, and replace φ by $1/s^2$ in order to obtain a quantity which can be calculated. Then

$$\begin{aligned} \theta_S &= \frac{1}{\Lambda} \operatorname{tr} s^{-2}X'X(s^{-2}X'X + R'\Psi^{-1}R)^{-1} \\ \theta_P &= \frac{1}{\Lambda} \operatorname{tr} R'\Psi^{-1}R(s^{-2}X'X + R'\Psi^{-1}R)^{-1} \end{aligned} \tag{5.19}$$

are the estimated shares of sample and prior information, respectively, in our posterior knowledge. For the example of Section 4 we have $\theta_S = 0.79$ as is easily verified from (4.9). Hence almost 80 per cent of the posterior precision is due to sample information, only slightly more than 20 per cent to prior information.

[10] It is worthwhile to mention that this measure is analogous to Hooper's [3] trace correlation coefficient, which serves to measure the degree to which the total variation in a number of jointly dependent variables of a linear econometric model is accounted for by the total variation in the predetermined variables; the term "total variation" is interpreted as the covariance matrix. Also see T. W. Anderson [2, p. 223] in a different context.

It will also be observed that the residual variance estimator $\hat{\sigma}^2$ of (2.21) can be expressed elegantly in terms of θ_S:

$$\hat{\sigma}^2 = \frac{(y - X\hat{\beta})'(y - X\hat{\beta})}{T - \theta_S \Lambda} \cdot \tag{5.20}$$

This result shows that the loss of degrees of freedom, which is Λ in the case of sample least-squares estimation, is reduced here in proportion to the share of the prior information in our posterior knowledge.

APPENDIX

To derive $\hat{\sigma}^2$ as defined in (2.21) and to prove that its bias is of a higher order of smallness than T^{-1}, we start by writing the vector of $\hat{\beta}$-estimated disturbances in the form

$$y - X\hat{\beta} = u - X(\hat{\beta}^* - \beta) - X(\hat{\beta} - \hat{\beta}^*). \tag{A.1}$$

The elements of the first vector (u) are $O(1)$, those of the second $O(T^{-1/2})$ and those of the third $O(T^{-1})$ in probability, see (2.19). The sum of squares of these estimated disturbances is

$$u'u - 2u'X(\hat{\beta}^* - \beta) + (\hat{\beta}^* - \beta)'X'X(\hat{\beta}^* - \beta)$$

plus the following three terms:

$$-2u'X(\hat{\beta} - \hat{\beta}^*) + 2(\hat{\beta}^* - \beta)'X'X(\hat{\beta} - \hat{\beta}^*) + (\hat{\beta} - \hat{\beta}^*)'X'X(\hat{\beta} - \hat{\beta}^*). \tag{A.2}$$

Now the first term in (A.2) is $O(T^{-1/2})$ in probability; this follows from the fact that the elements of $u'X$ are $O(T^{1/2})$ in probability [since the mean of this vector is zero and its covariance matrix is $\sigma^2 X'X = O(T)$] and that $\hat{\beta} - \hat{\beta}^* = O(T^{-1})$. As to the second term, we have $\hat{\beta}^* - \beta = O(T^{-1/2})$, $X'X = O(T)$, $\hat{\beta} - \hat{\beta}^* = O(T^{-1})$, hence this term is $O(T^{-1/2})$ in probability. The third is $O(T^{-1})$ as is easily verified. It follows that the sum of squares of the $\hat{\beta}$-estimated disturbances is

$$\begin{aligned}(y - X\hat{\beta})'(y - X\hat{\beta}) &= u'u - 2u'X(\hat{\beta}^* - \beta) \\ &\quad + (\hat{\beta}^* - \beta)'X'X(\hat{\beta}^* - \beta) + O(T^{-1/2}).\end{aligned} \tag{A.3}$$

The expectation of the first term, $u'u$, is $T\sigma^2$; that of the third is

$$\begin{aligned}E[(\hat{\beta}^* - \beta)'X'X(\hat{\beta}^* - \beta)] &= \operatorname{tr} X'XE[(\hat{\beta}^* - \beta)(\hat{\beta}^* - \beta)'] \\ &= \operatorname{tr} X'X(\varphi X'X + R'\Psi^{-1}R)^{-1},\end{aligned}$$

and that of the second:

$$\begin{aligned}-2E[u'X(\hat{\beta}^* - \beta)] &= -2 \operatorname{tr} XE[(\varphi X'X + R'\Psi^{-1}R)^{-1}(\varphi X'u + R'\Psi^{-1}v)u'] \\ &= -2 \operatorname{tr} X'X(\varphi X'X + R'\Psi^{-1}R)^{-1}.\end{aligned}$$

On combining these results we obtain

$$\begin{aligned}&E[(y - X\hat{\beta})'(y - X\hat{\beta})] \\ &= T\sigma^2 - \operatorname{tr} X'X(\varphi X'X + R'\Psi^{-1}R)^{-1} + O(T^{-1/2}) \qquad \text{(A.4)} \\ &= \sigma^2\{T - \operatorname{tr} \varphi X'X(\varphi X'X + R'\Psi^{-1}R)^{-1}\} + O(T^{-1/2}) \\ &= \sigma^2\{T - \operatorname{tr} s^{-2}X'X(s^{-2}X'X + R'\Psi^{-1}R)^{-1}\} + O(T^{-1/2}).\end{aligned}$$

It follows immediately that

$$\hat{\sigma}^2 = \frac{(y - X\hat{\beta})'(y - X\hat{\beta})}{T - \operatorname{tr} s^{-2}X'X(s^{-2}X'X + R'\Psi^{-1}R)^{-1}}$$

has indeed a bias with respect to σ^2 which is of a higher order of smallness than $1/T$.

REFERENCES

[1] Aitken, A. C., "On least squares and linear combination of observations," *Proceedings of the Royal Society of Edinburgh*, 55 (1934–35), 42–8.

[2] Anderson, T. W., *An Introduction to Multivariate Statistical Analysis*. New York and London: John Wiley and Sons, Inc., 1958.

[3] Hooper, J. W., "Simultaneous equations and canonical correlation theory," *Econometrica*, 27 (1959), 245–56.

[4] Raiffa, H., and Schlaifer, R. *Applied Statistical Design Theory*. Boston: Harvard Business School, 1961.

[5] Schlaifer, R. *Probability and Statistics for Business Decisions*. New York: McGraw-Hill Book Company, 1959.

[6] Theil, H., *Economic Forecasts and Policy. Second Edition*. Amsterdam: North-Holland Publishing Company, 1961.

[7] Theil, H., and Goldberger, A. S., "On pure and mixed statistical estimation in economics," *International Economic Review*, 2 (1960), 65–78.

[8] Theil, H., and Nagar, A. L., "Testing the independence of regression disturbances," *Journal of the American Statistical Association*, 56 (1961), 793–806.