Teaching Econometric Theory from the Coordinate-Free Viewpoint

Gordon Fisher - Montreal, Canada

1. Introduction

The principal aim of this paper is to demonstrate how the coordinate-free methods of linear statistical models may be adapted to the analysis of econometric models, and to explain why such methods are useful for teaching purposes.

The application of coordinate-free methods to linear statistical models cast in Euclidean space was first introduced by Kruskal (1961). Later, the methodology was extended to multivariate analysis by Dempster (1968) and Eaton (1970). Drygas (1970) and Philoche (1971) give coordinate-free expositions of minimum-variance linear unbiased [or Gauss-Markov (GM)] estimation, and Seber (1964a,b,c and 1980) makes use of coordinate-free methods in a comprehensive treatment of the linear hypothesis.

2. Coordinate-free methods

A coordinate-free argument is an argument that does not depend on a specific set of coordinates. Consider, for example, the n×k real matrix X of rank k. The set of all linear combinations of the columns of X forms a vector subspace called the range-space of X; this is written $\mathfrak{R}[X] = \{\theta: \theta = X\lambda, \lambda \in \mathbb{R}^k\}$. The dimension of $\mathfrak{R}[X]$ is the rank of X, and since X has n rows, $\mathfrak{R}[X]$ is a subspace of $\mathbb{R}^n$. If Z = XM, M being a k×k non-singular real matrix, then Z has rank k and $\mathfrak{R}[X] = \mathfrak{R}[Z]$. Clearly there are many more matrices like Z which may be constructed as k linearly independent linear combinations of the columns of X, and each will generate the same subspace. Let this subspace be $\mathfrak{X}$; then $\mathfrak{X}$ has dimension k and the matrices X and Z are examples of different bases of $\mathfrak{X}$ in $\mathbb{R}^n$. If $x = X\alpha$, then α is said to be the coordinate vector of x relative to the basis X of $\mathfrak{X}$; if $x = Z\lambda$, then λ is the coordinate vector of x relative to the basis Z of $\mathfrak{X}$. If a basis is not specified, then the simple notation $x \in \mathfrak{X}$ suffices,

indicating location without reference to a basis and hence without reference to coordinates: such notation is thereby coordinate free.

The (n-k) dimensional subspace of $\mathbb{R}^n$ of all vectors orthogonal to $\mathfrak{X}$ is written $\mathfrak{X}^\perp$ and may be defined, relative to the basis X, as $\mathfrak{X}^\perp = \{\theta: X^T\theta = 0, \theta \in \mathbb{R}^n\}$; this subspace is also called the nullspace of X^T, written $\mathfrak{N}[X^T]$. Notice that $\mathfrak{R}[X]^\perp = \mathfrak{N}[X^T] = \mathfrak{R}[Z]^\perp = \mathfrak{N}[Z^T]$.

Turning to statistical models, let Y be a random vector which ranges over $\mathbb{R}^n$ according to $N(\mu, \Sigma\sigma^2)$, μ being the n×1 mean vector of Y and $\Sigma\sigma^2$ its positive-definite (p.d.) dispersion, σ^2 being a positive scalar. A linear model would put $\mu = X\beta$ for some suitable X, β being a k-tuple of coefficients. Equally, $\mu = Z\gamma$ for some M where $M\gamma = \beta$: the vector μ has coordinate vector β relative to the basis X of $\mathfrak{X}$, and coordinate vector γ relative to the basis Z of $\mathfrak{X}$. The least squares (OLS) estimator of μ, based upon the n×1 observation y of Y, is $\hat{\mu}$ and this may be written, in coordinate form, as

$$(1) \qquad \hat{\mu} = X(X^TX)^{-1}X^Ty = X\hat{\beta} = Z(Z^TZ)^{-1}Z^Ty = Z\hat{\gamma},$$

$\hat{\beta}$ and $\hat{\gamma}$ being the least squares estimators corresponding to β and γ. The expression $X(X^TX)^{-1}X^T$ is the matrix of the linear transformation P from $\mathbb{R}^n$ to $\mathfrak{X}$ relative to the basis X, and $Z(Z^TZ)^{-1}Z^T$ is the corresponding matrix of the same linear transformation relative to the basis Z. Generally, the matrix of the same linear transformation will be different relative to two different bases but, in the case of (1), $X(X^TX)^{-1}X^T = Z(Z^TZ)^{-1}Z^T$. This may be seen by substituting Z = XM into the last expression. What this demonstrates is that for every n×k matrix basis B of $\mathfrak{X}$, the n×n matrix P of the linear transformation $P_{\mathfrak{X}}$ always takes the form $P = B(B^TB)^{-1}B^T$.

In coordinate dependent terms, the distribution of Y is expressed as $N(X\beta, \Sigma\sigma^2)$ or $N(Z\gamma, \Sigma\sigma^2)$, and $\hat{\beta} = (X^TX)^{-1}X^Ty$ or $\hat{\gamma} = (Z^TZ)^{-1}Z^Ty$ are the estimated coordinate vectors relative to X and Z respectively. The corresponding coordinate-free expressions are $N(\mu, \Sigma\sigma^2)$, $\mu \in \mathfrak{X}$, $\hat{\mu} = P_{\mathfrak{X}}y$, and any convenient basis of $\mathfrak{X}$ suffices to compute $\hat{\mu}$. $P_{\mathfrak{X}}$ takes any y in $\mathbb{R}^n$ onto $\mathfrak{X}$ such that $(y-P_{\mathfrak{X}}y)$ is orthogonal to $\mathfrak{X}$, i.e. $(y-Py)^Tx = 0$ $\forall x \in \mathfrak{R}[X]$. For this reason, $P_{\mathfrak{X}}$ is called the orthogonal projection from $\mathbb{R}^n$ to $\mathfrak{X}$. Also, since $(I_n - P_{\mathfrak{X}})y \in \mathfrak{X}^\perp$, $(I_n - P_{\mathfrak{X}})$ is the orthogonal projection from $\mathbb{R}^n$ to $\mathfrak{X}^\perp$.

A more efficient estimator of μ than $\hat{\mu}$ is the generalised least squares (GLS) estimator

$$(2) \qquad \tilde{\mu} = X(X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}y = Fy,$$

where $F = X(X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1} = Z(Z^T\Sigma^{-1}Z)^{-1}Z^T\Sigma^{-1}$. Given Σ, F is the unique matrix of the linear transformation, $F_{\mathfrak{X}}$, taking any y in $\mathbb{R}^n$ onto $\mathfrak{X}$, such that $(y - F_{\mathfrak{X}}y)$ is orthogonal to $\Sigma^{-1}\mathfrak{X}$; that is, in coordinate dependent terms, such that $(y - Fy)^T\Sigma^{-1}x = 0$ for all $x \in \mathfrak{R}[X]$. While P is symmetric and idempotent ($P = P^T = P^2$), F is merely idempotent ($F = F^2$). $F_{\mathfrak{X}}$ is an oblique projection, F being the corresponding oblique projection matrix. In Figure 1, the vector y is put in $\mathbb{R}^2$ and $\mathfrak{X}$ is a line. $\Sigma^{-1}\mathfrak{X} = \mathfrak{R}[\Sigma^{-1}X]$ lies below $\mathfrak{X}$; at right-angles to it is $[\Sigma^{-1}\mathfrak{X}]^\perp = \mathfrak{R}[\Sigma^{-1}X]^\perp = \mathfrak{N}[X^T\Sigma^{-1}]$. Thus $F_{\mathfrak{X}}$ is seen to take y from $\mathbb{R}^2$ to $\mathfrak{X}$ along $[\Sigma^{-1}\mathfrak{X}]^\perp$. Of course, $Fy = X\tilde{\beta}$, where $\tilde{\beta}$ is $(X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}y$, and this is the coordinate dependent expression for $\tilde{\mu}$; the

corresponding coordinate-free expression is $F_{\mathfrak{X}}y$ and this may be calculated, for given Σ, with any chosen basis of $\mathfrak{X}$.

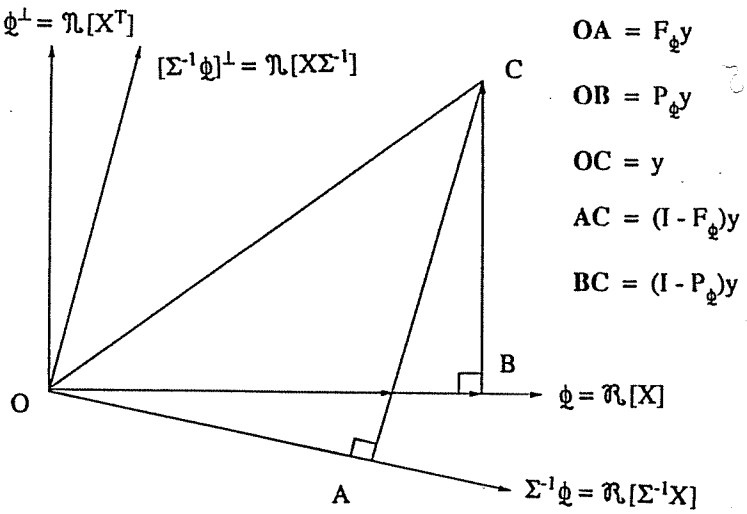

FIGURE 1

The simple geometry of OLS and GLS in Figure 1 makes use of the orthogonality conditions (OLS) $x^Tz = 0$ between two coordinate vectors x and z in $\mathbb{R}^n$ and (GLS) $x^T\Sigma^{-1}z = 0$. The expression x^Tz is called the natural scalar product between x and z and, when x and z are in coordinate-free form, this will be written (x,z) $\forall x, z \in \mathbb{R}^n$. The expression $x^T\Sigma^{-1}z$ is a special case of the scalar product induced by Σ on $\mathfrak{R}[\Sigma]$ and, when x and z are in coordinate-free form, this is written $(x,z)_{\Sigma}$. When Σ is p.d. then $(x,z)_{\Sigma} = (x,\Sigma^{-1}z)$ $\forall x, z \in \mathbb{R}^n$; but when Σ is non-negative definite (n.n.d.) then $(\Sigma x,\Sigma z)_{\Sigma} = (x,\Sigma z)$ $\forall x, z \in \mathbb{R}^n$ (Philoche, 1971). When $\mathbb{R}^n$ is endowed with a scalar product, either $(.,.)$ or $(.,.)_A$, for some appropriate n×n p.d. or n.n.d. matrix A, then $\mathbb{R}^n$ is said to constitute a Euclidean space, written $\mathfrak{E}$.

In review the principal advantage of carrying through an argument in coordinate vectors is that the resulting formulae are directly computable. This is clearly of practical value. Yet there is also a real disadvantage, because coordinates ultimately clutter a theoretical argument and thereby help mask the underlying mathematical structure. One advantage of a coordinate-free argument is that it gives emphasis to the mathematical, and particularly the geometrical, structure and hence has powerful intuitive appeal. A second advantage is that coordinate-free arguments are characteristically simple and elegant and thereby permit, didactically speaking, a rapid development of theoretical ideas. When translated into classroom time, these two advantages imply that a coordinate-free approach allows a theoretical argument to be efficiently advanced to the benefit of understanding, leaving time for a discussion of practical matters, like computing, as a side issue involving the choice of an appropriate basis.

3. Application in econometrics

For a wide range of estimation methods in econometrics, choice of method may be cast in terms of choice of scalar product. Since the mathematical analysis may be carried out for any admissible scalar product, it follows that one mathematical analysis suffices for the whole range of scalar products, and hence for a corresponding range of econometric estimation techniques.

To illustrate, consider a characteristic econometric problem, that of estimating and testing one equation of a linear system comprising M interdependent or endogenous variables and K weakly exogenous variables. The one equation will be assumed to contain (m+1) endogenous, k weakly exogenous variables, and a general p.d. error dispersion.

Two critical problems immediately arise: how to accommodate (i) a general error dispersion, and (ii) interdependence into estimation and inference. Each problem will be tackled by considering an appropriate transformation of the model and how such transformations determine the scalar product with which to work. A transformation to accommodate a general error dispersion is called a transformation to standard form (TSF) and a transformation to accommodate interdependence is called a transformation for orthogonality (TFO). If Q is a TSF which is applied first, and B is a TFO, which is applied second, their combined effect on the natural scalar product (w,z) of two vectors w and z in $\mathfrak{C}$ is

$$(3) \qquad (BQw, BQz) = (w, Q^TB^TBQz) = (w, Az)$$

where $A = Q^TB^TBQ$ (B and Q being treated as matrices).

Four didactic considerations are thrown up by use of the scalar product (., A.).

(i) It happens that A is characteristically n.n.d. and hence singular. Thus (., A.) is not a proper inner product. Such singularity leads to a consideration of the scalar product induced by A on $\mathfrak{R}[A]$ mentioned in Section 2.3 above: $(., A.) = (., .)_A$ when A is singular. If the regression subspace is $\mathfrak{X}$, then to ensure unique estimation it is necessary and sufficient that $\mathfrak{X} \cap \mathfrak{N}[A] = \phi \Leftrightarrow \mathfrak{X}^{\perp} + \mathfrak{R}[A] = \mathfrak{C} \Rightarrow (K - k) \geq m$, a necessary condition for identifiability of the equation. If $m = 0$, leaving only one endogenous variable, then $B = I_n$. If, then, Σ is n.n.d., there is no guarantee that the regression subspace $\mathfrak{X}$ will be a proper subspace of $\mathfrak{R}[\Sigma]$, which is the subspace over which the endogenous variable ranges. In this case it is necessary to redefine the GLS estimator (2) and, in particular, to define the orthogonal projection from $\mathfrak{R}[\Sigma]$ to $\mathfrak{X} \cap \mathbb{R}[\Sigma]$.

(ii) A second question arising from the use of (., A.) is: When is estimation governed by (., A.) different from estimation governed by (., .)? When $m = 0$, $B = I_n$ and Σ is p.d., this amounts to asking when are the OLS estimators in (1) and the estimators in (2) the same? The answer to this question is an invariance condition which is stated below in a general form of Kruskal's (1968) Theorem.

(iii) When considering a model comprising variables classified as endogenous and exogenous with a general error dispersion structure, it is clearly very important to give consideration to specification testing. This is especially true if variables classified as endogenous could conceivably be treated as exogenous, since there are potential gains in efficiency by such a change. Such specification tests are referred to as Hausman (1978)

efficiency by such a change. Such specification tests are referred to as Hausman (1978) tests and it is natural to ask: How are classical significance tests and Hausman tests related? A general result which has a simple geometric interpretation can be established in answer to this question; and once the result has been stated for one scalar product, it is straightforward to extend it to another.

(iv) Finally, having established a general understanding of estimation and inference under TSFs and TFOs, it is useful to consider how the methods developed may be extended to more general situations. Here the notion of choosing estimators by minimisation of a squared length defined by a chosen scalar product is crucial in providing a natural intuition for extension into, for example, non-linear methods, M- and analogue estimators.

4. A logical course outline

4.1 *Linear estimation*

The course begins with classical linear theory and then moves to adaptations of linear theory in econometrics. The linear theory is based upon $y \sim N(\mu, \Sigma\sigma^2)$ with $\mu \in \mathfrak{X}$. The main theorems are outlined below, without proofs. The symbol $\mathbb{L}$ denotes the set of linear transformations on $\mathfrak{E}$.

Definition 1. Let $Q \in \mathbb{L}$ be n.n.d. The scalar product induced by Q on $\mathfrak{R}[Q]$ is $(.,.)_Q$, defined by

$$(Qx, Qz)_Q = (x, Qz) \qquad \forall x, z \in \mathfrak{E}.$$

When Q is p.d., $\mathfrak{R}[Q] = \mathfrak{E}$ and $(.,.)_Q$ is determined entirely on $\mathfrak{R}[Q]$ by

$$(x, z)_Q = (x, Q^{-1}z) \qquad \forall x, z \in \mathfrak{E}. \qquad \square$$

On $\mathfrak{R}[Q]$, $(x, Qx) = 0$ only when $x = 0$ and $(x, Qz) = (x, z)$ $\forall x \in \mathfrak{R}[Q]$ and $\forall z \in \mathfrak{E}$.

Definition 2. Let $\mathfrak{X} \subset \mathfrak{E}$ be a subspace and $Q \in \mathbb{L}$ be n.n.d. $\mathbb{P}(\mathfrak{X})$ is the set of all projections on $\mathfrak{X}$, that is if $P \in \mathbb{P}(\mathfrak{X})$

$$Px = x \qquad \forall x \in \mathfrak{X}. \tag{4}$$

$P \in \mathbb{P}_Q(\mathfrak{X})$ if $P \in \mathbb{P}(\mathfrak{X})$ and

$$P\mathfrak{R}[Q] \subset \mathfrak{R}[Q] \tag{5}$$

$$(z - Pz, x)_Q = 0 \qquad \forall z \in \mathfrak{R}[Q], \ \forall x \in \mathfrak{R}[Q] \cap \mathfrak{X}. \tag{6}$$ $\square$

Equation (4) implies $P = P^2$ and defines the set of projections from $\mathfrak{E}$ on $\mathfrak{X}$; (4)-(6) defines the set of orthogonal projections, relative to $(.,.)_Q$, from $\mathfrak{R}[Q]$ on $\mathfrak{X} \cap \mathfrak{R}[Q]$,

save when Q is p.d. whereupon the set is from $\mathcal{E}$ to $\mathfrak{X}$.

Theorem 1. (Philoche, 1971) *Let* $\mathfrak{X} \subset \mathcal{E}$ *be a subspace,* $Q \in \mathbb{L}$ *be n.n.d. and let* $P_0 \in \mathbb{P}_Q(\mathfrak{X})$. *Then* $PQP^T - P_0QP_0^T$ *is n.n.d.* $\forall P \in \mathbb{P}(\mathfrak{X})$.

□

Consider now the random vector Y which ranges over $\mathcal{E}$ according to the family of probability measures $\{\pi : \pi \in \Pi\}$. $E_\pi[Y] = \mu$ and $D_\pi[(Y, x), (Y, z)] = (x, \Sigma_0 z)$, $\Sigma_0 = \Sigma\sigma^2$, σ being scalar or $D[Y] = \Sigma\sigma^2$.

Theorem 2. (Philoche, 1971) *Let* $\mathfrak{X} \subset \mathcal{E}$ *be a subspace and,* $\forall \pi \in \Pi$, *let* $\mu \in \mathfrak{X}$ *and* $\Sigma \in \mathbb{L}$ *be n.n.d. Let* $P_0 \in \mathbb{P}_\Sigma(\mathfrak{X})$. *Then*

(i) $D_\pi[(P_0Y,x), (P_0Y,x)] = \inf_{P \in \mathbb{P}(\mathfrak{X})} D_\pi[(PY,x), (PY,x)]$, $\forall x \in \mathcal{E}, \forall \pi \in \Pi$;

(ii) *if* $P_0, P_1 \in \mathbb{P}_\Sigma(\mathfrak{X})$, *then* P_0Y *and* P_1Y *are almost everywhere equal,* $\forall \pi \in \Pi$.

□

Theorem 2 is a general version of the GM theorem on least squares. Thus P_0Y is the minimum variance linear unbiased estimator of $\mu \in \mathfrak{X}$, P_0 being the generalised least squares (GLS) projector on $\mathfrak{X}$, namely the orthogonal projection on $\mathfrak{X}$ relative to $(.,.)_\Sigma$. This may be shown using (1) and (2).

The next theorem answers the question: When is the OLS estimator of μ also the GM estimator?

Theorem 3. (Kruskal, 1968; Philoche, 1971) *Let* $P \in \mathbb{L}$ *satisfy* $(z - Pz, x) = 0$ $\forall z \in \mathcal{E}$ *and* $\forall x \in \mathfrak{X}$. $P \in \mathbb{P}_\Sigma(\mathfrak{X})$ *if and only if* $\Sigma\mathfrak{X} \subset \mathfrak{X}$.

□

If $x \in \mathfrak{X}$ and $\Sigma x \in \mathfrak{X}$, $\forall x \in \mathfrak{X}$, then $\Sigma\mathfrak{X} \subset \mathfrak{X}$ and $\mathfrak{X}$ is invariant under Σ. Theorem 3 may be illustrated using $\Sigma = [I_n(1-\rho) + ee^T\rho]$, $|\rho| < 1$, where e is the equi-angular line in $\mathcal{E}$, i.e. the (n×1) vector each of whose elements is unity.

A useful theorem for corroborating invariance is:

Theorem 4. (Halmos, 1958) *If* $\mathcal{V} \subset \mathcal{E}$ *is a subspace which is invariant under* $A \in \mathbb{L}$, *then* $PAP = AP$ *for every projection* P *on* $\mathcal{V}$. *Conversely, if* $PAP = AP$ *for some projection* P *on* $\mathcal{V}$, *then* $\mathcal{V}$ *is invariant under* A.

□

From Theorem 4 a number of computable conditions for checking invariance follow. Theorem 4 may be illustrated with various simple examples from econometrics, for example seemingly unrelated regressions.

4.2 *Linear hypothesis*

When the scalar product is $(.,.)$, $\|x\| = \sqrt{(x,x)}$ denotes the length of x; when the scalar product is $(.,.)_\Sigma$, length of x is $\|x\|_\Sigma = \sqrt{(x,x)_\Sigma}$. If $\mathcal{V}$ is a subspace of $\mathcal{R}(\Sigma)$, $V^\# = \{x: (x,z)_\Sigma = 0, x \in \mathcal{R}[\mathcal{E}], z \in V\}$.

Let $H : \mu \in \mathfrak{X} \cap \mathcal{R}[\Sigma]$ be the maintained hypothesis when it is desired to test $H_0 : \mu \in \mathcal{V}$, $\mathcal{V}$ being determined by $r < \dim\{\mathfrak{X} \cap \mathcal{R}[\Sigma]\} = k$ linearly independent restrictions $(a_i,\mu)_\Sigma = 0$, i = 1,2,...,r. Now

$$\mathfrak{R}[\Sigma] = \mathcal{V} \oplus [\mathcal{V}^{\#} \cap \mathfrak{X}] \oplus \mathfrak{X}^{\#}$$

and dim $\mathcal{V}$ = (k-r), dim $[\mathcal{V}^{\#} \cap \mathfrak{X}]$ = r and, setting dim $\mathfrak{R}[\Sigma]$ = n, dim $\mathfrak{X}^{\#}$ = (n-k).

$$(7) \qquad F = \frac{\|P_r y\|^2_\Sigma}{\|(I_n - P_0)y\|^2_\Sigma} \cdot \frac{n-k}{r} = \frac{\|(P_0 - P_{k-r})y\|^2_\Sigma}{r\hat{\sigma}^2}$$

where P_{k-r}, P_r and P_0 are the orthogonal projections, relative to $(.,.)_\Sigma$, on $\mathcal{V}$, $[\mathcal{V}^{\#} \cap \mathfrak{X}]$ and $\mathfrak{X} \cap \mathfrak{R}[\Sigma]$ respectively and $\hat{\sigma}^2 = \{\|(I_n - P_0)y\|^2_\Sigma / (n-k)\}$.

If $X\beta = X_1\beta_1 + X_2\beta_2$, X_1 having k_1 and X_2 having k_2 columns such that $X = [X_1 : X_2]$ has rank $k_1 + k_2 = k$, and if $\Sigma = I_n$, consider testing $H_0 : \beta_2 = 0 \Leftrightarrow \mu \in \mathfrak{X}_1, \Leftrightarrow F_2\mu = 0, \mu \in \mathfrak{X}$ where $\mathfrak{X}_i = \mathfrak{R}[X_i]$, P_i and M_i the orthogonal projections on $\mathfrak{X}_i$ and $\mathfrak{X}_i^{\perp}$ respectively, and $F_i = X_i(X_i^T M_j X_i)^{-1} X_i^T M_j$, $i \neq j$, $i,j = 1,2$. A corresponding hypothesis is $H_1 : (P_1 - F_1)\mu = 0, \mu \in \mathfrak{X}$ where $P_1 y = X_1\tilde{\beta}_1$, the OLS estimate of $X_1\beta_1$ on H_0, and $F_1 y = X_1\hat{\beta}_1$, the corresponding estimate on H. $H_0 \equiv H_1$ if and only if $\mathfrak{X}_1^{\perp} \cap \mathfrak{X}_2 = \phi$ and $k_1 \geq k_2$. When these conditions are satisfied

$$(8) \qquad \frac{\|(P-P_1)y\|^2}{k_2\hat{\sigma}^2} = \frac{\|M_1F_2y\|^2}{k_2\hat{\sigma}^2} = \frac{(\tilde{\beta}_1 - \hat{\beta}_1)^T[\hat{D}(\tilde{\beta}_1 - \hat{\beta}_1)]^{-1}(\tilde{\beta}_1 - \hat{\beta}_1)}{k},$$

$\hat{D}[.]$ denoting estimated dispersion on H. This is the basis of Durbin (1954) and Hausman (1978) specification testing. In all cases $\tilde{\beta}_1$ and $\hat{\beta}_1$ are consistent on H_0 but $\tilde{\beta}_1$ is more efficient; on $H_\alpha : \beta_2 \neq 0$, $\hat{\beta}_1$ is consistent but not $\tilde{\beta}_1$.

4.3 *Interdependent models*

The principal difference between the methods of Sections 4.1 and 4.2 and econometric methods is that in a linear model like $y = X\beta + \varepsilon$ the former requires X to be fixed or $E[\varepsilon|X] = 0$, whereas in econometrics X is bedevilled by interdependence with ε. This implies *inter alia* that asymptotic methods become necessary. The problem of interdependence is generally tackled by the use of instrumental variables (IVs). Let Z represent the n×K matrix of IVs, $Z = [X_1 : Z_2 : Z_3]$ the blocks having respectively k_1, K_2 and K_3 columns such that Z has rank $(k_1 + K_2 + K_3) = K$ and $K_2 \geq k_2$. The principal property of Z is that each element z_{ij} is uncorrelated with the "contemporaneous" element of ε, ε_i. Two approaches to making use of Z are considered: the first, called the GM approach, is due to Fisher (1966, 1972) and Dhrymes (1968), and the second, due to Sargan (1958, 1959, 1961), is called the GMM approach.

If all the IVs are used, three important examples of the GM approach are the generalised two-stage least squares (G2SLS) method (Theil, 1958), Madansky's generalised IV estimation (M-GIVE) method (Madansky, 1964) and Sargan's generalised IV estimation (S-GIVE) method (Sargan, 1958, 1959, 1961; Amemiya, 1966).

G2SLS: $\text{TSF} = Q:\ Q\Sigma Q^T = I_n;\ \text{TFO} = B_n^T = CZ^TQ^T:\ C(Z^T\Sigma^{-1}Z)C^T = I_K$

$A_n = \Sigma^{-1}Z(Z^T\Sigma^{-1}Z)^{-1}Z^T\Sigma^{-1} \Rightarrow \text{use } (.,.)_{A_n}.$

M-GIVE: $\text{TSF} = Q:\ Q\Sigma Q^T = I_n;\ \text{TFO} = B_n^T = CZ^TQ^{-1}:\ C(Z^T\Sigma Z)C^T = I_K$

$A_n = Z(Z^T\Sigma Z)^{-1}Z^T \Rightarrow \text{use } (.,.)_{A_n}.$

S-GIVE: $\text{TSF} = Q:\ Q\Sigma Q^T = I_n;\ \text{TFO} = B_n^T = CZ^T\ C(Z^TZ)C^T = I_K$

$A_n = QZ(Z^TZ)^{-1}Z^TQ^T \Rightarrow \text{use } (.,.)_{A_n}.$

In the GMM approach, the absence of correlation between the IVs and the equation errors suggests that such a property should hold in the corresponding sample moment $n^{-1}Z^T(y - X\beta)$. The basis of the method is to find an estimate $\hat{\beta}$ which minimises the quadratic length from the origin of $Z^T(y - X\beta)$ or some transformation of it or $(y - X\beta)$. Thus in M-GIVE $\|CZ^T(y - X\beta)\|^2$ is minimised, where $C^TC = (Z^T\Sigma Z)^{-1}$, and in S-GIVE $\|CZ^TQ(y - X\beta)\|^2$ is minimised where $C^TC = (Z^TZ)^{-1}$. GMM estimation may be looked at as applying a TFO before a TSF, not the other way round as in GM estimation.

The adequacy of IV specifications may be tested by application of (8). Suppose, for example, that $\tilde{\beta}$ is the IV estimator using all the z's, while $\hat{\beta}$ is the IV estimator using the subset $[X_1 : Z_2]$. There may be some doubt that the columns of Z are genuine IVs, whereas no such doubt is held for the subset. Let $P = Z(Z^TZ)^{-1}Z^T$ and P_0 be the corresponding projection matrix for the subset. Then if $PX = \tilde{X}$ and $P_0X = \hat{X}$, $\hat{P}$ and $\tilde{P}$ would be the orthogonal projections on $\mathcal{R}[\hat{X}]$ and $\mathcal{R}[\tilde{X}]$; let $\tilde{M} = (I_n - \tilde{P})$ and $\hat{M} = (I_n - \hat{P})$. Then, corresponding to (8) with $\Sigma = I_n$

$$\{\sqrt{n}\,(\tilde{\beta} - \hat{\beta})^T [(\hat{X}^T\hat{X}/n)^{-1} - (\tilde{X}^T\tilde{X}^T/n)^{-1}]^- (\tilde{\beta} - \hat{\beta})\sqrt{n}\} / \sigma^2$$

$$= y^T \tilde{M} \hat{X} (\hat{X}^T\tilde{M}\hat{X})^- \hat{X}\tilde{M} y^T / \sigma^2$$

$$= y^T (P_{\mathcal{Q}} - \tilde{P})\, y / \sigma^2,$$

where $P_{\mathcal{Q}}$ is the orthogonal projection on $\mathcal{Q} = \mathcal{R}[\hat{X} : \tilde{X}]$ and $(.)^-$ denotes any generalised inverse. Using σ^2 itself or σ^2 replaced by a consistent estimate of it, the test-statistic has a central χ^2-distribution on $H_0 : \text{plim}(\tilde{\beta} - \hat{\beta}) = 0$ with degrees of freedom determined by rank $(P_{\mathcal{Q}} - \tilde{P})$. On H_0, then, the "extra" IVs are valid. The test-statistic may be calculated by testing for the exclusion of $\hat{X}$ in the regression $y = \tilde{X}\theta + \hat{X}\lambda + v$, in the same way as in Durbin's (1954) test noted above.

Extensions to non-linear regressions of the kind $y = f(X;\beta) + \varepsilon$ may be obtained, when Σ is p.d. and $\varepsilon \sim (0,\Sigma)$, by considering

$$\inf_{\beta}\ [y - f(X;\beta)]^T A_n [y - f(X;\beta)]$$

where A_n is a scalar product matrix of the kind given in this section.

Finally, G2SLS, M-GIVE and S-GIVE are examples of choosing a method of estimation by choosing a scalar product. Various results are also possible concerning applications of Theorem 3.

5. Summary

The aim of the paper has been to demonstrate five qualities of a coordinate-free presentation of econometric theory. First, the structure is clarified, thereby simplifying the mathematics and reducing tedium. Second, relations between methods are exposed and given coherence. Third, the geometrical structure provides a powerful intuition. Fourth, extension to non-linear methods and GMM is straightforward. Fifth, a coordinate-free exposition allows a speedy development of theory, permitting more time for practical applications.

References

Amemiya, T (1966) Specification analysis in the estimation of parameters of a simultaneous equation model with autoregressive errors. *Econometrica* **34**, 283-306.

Dempster, A P (1969) *Elements of Continuous Multivariate Analysis.* Addison-Wesley, Reading, Mass.

Dhrymes, P J (1968) Alternative asymptotic tests of significance and related aspects of 2SLS and 3SLS estimated parameters. *Review of Economic Studies* **33**, 213-226.

Durbin, J (1954) Errors in variables. *Review of the International Statistical Institute* **22**, 23-32.

Drygas, H (1970) *The Coordinate-Free Approach to Gauss-Markov Estimation.* Springer-Verlag, Heidelberg.

Eaton, M (1970) Gauss-Markov estimation for multivariate linear models : a coordinate-free approach. *The Annals of Mathematical Statistics* **41**, 528-538.

Fisher, G R (1966) On generalized least squares. *Econometrica* (Supplement) **34**, 77.

Fisher, G R (1972) The algebra of estimation in linear econometric systems. *International Journal of Mathematical Education in Science and Technology* **3**, 385-403.

Halmos, P R (1958) *Finite-Dimensional Vector Spaces.* Van Nostrand, Princeton.

Hausman, J A (1978) Specification tests in econometrics. *Econometrica* **50**, 697-720.

Kruskal, W (1961) The coordinate-free approach to Gauss-Markov estimation, and its application to missing and extra observations. *Proceedings of the IVth Berkeley Symposium on Mathematical Statistics and Probability* **1**.

Kruskal, W (1968) When are Gauss-Markov and least squares estimators identical? A coordinate-free approach. *Annals of Mathematical Statistics* **39**, 70-75.

Madansky, A (1964) On the efficiency of three-stage least squares estimation. **32**, 51-56.

Philoche, J -L (1971) A propos du théorème de Gauss-Markov. *Annales de l'Institut Henri Poincaré* **7**, 271-281.

Sargan, J D (1958) The estimation of economic relationships using instrumental variables. *Econometrica* **26**, 393-415.

Sargan, J D (1959) The estimation of relationships with autocorrelated residuals by the use of instrumental variables. *Journal of the Royal Statistical Society B* **21**, 91-105.

Sargan, J D (1961) The maximum-likelihood estimation of economic relationships with autoregressive errors. *Econometrica* **29**, 414-426.

Seber, G A F (1964a) The linear hypothesis and idempotent matrices. *Journal of the Royal Statistical Society B* **26**, 261-266.

Seber, G A F (1964b) Linear hypotheses and induced tests. *Biometrika* **51**, 41-48.

Seber, G A F (1964c) The linear hypothesis and large sample theory. *The Annals of Mathematical Statistics* **35**, 773-779.

Seber, G A F (1980) *The Linear Hypothesis : A General Theory*. Griffin, London.

Theil, H (1958) *Economic Forecasts and Policy*. North-Holland, Amsterdam.