QUANTIFYING LONG RUN AGRICULTURAL RISKS AND EVALUATING FARMER RESPONSES TO RISK

Proceedings of a Seminar sponsored by
Southern Regional project S-~~180~~232
"Quantifying Long Run Agricultural Risks and Evaluating Farmer Responses to Risk"
Sanibel Island, Florida
April 9 - 12, 1989

Agricultural Economics Department
Texas A&M University
College Station, Texas

July 1989

THE MEASUREMENT OF ECONOMIC VARIABILITY WITH POOLED DATA: CONCEPTUAL AND METHODOLOGICAL ISSUES

John M. Antle*

Prices, quantities and other variables are often represented as random variables in economic theory. When researchers translate theoretical propositions into testable hypotheses they need to devise empirical characterizations of the distributions of economic variables. There are many ways to do this, but the usual method is to describe distributions using the means, variances, covariances, and possibly higher order moments. The objective of this paper is to discuss some of the concpetual and methodological issues that arise in quantifying these higher central moments of economic random variables.

One might well ask why estimation of higher moments is any different than estimation of the mean of a random variable. Indeed, from the point of view of general estimation theory there is no difference. For example, the maximum likelihood approach produces estimates of every characteristic of a distribution. But as a practical matter, applied econometrics has focused much more on the estimation of the conditional mean (as in least squares estimation) than on the estimation of higher moments. Moreover, when standard econometrics has been concerned with estimation of higher moments, the objective was to obtain more efficient estimates of the conditional mean (as in the estimation of a covariance matrix for a feasible generalized least squares estimator). It has only been since theory has shown that higher moments matter for analysis of behavior and policy that econometricians have become concerned with estimation of higher moments on an equal basis with conditional means.

Another reason to focus on the measurement of variability is that a variety of ad hoc methods have been used to measure variability, and often these methods have not been critically assessed. For example, a researcher may need an estimate of a price covariance matrix to use a risk programming model of a farm's cropping decisions. The available data may be time series at the county or state level. The researcher may therefore compute sample variances and covariances from the aggregate data. Or a policy maker may need to evaluate national policies over time using data from a large number of heterogeneous regions, and may use sample variances and covariances across regions.

Figure 1 classifies policy and data by level of aggregation. The cells in Figure 1 can be classified into three types. The cells FF and AA along the main diagonal represent those cases in which data are available at the same level of aggregation that is used in analysis. In this case the measurement of variability with pooled data can be addressed with a set of established estimation prodcedures. However, it must be emphasized that different kinds of variability are being measured at each level. At the farm level, pooled data capture variation withing

---

*John M. Antle is an associate professor of agricultural economics, Montana State University.

the farm population across farms and over time; at the aggregate level, pooled data represent variation across regions and over time. The cell AF below the diagonal represents the "aggregation problem" in which data are available at a disaggregate level but analysis is desired at the aggregate level. The cell FA above the diagonal represents what might be called the "disaggregation problem," as it involves the use of aggregate data to undertake economic analysis at a disaggregate level.

After presenting a stylized production model to be used in the analysis, the paper overviews some issues in the measurement of economic variability with pooled data for cases in which aggregation (or disaggregation) is not an issue (the cases on the diagonal of Figure 1). The remainder of the paper briefly addresses the additional problems posed by aggregation and disaggregation.

A Production Model and the Policy Problem

Variability is defined as the properties of the covariance matrix of a random vector. The definitional issues that arise in the debate over how "risk" should be measured are not addressed. Generalizations to other higher moments, and to mean-, variance-, and n^{th}-moment-preserving spreads are possible, but will not be discussed.

The problem addressed in this paper may be stated in relation to a disaggregate and a more aggregate level of data and analysis. For purposes of discussion, the disaggregate level is the individual decision maker, the farm, and the aggregate level is the region. At the farm level there are $i=1,\ldots,N_r$ units in region r, and there are $r=1,\ldots,R$ regions. Associated with each farm is a quantity of output y, input x, and acreage n. Each unit of land has associated with it an environmental attribute a.

The model of individual economic agents is based on the optimal allocation of land and other inputs into production as functions of prices, policies, technology, and other farm charcteristics (capital stocks, risk attitudes, etc.). Define p as the vector of output and input prices, ψ as a vector of policy parameters, and ω a vector of farm characteristics (if prices are unknown at the time decisions are made, then interpret p as a vector of price distribution parameters and interpret output as revenue). Define x_j^i as the input allocation of farmer i to acre j, x^i as the vector of the x_j^i, a_j^i as the environmental attribute of acre j managed by farmer i, and a^i as the vector of a_j^i. Define the indicator function δ_j^i such that

$$\delta_j^i = \begin{cases} 1 \text{ if acre j is in production} \\ 0 \text{ otherwise,} \end{cases}$$

and let $\delta^i = \{\delta_j^i\}$. Also define the vector of attributes of land in production on farm i as $a(\delta^i) = (a_1^i\delta_1^i, a_2^i\delta_2^i, \ldots)$.

The ith farmer's decision problem is

$$\max \quad J(x^i, a(\delta^i) | p, \psi, \omega^i)$$

where J is the farmer's objective function and ω^i is a vector of the ith farm's characteristics, including the production technology and the farmer's behavioral attributes. In addition to setting prices, policy may impose a set of inequality constraints on land use. For example, the total acreage in production on farm i is $n^i = \sum_j \delta_j^i$, and a diversion requirement for participation in commodity programs imposes a constraint on n^i. The solution to the above maximization problem generates the demand functions $x_j^i = x(p,\psi,\omega^i,a_j)$ and $\delta_j^i = \delta(p,\psi,\omega^i,a_j)$.

Define yield on acre j of farm i as $y_j^i = y(x_j^i,a_j^i,\varepsilon_j^i)$, where ε_j^i is a random variable representing weather, etc. The vector ω^i of farm characteristics and the vector ε^i of production shocks can be interpreted as random variables in the farmer population defined by a parameter vector θ. The parameter vector θ will change over time as farmers adopt new technology, make capital investments, and enter and exit the industry. Thus, while the distribution of ω is exogenous in the current decision period, in the longer run it may be endogenous to prices and policy parameters. The distribution of the random variables ω and ε induce a joint distribution for output, input, land use, and environmental attributes of land. Let this distribution function be $\phi(y,x,n,a|p,\psi,\theta)$.

Policy analysis can be conducted at the level of the indivdual firm or at a more aggregate level. Analysis of the welfare effects at the farm level of policy changes or other exogenous changes can be conducted using the firm's objective function. Substituting the optimal levels of x and δ into the objective function $J(x^i,a(\delta^i)|p,\psi,\omega^i)$ gives the function $J^*(p,\psi,\omega^i)$. The effects of changes in policy parameters on farm-level welfare can be evaluated using this function.

For analysis of welfare at the regional level, suppose that policy makers define aggregate welfare as a function of the population mean outputs and inputs and the variances and covariances of those variables. Define the column vector v as the stacked vector containing y, x, n, and a. Viewing v as a random vector in the population of farms in a region in each production period,

$$E(v|p,\psi,\theta) = \int v \, d\phi(y,x,n,a|p,\psi,\theta)$$

$$Var(v|p,\psi,\theta) = \int (v - E(v))(v - E(v))' \, d\phi(y,x,n,a|p,\psi,\theta)$$

are the mean vector and covariance matrix associated with these variables in the region at a point in time. These means, variances, and covariances are defined conditionally on the prevailing prices, policy parameters, and characteristics of the producer population. Assuming policy is concerned with the means, variances and coavariances, the policy problem can be defined as:

$$(1) \quad \max_{\psi} \; w[E(v|p,\psi,\theta),Var(v|p,\psi,\theta)].$$

For example, in setting pesticide policy, policy makers could be trading off mean productivity and variance of productivity against environmental quality.

Alternatively, policy analysis can be conducted using aggregates of outputs, inputs, and other variables. Define Y, X, L, and A as aggregates of output, input, acreage, and environmental attributes, and define the vector V composed of the aggregates. These aggregates can be viewed as random variables with mean $E(V) = M$ and variance $E[(V - M)(V - M)'] = Var(V)$. If policy makers base policy decisions on aggregates, the policy problem can be defined as:

$$(2) \quad \max_{\psi} \; W[E(V|\psi), Var(V|\psi)]$$

It is important to distinguish the policy problems in (1) and (2). The former problem requires information about the variability within the population at a given point in time and over time. Such information can be obtained from cross-sectional or pooled farm level data. The latter problem requires information about the variability in the aggregates. Variability in aggregates is measured across regions and over time, and can be obtained from pooled aggregate data. In general, the measures of variability obtained at one level of aggregation will not correspond to those obtained at another level of aggregation.

Measuring Economic Variability: Basic Issues

A set of basic issues arise in the measurement of economic variability. For simplicity, the general problem is reduced to that of estimating the covariance matrix of a single variable, say y_{rt}, $r=1,\ldots,R$ and $t=1,\ldots,T$, with covariance matrix Ω. A series of issues arise in estimation of this matrix with pooled data.

Dimensionality and Degrees of Freedom

Let the random variables y_{rt} be arranged in a (RTx1) vector $y = (y_1', y_2', \ldots, y_R')'$, where y_r is the (T x 1) vector of time series observations from region r. Thus $E(yy') = \Omega = [\Omega_{ij}]$ where $\Omega_{ij} = E(y_i y_j')$, $i,j = 1,\ldots,R$. Ω is the (RTxRT) matrix made up of the (TxT) submatrices Ω_{ij}. Alternatively, we can arrange the variables so that $y = (y_1', y_2', \ldots, y_T')'$, where the y_t are the (R x 1) vectors of cross-section observations from all R regions in period t. In this case we can define $\Omega_{st} = E(y_s y_t')$, $s,t\ j= 1,\ldots,T$ and set $\Omega = [\Omega_{st}]$. In either case, by the symmetry of Ω there are a total of $(R^2T^2 + RT)/2 - 1$ distinct parameters in Ω (one parameter is an arbitrary scale factor). Since there are only RT observations on the y_{rt} in a pooled data set, it is obvious that it is impossible to estimate a fully general covariance matrix, because the number of parameters exceeds the number of observations and increases with the number of observations. For example, if R=20 and T=10, RT=200 but there are 20,100 distinct parameters in the full covariance matrix. Clearly, some simplifying assumptions are needed to make the estimation of Ω possible.

In general, the dimensionality problem can be stated as follows. To define a covariance matrix one needs the means, variances, and

covarainces for every random variable. The total number of parameters to be estimated must not exceed the number of observations, so there must be less than RT parameters involved in defining the means, variances, and covariances.

Thus with pooled data it is possible to estimate a covariance matrix only if the heterogeneity and dependence of the random variables are restricted. It is possible to allow some limited degree of herogeneity or dependence in either the time dimension or the cross section dimension. Note that each block Ω_{ij} contains $(T^2+T)/2$ distinct elements and each block Ω_{rs} contains $(R^2+R)/2$ distinct elements. Thus, even under extreme simplifying assumptions, such as $\Omega_{ij} = \Omega_{oo}$ for all i and j, there are still likely to be a large number of parameters to be estimated, possibly more parameters than there are observations.

The above example suggests that one way to make estimation of a covariance matrix possible is to impose enough structure on it to reduce the number of free parameters. An alternative approach, used in the literature on heteroscedatic regression models, is to assume that elements of the covariance matrix are functions of a set of observable exogenous variables. If the model is linear, this assumption reduces the number of parameters to be estimated to the number of explanatory variables.

Heterogeneity and Dependence

Two basic properties of random variables are represented by a covariance matrix, heterogeneity and dependence. Heterogeneity occurs in both the time and spatial dimensions, and has to do with changes in the distributions defining the random variables. Dependence refers to the degree of relationship among random variables in either time or spatial dimensions.

Heterogeneity in the time dimension is known as the stationarity property of a time series. The sequence of random variables y_{rt}, $t=1,\ldots,T$ is said to be stationary if its joint probability distribution function is time invariant; it is said to be weakly stationary if the mean is constant and the autocovariance between y_{rs} and y_{rt} is a function of $|t-s|$. This discussion considers only the covariance matric so weak stationarity suffices.

The above discussion of dimensionality suggests why stationarity is important to time series analysis. If a stochastic process is weakly stationary, then there is one covariance matrix defining the "variability" of the process through all points in time. In the notation introduced above, the covariance matrix Ω_{tt} would be the same for all t, for example, greatly reducing the dimensionality of the estimation problem.

The difficulty with economic time series is that they tend to be nonstationary. For example, productivity changes over time with the introduction of new technology, so output or yield distributions are often nonstationary. How then do we deal with nonstationary series? We find some way to transform a nonstationary series into a stationary one. Time series analysts are fond of detrending and differencing data, for

example (the latter procedure leads to processes that are "integrated" in time series parlance). Adding time trends to econometric production models is another method to transform a nonstationary series into a stationary one. The problem with using such ad hoc, and one might say simplistic, techniques, is that they are not likely to capture the shifts in the conditional distribution caused by events such as technological change or policy change that do not typically occur uniformly over time. The consequence of working with nonstationary series is that estimates are biased.

Heterogeneity in variability in the cross-sectional dimension is heteroskedasticity. Conceptually, of course, one could define nonstationarity in the spatial as well as the time dimension. However, the case of heteroskedasticity usually refers to a situation wherein means (or conditional means) are constant but there are different variances (or, more generally, different covariance matrixes). Thus heteroskedasticity can be thought of as a case in which a random variable is nonstationary but only in the variance dimension.

Dependence in data means that there are nonzero off-diagonal elements in the covariance matrix (but note that the converse is not necessarily true). If we refer back to our example above with $\Omega = [\Omega_{ij}]$, $i,j = 1,\ldots,R$, then the off-diagonal elements of the Ω_{ii} matrixes represent the covariances over time in a given region, and the off-diagonal elements of the Ω_{ij}, $i \neq j$ are the covariances between time periods and regions. The diagonal elements in these matrixes measure the covariances between regions at the same points in time.

Dependence is important from an estimation point of view for at least two reasons. First, as we discussed above, it is not possible to estimate a covariance matrix without imposing structure on it, and one way to do so is by making assumptions about the form of dependence. The strongest assumption is that random variables are independent; a weaker assumption is that there is first-order autocorrelation in the time or spatial dimensions; a still weaker assumption is higher order autocorrelation. Such assumptions greatly reduce the dimensionality problem by substituting one or a few autocovariance parameters for a large number of covariances.

A second reason why dependence is important has to do with large sample properties of estimators. Asymtotically, the property of statistical indpendence has been generalized to the concept of ergodicity. In hueristic terms, two random variables are ergodic if they become "less dependent" as they get farther apart in time (or space) so that they are independent in the limit. It can be shown that ergodicity is necessary for general versions of central limit theorems to hold (see White). For this reason dependence plays a central role in large sample estimation theory.

Thus it can be concluded that it is necessary both for dimensionality reasons and for estimation purposes to impose restrictions on the forms of dependence in economic data.

Estimation Methods

There are two basic approaches to estimation of the moments of random variables. One is to estimate the moments directly, the other is to estimate the probability distribution function and then compute the moments of that distribution.

Method of Moments. The most obvious way to estimate population moments is to use the corresponding sample moments. This is the classical "method of moments" and under certain circumstances can be used to obtain estimates with desirable properties. The sample moments, scaled for degrees of freedom, are unbiased estimators. In some cases they correspond to maximum likelihood estimates and are therefore also efficient (Kendall and Stuart).

The disadvantage of the method of moments is that only a small number of the elements of a general covariance matrix can be estimated, and the estimated variances and covariances must be assumed to be constant across individuals and time. Consider, for example, a sample of observations on the variable y_{rt}. Defining y_r as the (T x 1) column vector of the y_{rt} for each region, a mean can be computed for each region. Using these means, sample variances and covariances $s_{ij} = y_i'A_T y_j$ can be computed for each region, where $A_T = I_T - (1/T-1)J_T$, with I_T the T-dimensional identity matrix and J_T the T-dimensional matrix containing 1 in each cell. Thus, for the covariance structure with $\Omega_{ij} = \sigma_{ij}I_T$, σ_{ij} a scalar, an estimate of the covariance matrix $\Omega = [\sigma_{ij}] \otimes I_T$ can be written as $S = [s_{ij}] \otimes I_T$. Note, however, that this estimate of Ω may not positive definite. The rank of A_T is (T-1), so the rank of the (R x R) matrix $[s_{ij}]$ cannot exceed the minimum of R and (T-1). Thus if R > (T-1), $[s_{ij}]$ is a singular matrix with rank (T-1) and thus S is singular with rank T(T - 1). In other words, it is possible to obtain a positive definite estimate of the covariance matrix $\Omega = [\sigma_{ij}] \otimes I_T$ only if the number of regions does not exceed the number of observations over time less one.

Least Squares Estimates of Moments. The least squares estimate of a model of the form $y = \mu_i(X,\beta_i) + u$ can be interpreted as an estimate of the conditional mean of y given X. Since it follows that $E(u^i)$ is the ith central moment of y, one can posit the model $u^i = \mu_i(X,\beta_i) + u_i$, $E(u_i) = 0$, and estimate this model by replacing u with the residual from the mean regression. The same principle can be employed to obtain estimates of covariances. This is the kind of procedure suggested in the literature on heteroscedasticity for variance estimation (see Judge et al.) and generalized by Antle (1983) to higher moments and covariances.

There are several advantages to this kind of procedure. First, the functional relationships between moments and exogenous variables can be directly estimated as part of the model. Second, as noted above, this approach reduces the number of parameters to be estimated and thus helps solve the degrees-of-freedom problem. Third, the estimates can be shown to have desirable large sample properties (consistency, asymptotic normality). The disadvantage of this approach is that these estimates of variances, covariances, and other higher moments can be shown to be biased in small samples.

The degrees of freedom issue also arises with this type of model. In the covariance structure $\Omega = [\sigma_{ij}] \otimes I_T$, if there are k_i parameters in the mean function for each region, then the number of regions R must be less than or equal to $(T-k_i)$ for the matrix $[\sigma_{ij}]$ to be nonsingular.

A more general case is where $\Omega = [\Omega_{ij}]$ and the blocks Ω_{ij} are diagonal heteroskedastic, with the variances specified as a function of exogenous variables with k_2 parameters, and that the covariances specified as a function of k_3 parameters. For the estimate of Ω to be nonsingular it is necessary that $k_2+k_3 \leq R(T-k_1)$.

The Autoregressive Conditionally Heteroskedastic (ARCH) Model. Engle has extended the heteroskedastic model to model time series so that the conditional variance of the series depends on past realizations. Suppose for example that y_t is distributed $N(x_t\beta,h_t)$, where

$$h_t = \pi_0 + \pi_1 y_{t-1}^2$$

Thus the variance of the process evolves over time as a function of past realizations of y_t. Engle shows that this type of model can be estimated using simple regression methods but the estimates are inefficient; efficient estimates can be obtained using maximum likelihood procedures. For pooled data the linear ARCH model could be extended by defining y_t as a vector of cross-sectional observations y_{rt} such that y_t is distributed $N(x_t\beta,H_t)$, where the (R x R) matrix H_t has elements

$$h_{rmt} = \pi_{rm0} + \pi_{rm1} y_{rt-1} y_{mt-1}.$$

Parametric and Nonparametric Distribution Estimation. Another approach is to fit a parametric distribution to data, using one of many possible methods, and to then use the estimated distribution function to numerically compute variances and covariances. For example, maximum likelihood methods can be used to fit general classes of distributions, such as the Pearson system (Kendall and Stuart). It is also possible to fit a function as an approximation to an unspecified distribution function (Taylor). There are also a variety of methods being developed that fit empirical distributions without maintaining any distributional assumptions known in statistics as distribution-free or nonparametric methods. These methods have the advantage of not requiring that any parametric restrictions be imposed on the form of the distributions being estimated. However, they have several disadvantages. First, they are generally computationally burdensome, especially if multivariate distributions are involved. Second, these methods are not well suited for estimating conditional distributions of the type that economic random variables usually involve.

An Example Using Variance Components

Consider the following example from Antle (1989). The model takes the form

$$(3) \quad \beta_{0t} + \sum_{i=1}^{m} \beta_i D_{ijt} = u_{jt},$$

where

$$u_{jt} = \varepsilon_{0jt} - \sum_{i=2}^{m} \varepsilon_{ijt} D_{ijt}, \quad E(\varepsilon_{ijt}) = 0.$$

The first equation represents a model of the conditional mean of a random variable which is a function of variables D_{1jt}, with $E(u_{jt}) = 0$.

With pooled data it is possible to estimate several block covariance structures, although it is not possible to estimate a full (unrestricted) covariance matrix because of the dimensionality constraints discussed above. It is possible, for example, to estimate a model which has a block-wise heteroskedastic structure with nonzero off-diagonal terms in each block. Each block can refer either to a time period or to an individual. For example, it could be assumed that each observation is correlated across time but independently distributed across individuals. Thus, each block in the covariance matrix refers to an individual, and covariances within each block are between observations over time for that individual.

Let model (3) be specified with $\beta_{0t} = \beta_{00} + \beta_{01}d_t$, where d_t is a vector of time dummy variables to measure year-specific changes in the intercept. Defining $\beta_1 \equiv 1$ in equation (1), the model can be written in the vector form as

$$D_{1j} = D_{2j}\beta + u_j, \quad j=1,\ldots,N,$$

where D_{1j} is a (T x 1) vector, D_{2j} is a (T x g) matrix of the $-D_{1jt}$ for farm j, β is the corresponding (g x 1) vector of parameters, and u_j is the (T x 1) vector of the u_{jt}. Stacking the D_{1j} and the u_j into (NT x 1) vectors D_1 and u, and the D_{2j} into an (NT x g) matrix D_2, the model can be written

$$D_1 = D_2\beta + u.$$

The covariance matrix of u is $E[uu'] = \Omega$, an (NT x NT) matrix. Following the assumption that observations are independent across individuals but correlated over time, Ω has the block diagonal structure $\Omega = \text{diag}[\Omega_1,\ldots,\Omega_N]$ where Ω_j is a (T x T) matrix. In order to estimate a distinct term for each element of these blocks, it can be assumed that each variance and covariance term can be decomposed into an overall effect, a time effect, and a farm effect. That is, defining the (t,u) element of the jth block as Ω_{jtu}, it was assumed that $\Omega_{jtu} = \tau + \tau_{tu} + \tau_j$, where τ is a scalar parameter measuring the overall effect, τ_{tu} is a scalar parameter representing the time effect, and τ_j is a scalar representing the farm effect. Note that since each block Ω_j is a symmetric (T x T) matrix, there are $(T^2+T)/2-1$ distinct time effects but just N-1 distinct farm effects.

To estimate and test the block covariance structure, a consistent estimate of the error vector u can be computed from a consistent estimator β^c as $u^c = D_1 - D_2\beta^c$, where β^c is obtained from a procedure such as least squares regression. Taking products of the residuals corresponding to the nonzero terms in the covariance matrix Ω, regressing them on a constant term, a set of time dummy variables, and a set of farm dummy variables, yields consistent estimates of the parameters τ, τ_{tu}, and τ_j, and hence of the elements Ω_{jtu} of the covariance matrix. To test the validity of the specification, Wald tests for the significance of the time effects (all $\tau_{tu}=0$), of the farm effects (all $\tau_j=0$), and both time and farm effects, can be constructed. Note that when neither farm nor

time effects is significant, the data do not support the block heteroskedastic covariance specification.

If the covariance matrix is not block heteroskedastic, it may be diagonal heteroskedastic. A test of this specification can be derived from the equation below (3). Squaring u_{jt} and taking its expectation,

$$E(u_{jt}^2) = \sigma_{ojt} + \sum_{k=2}^{m} \sigma_{okjt} D_{kjt} + \sum_{i=2}^{m} \sum_{k=2}^{m} \sigma_{ikjt} D_{ijt} D_{kjt}$$

where

$$E(\varepsilon_{ijt}\varepsilon_{kjt}) = \sigma_{ikjt}.$$

Several assumptions need to be made to facilitate estimation: $\sigma_{okjt} = 0$; the intercept $\sigma_{ojt}=\sigma_{ot}$ can be assumed to vary with time by including an intercept and time dummies in the equation; and it can be assumed that the covariances σ_{ikjt} are equal across individuals, hence $\sigma_{ikjt} = \sigma_{ik}$. Regressing squared residuals e_{jt}^2 on an intercept, time dummies, and terms $D_{ijt}D_{kjt}$, i,k = 2,...,m, yields consistent estimates of the parameters of the above equation under these assumptions. Fitted values of this regression are thus consistent estimates of the diagonal terms in the covariance matrix Ω under the hypothesis that it is diagonal heteroskedastic. This hypothesis can be evaluated with a Wald test for the significance of the regression's slope coefficients.

Aggregation and Disaggregation

Following Stoker (1982), aggregation can be defined as the 'adding up' of attributes of individuals in the population to obtain summary statistics for the population which can not be differentiated by the individual outcomes that are aggregated. Following Antle (1986), it is shown in this section that aggregate measures of variability can be derived that are analogous to their microeconomic counterparts. The aggregate functions depend on the parameters defining the distribution of individual attributes in the population of producers. An implication of this analysis is that the aggregate relationships will change when the producer population's attributes change, just as their disaggregate counterparts do.

Despite these similarities between the general structure of the disaggregate and aggregate models, direct connections between the farm level and aggregate variability can be established only under stringent

conditions. For example, if linear aggregation is possible, the aggregate variables are constructed so that they behave asymptotically like sample means. A central limit argument can then be used to relate variances and covariances of individuals to those of the aggregates. Letting Y be aggregate output, and if χ and σ_y^2 are the population mean and variance of farm-level output, then the distribution of Y has mean χ and variance σ_y^2/N, where N is the population size. In general, however, this kind of relationship between the two levels of aggregation does not exist.

Define the distribution of farm characteristics in the population as $A(\omega|\theta)$, so that the mean of ω is

$$\zeta \equiv \int \omega \, dA(\omega|\theta) = b(\theta).$$

Recall that the joint distribution of yield (y), input (x), acreage (n), and environmental attributes (a) is given by $\phi(y,x,n,a|p,\psi,\theta)$. The expected farm output, input, acreage, and environmental attributes in the population are therefore

$$\chi = \int ny \, d\phi(y,x,n,a|p,\psi,\theta)$$

$$\mu = \int nx \, d\phi(y,x,n,a|p,\psi,\theta)$$

$$\lambda = \int n \, d\phi(y,x,n,a|p,\psi,\theta)$$

$$\alpha = \int a \, d\phi(y,x,n,a|p,\psi,\theta).$$

Recall that the aggregates of output, input, acreage, and environmental attributes are Y, X, L and A, and that V is a vector of these aggregates. The aggregate variables are assumed to be functions of the firm-level quantities y, x, n and a. Therefore,

$$E(V|p,\psi,\theta) = \int V \, d\phi(y,x,n,a|p,\psi,\theta) \equiv M(p,\psi,\theta).$$

By the same logic,

$$E[(V-M)(V-M)'] = \int (V-M)(V-M)' \, d\phi(y,x,n,a|p,\psi,\theta) \equiv \Sigma(p,\psi,\theta).$$

The previous two equations demonstrate that the aggregate means and covariance matrix can be expressed as functions of the vectors of prices, policy parameters, and firm characteristics. Thus, in general, changes in any of these vectors induce changes not only in the means but also in the elements of the aggregate covariance matrix.

Assume the aggregate data satisfy

$$\text{plim } Q = \chi, \ \text{plim } X = \mu, \ \text{plim } L = \lambda, \ \text{plim } A = \alpha, \ \text{plim } Z = \zeta,$$

where Z is a vector of aggregate firm characteristics. Also define a sample estimator of S such that plim $S = \Sigma$, and assume further that the function $b(\theta)$ is invertible such that the function $\theta = h(\zeta)$ exists. Then using the above results, aggregate functions can be defined such that, for data aggregated over a large number of firms,

$$V \approx M(p,\psi,\theta) = M(p,\psi,h(\zeta)) \approx M(p,\psi,Z)$$

$$S \approx \Sigma(p,\psi,\theta) = \Sigma(p,\psi,h(\zeta)) \approx \Sigma(p,\psi,Z).$$

Thus, if aggregates converge to the population means, it is possible to define aggregate supply and factor demand functions that can be expressed as functions of price, policy parameters, and observable aggregate characteristics of the firms in the industry. Similarly, there exists a covariance matrix of the aggregate variables, and the aggregate variances and covariances can be expressed as functions of these same variables.

Nonstationarity and Detrending

The discussion of the previous sections established that variability measures have a similar structure whether they are defined for the population or for aggregates. The conditional property of distributions of economic variables means that their variances and covariances generally are functions of the exogenous "forcing variables" in the system such as prices and policy parameters. This fact means that economic random variables are likely to be nonstationary unless their "conditionality" is modeled correctly.

By way of illustration, suppose that aggregate time series data are used to estimate the variance of output. A typical procedure (e.g. Hazell) is to detrend the data and then compute the sample variance from the detrended data. Recalling from the previous section that Y is a function of prices, policy variables, and population characteristics, it can be seen that the Y series is nonstationary. Detrending removes the nonstationarity if these explanatory variables are all linear functions of the same trend. If they do not meet this stringent condition there is a different population mean χ_t and variance σ_{yt}^2 for each time period, and the sample variance based on the detrended data is a biased estimator. Letting Y be the sample mean of the data, the expectation of the sample variance is:

$$E\left[(1/T)\sum_{t=1}^{T}(Y_t - Y)^2\right]$$

$$= E\left[(1/T)\sum_{t=1}^{T}\{(Y_t - \chi_t) + (\chi_t - Y)\}^2\right]$$

$$= \sum_{t=1}^{T}\sigma_{yt}^2/NT + E\left[(1/T)\sum_{t=1}^{T}\{(Y_t - \chi_t)(\chi_t - Y) + (\chi_t - Y)^2\}\right]$$

The second term on the right hand side of the above equation is generally nonzero if the detrending does not remove the nonstationarity from the series. The second term in the summation is positive. The first term in the summation is nonzero if deviations of Y from its mean are systematically related to deviations of χ_t from Y. In many cases there may be no such relationship and it can be concluded that the sample variance is a biased estimator for the σ_{yt}^2 and is an upward biased estimator of the average of the σ_{yt}^2.

This result, and other similar results that can be developed for covariances, suggests that researchers be cautious in estimating and interpreting measures of variability in aggregate data.

Conclusions

This paper addresses some of the conceptual and methodological issues that arise in measuring economic variability with pooled data. Variability was defined as the covariance matrix of a random vector, and was assumed to be important to policy analysis through a policy objective function depending on population means and the covariance matrix of economic variables. A model of production decision making was used to illustrate the analysis. This model shows that economic variables are defined with distribution functions conditioned on vectors of prices, policy variables, and characteristics of the producer population. The distinction was drawn between variability within the population and variability in aggregate data across regions and over time.

Estimation of a covariance matrix was shown to require the imposition of restrictions on the heterogeneity and dependence of the random variables, to reduce the dimensionality of the estimation problem and to achieve estimators with desirable properties. These are standard statistical results.

Aggregate covariance matrixes were shown to be functions of the prices, policy parameters, and population characteristics with a structure similar to their farm-level counterparts. An important implication for estimation is that aggregate relationships are likely to be nonstationary. Unnless care is taken to adequately model the processes generating aggregate data, biased estimates of variability will be obtained. It is not likely to be adequate, for example, to simply detrend data and compute sample variances and covariances as estimates of variability and covariation.

The conclusion to be drawn from the analysis presented in this paper is that measurement of variability requires as much attention to development of a theoretical model and to its empirical implementation as does the more familiar problem of measuring a conditional mean. The constraints imposed by the dimensionality problem require that the researcher carefuly weigh the tradeoffs involved in the modeling assumptions that are made.

		Data	
		Farm level	Aggregate
Policy	Farm level	FF	FA
	Aggregate	AF	AA

Figure 1. Classification of Policy and Data by Level of Aggregation

Summary of Notation

subscript j denotes an acre of land
subscript t denotes time
superscript i denotes a farm in a region
subscript r denotes a region

y	yield
x	quantity of input
δ	indicator function, δ=1 if acre is in production, 0 otherwise
n	number of acres in production
a	environmental attribute of an acre
p	price vector
ψ	policy parameter vector
ω	vector of farm characteristics
ε	random production disturbance
v	vector of y,x,n and a
ϕ	distribution function of v
w	policy function defined over moments of v
Y	aggregate output
X	aggregate input
L	aggregate land in production
A	aggregate land attribute
V	vector of Y,X,L and A
M	expectation of V
W	aggregate policy function
$\sum$	covariance of V
S	estimate of $\sum$
Ω	covariance matrix
σ	variance or covariance
u	random error
μ_i	ith moment
β	parameter vector
τ	parameter
ζ	expectation of ω
χ	expectation of farm output
μ	expectation of farm input
λ	expectation of n
α	expectation of a
$\bar{Y}$	sample mean of Y

References

Antle, J.M. "Aggregation, Expectations, and the Explanation of Technological Change." Journal of Econometrics 33(1986):213-236.

_____. "Nonstructural Risk Attitude Estimation." American Journal of Agricultural Economics 71(1989): in press.

_____. "Testing the Stochastic Structure of Production: A Flexible Moment-Based Approach." Journal of Business and Economic Statistics 1 (1983): 192-201.

_____, and R.E. Just. "Effects of Commodity Policy Structure on Resources and the Environment." In N. Bockstael and R. Just, editors, Commodity and Resource Policy in Agricultural Systems. (Springer-Verlag, in press). Paper prepared for presentation at the Commodtiy and Agricultural Resource Policy Conference, Baltimore, Maryland, May 5, 1989.

Engle, R.F. "Autoregressive Conditional Heteroskedasticity with Estimates of the Variance of United Kingdom Inflation." Econometrica 50(1982):987-1008.

Hazell, P.B.R., "Sources of Increased Instability in Indian and U.S. Cereal Production." American Journal of Agricultural Economics 66(1984):302-311.

Judge, G.G., et al. The Theory and Practice of Econometrics. (New York: John Wiley and Sons, 1985).

Kendall, M. and A. Stuart. The Advanced Theory of Statistics. (New York: Macmillan Co., 1976).

Taylor, C.R. "A Flexible Method for Empirically Estimating Probability Functions." Western Journal of Agricultural Economics 9(1984):66-75.

While, H. Asymptotic Theory for Econometricians. (New York: Academic Press, 1984).