Estimating Multiple-Output Production Functions for the CLASSES Model

Chris Barrett and Heidi Hogset (Cornell University)
July 2003

This informal memo is meant to help the BASIS CRSP project team think through some of the econometric and data issues involved in estimating the multiple-output production functions we will use to calibrate the CLASSES model for different sites. Much of this will be extremely basic to most of you. We have tried to write this in a style accessible to any team member likely to work with the data on these estimations. We hope this memo will ignite discussion and collaboration among the team on these tasks.

There will be small differences across sites due to differences in data and the underlying agroecologies. A key objective of this memo is to try to standardize, to the maximum degree that is feasible and appropriate, the methods we are using in each site so that differences in estimates are more likely to reflect true differences rather than be artifacts of the statistical methods employed. Because of the considerable differences across sites, we need to estimate production functions specific to each site from the relevant data. We will not pool the data across sites.

Intercropping is common in our Kenya and Madagascar research sites. Yet intercropping is qualitatively different from some kind of "mixed monocropping". The fact that two or more crops are raised together on the same patch of land generates interactions between the crops that may be synergistic, antagonistic, or both. For example, the crops may compete for light and soil moisture, but may also help create a favorable micro environment for each other, increasing accessibility of soil nutrients, or decreasing the severity of pest infestations, etc. Moreover, for most inputs, it is impossible to target input use to one crop, excluding the other(s). Most inputs are applied jointly, and their effects on the respective crops will depend not only on how each crop responds to the input individually, but also on how the input affects the interaction between them. Thus, for our purposes, intercropping entails a single crop production activity, producing multiple outputs.

In economics, multiple-output production functions are often estimated in the dual form, i.e., represented by their cost or profit functions (e.g., R.D.Weaver *AJAE* 1983). But where production is mostly for subsistence, and only a small fraction of agricultural inputs and outputs are traded, that approach may be inappropriate because dual methods rely fundamentally on input and output prices, which are commonly not observed. Imputing prices where no transactions are observed will necessarily introduce an errors in variables problem. Indeed, even just using market-level observations that are not specific to the individual farmer will introduce errors in variables problems due to variability in market prices and in transactions costs across farmers. For multiple reasons, then, dual approaches are probably inadvisable in smallholder agriculture applications.

In calibrating the CLASSES model, we therefore want to specify and estimate multiple-output production functions in primal form, i.e., conversion functions that map inputs into outputs. The primal approach nonetheless suffers its own potential problems of endogeneity of inputs. Observed input application rates commonly respond to unobserved production shocks (e.g., people work less on a field that has been trampled and destroyed by wildlife, so regressing output on labor would generate a spurious positive relation between the variables, biasing upwards estimated output elasticities with respect to labor). So we need to take some care to instrument for highly endogenous inputs – those allocated after planting – in estimating primal production functions so as to minimize the endogeneity problem. We will come to this below.

Multiple-Output Production Functions

A direct and easily interpretable way to estimate a multiple-output production is to formulate the production system as a system of equations (e.g., of equations that are linear in their estimable parameters), and estimate them simultaneously (e.g., by feasible generalized least squares). However, this is computationally demanding, and may be infeasible with our data (based on Heidi's quick attempts with 2002 Madzuu production data). An alternative is to estimate a system of linear equations by estimating it equation by equation. But that entails the implicit assumption that all disturbances are independent, which does not make sense for intercropped crops.

An easier and more elegant way to do it may be the ray production function approach recommended by Mickael Löthgren (1997 *Economic Letters* 57: 255-259 … a copy is attached), which is closely related to the distance function with which some of you may already be familiar (R. Shephard 1970, *Theory of Cost and Production Functions*). This approach offers a reasonably straightforward generalization of the standard single output stochastic frontier production model.

The ray production function approach is pretty simple; it sounds more complicated than it is. Total output quantity (i.e., the combination of all products, 1…p) is represented by the Euclidian norm (i.e., distance) of the quantities, and a direction measure (i.e., the polar coordinates, or angle, related to the ratio of one crop output to the others) for all but a reference output. More specifically, the transformation of the observed outputs into a composite output – the norm – to be used as a dependent variable is done this way:

$$y = \|y\| m(\theta) \tag{1}$$

where

$$\|y\| = \sqrt{\sum_{i=1}^{p} y_i^2} \tag{2}$$

and

$$m_i(\theta) = y_i / \|y\| = \cos(\theta_i) \prod_{j=0}^{i-1} \sin(\boldsymbol{\theta}_j) ; \sin(\theta_0) = \cos(\theta_p) = 1 \tag{3}$$

which implies

$$\theta_i(y) = \cos^{-1}(y_i / \|y\| \prod_{j=0}^{i-1} \sin(\boldsymbol{\theta}_j)) \tag{4}$$

For a system with two outputs, that means:

$$m_1(\theta) = \cos(\theta_1) = y_1 / \sqrt{y_1^2 + y_2^2} \tag{5}$$

$$m_2(\theta) = \sin(\theta_1) = y_2 / \sqrt{y_1^2 + y_2^2} \tag{6}$$

$$\cos(\theta_2) = 1 => \theta_2 = \frac{\pi}{2} \tag{7}$$

The Euclidian norm, $\|y\|$, is used as dependent variable in the multi-output production function regression, while the polar-coordinate angles, which serve as directional variables, θ_i, are included in the vector of regressors, with the exception of θ_p. In the case of just two outputs, θ_2 is therefore fixed and omitted, leaving just θ_1 is the only directional variable included as a regressor. This transformation of the data implies that the production of the two intercropped crops is viewed as inseparable - this is one production process (with only one residual or disturbance) resulting in two outputs. The Euclidian output vector norm lets any number of outputs be represented by one single number representing total output, and a set of polar-coordinate angles indicating the relationship between the components and the total. Note that the estimated multi-output distance measure reflects an iso-output surface along which an infinite combination of the p products could be produced simply by varying the output ratio between crops. Since the CLASSES model will use composites anyway (e.g., "field crops" comprised of beans, maize and sorghum) within which we are not looking to model tradeoffs or reallocations of inputs (or investment), this has no significant downside for our purposes.

With this definition of the multi-output dependent variable, estimation of the production function follows the very familiar procedure used for single output production functions: just specify a functional form and the independent variables, then estimate, do diagnostic tests to ensure core assumptions hold with respect to the properties of the regression residuals, then conduct inference (i.e., hypothesis testing). For the purposes of calibrating CLASSES, we are mainly interested in just obtaining good estimates for the multi-output production function, which can be written as (dropping subscripts)

$$\lVert y \rVert = f(x,\theta,\varepsilon) \qquad (8)$$

where x is a mx1 vector of inputs, θ is the (p-1)x1 vector of polar coordinates, and ε is the stochastic regression disturbance.

Panel Data Estimation of Production Functions

With the exception of the Baringo, Embu and Marsabit cropping data, we have multiple observations over time on each household. In most cases, this also means multiple observations on each plot. One of the helpful characteristics of panel data (multiple observations in time on the same cross-sectional units) is that they permit us to control for unobserved but time invariant characteristics of observational units. For example, if one farmer is especially skilled and another is especially inept, these characteristics were the same (or nearly so) in each period for which we have data and – most importantly – these farmer-specific, unobserved characteristics are correlated with observed variables (e.g., the more skilled farmer uses inorganic fertilizer and more labor, the less skilled one uses no fertilizer and less labor), then by incorporating a farmer-specific dummy variable we can control effectively for these time invariant skill differences (and everything else about the farmer that doesn't change over time that we haven't already controlled for). Without that control for what are commonly termed "fixed effects", our estimates of the production function parameters will generally be biased and inconsistent. For plot-specific data, we can go one better, in that plot-specific dummy variables (which render household-specific dummy variables redundant … we don't use both) control for time invariant soil, light and hydrological characteristics of plots.

So for the data for which we have multiple observations over time, sample size permitting, we want to use fixed effects estimation to control for time invariant unobserved heterogeneity across plots and farmers. That means we add a dummy variable taking a value of one for each observation on a given unit for which we have multiple observations, and value zero otherwise. We drop the constant term, since each unit will effectively have its own intercept estimate. If sample size is too small, we may have to go to village-specific fixed effects, to capture local soils, pest and hydrological conditions. (For those of you who know the difference between fixed and random effects in panel data estimation, we could go either way here – equivalently, do both and test which fits the data best – but fixed effects are simpler for those less familiar with these methods, I think. Hence this direction here.)

Because we pool observations from multiple periods together in estimating panel data production models, we also have to worry about intertemporal variation in key factors of production, such as rainfall, temperature, cloud cover, etc. We obviously cannot observe these variables for each production unit, but on the assumption that they're reasonably covariate among all production units within a location in a given year, we can simply include a year-specific dummy variable (e.g., a dummy for 2002, which effectively shifts the production function up or down for 2002 relative to the first year's observations).

Inputs and Instruments

Intercropping entails raising several crops on the same piece of land. Thus, many inputs cannot be allocated to any one crop in isolation within the system. This is true for application of labor to pre-planting field preparations, as well as weeding labor and application of fertilizers and pesticides. However, only maize seeds produce maize seedlings, and only bean seeds produce bean seedlings, etc. The crops also do not necessarily reach maturity simultaneously, and they may be harvested at different times. Thus, harvesting labor may be allocated to the crops individually. So it makes sense to think of some inputs as output-specific, while others are not. If we have data disaggregated by crop, it is easy enough to test the hypothesis that one can aggregate inputs across crops on the plot (and thereby conserve degrees of freedom in estimation). We should do this where possible.

The key inputs are land (plot size), labor (hours as well as plot manager experience and education), nonlabor variable inputs (animal traction, inorganic fertilizer, manure, pesticides and other chemicals) and exogenous environmental factors (rainfall, slope, soil type, etc.), in addition to the p-1 output polar coordinates, $\theta_1 , \ldots, \theta_{p-1}$. Our surveys asked detailed questions on the environmental variables in order that we can control for these adequately since their omission typically biases production function parameter estimates otherwise (see, for example, the attached 2002 *JDE* paper by Sherlund et al.).

Given the focus of our project on the interactions between smallholder management of soils and welfare dynamics, it is especially important for us to include measures of environmental conditions in our production function estimates. There is information about the respondents' own assessment of soil fertility in each data set. There is good evidence from elsewhere in Africa that farmers' subjective assessments of soil quality match soil chemistry test results reasonably accurately (see February 2003 special issue of *Geoderma*). So we can use simple categorical variables for soil quality (good, average, poor). Soil tests are being run for most plots in our samples this summer, so we can subsequently go back and replace these coarse, farmer-reported variables with a continuously measured soil chemistry indicator variable, such as soil organic matter (SOM). However, there is no particular reason to wait on those results at this stage, although site-specific teams should avail themselves of those data as they become available. The data also contain information on slope, the presence of soil and water conservation structures, plot use history, and other covariates that affect erosion, soil moisture, etc. Take advantage of these data.

Besides omitting relevant variables – such as unobserved farmer or plot characteristics for which we can control in panel data using fixed effects, or environmental variables for which we should have reasonable observations – the biggest problem we are likely to have with data on inputs arises from endogeneity. In so far as farmers choose their inputs partly in response to unobserved conditions or shocks that will cause observed output to deviate from predicted output, this will violate the standard regression assumption that the independent variables are statistically independent of the residuals. The inputs are themselves endogenous and, as a consequence, parameter estimates will be biased and inconsistent. The standard approach for fixing this problem involves instrumental variables (IV) estimation. The most commonly used form of IV estimation, and the variant we will use, is two stage least squares (2SLS). In 2SLS, one regresses the endogenous independent variable on a vector of exogenous "instruments", then uses the fitted value from that regression as an independent variable in a second regression of the dependent variable of interest (in our case, $\|y_i\|$) on the independent regressors. The first stage strips the independent regressor of its endogenous component. The basic principle of IV estimation is that one needs instruments that are independent of the regression error term (i.e., truly exogenous or at least predetermined) and that one would not believe to have their own independent association with the dependent variable of interest. This gives us an instrumenting equation $x=g(z)+\varphi$, where z is a vector of instruments. We then use the predicted value of the endogenous independent variable (in this case hours worked) from the instrumenting equation, $x^*=g(z)$, as a regressor in the production function regression. In general, one wants to use as many legitimate instruments as possible so as to maximize goodness of fit. The better the fit in the instrumenting equation, the less loss of precision we suffer from stripping away some of the variation in the endogenous regressor. So don't worry about t-statistics on instruments in the first stage equation; if a variable makes a legitimate instrument and increases the regression r^2, include it in the instrumenting equation.

In our case, there will be at least two inputs for which we will want to instrument; human labor and animal traction labor. For labor effort (hours worked per plot), one reasonable instrumenting equation for which we should have appropriate data in each site would use as instruments household composition (# adult males, # adult females, # children), total area cultivated (not area of the plot, which is an independent regression in the production function), total number of plots cultivated, total number of livestock owned, and dummy variables for off-farm or non-farm work by type (e.g., salaried employment, agricultural day labor, etc.). The off-farm/non-farm dummies capture employability elsewhere, which affects on-farm labor allocation. We don't use actual hours worked in these off-farm/non-farm uses (where we have those data) because then we have drifted back into the endogenous labor allocation problem. Animal traction (hours) would use the same sort of instrumenting regression.

Since fertilizer application in most of these systems is done mostly early in the plant growth cycle and because relatively few people use inorganic fertilizer anyway, it likely causes negligible problems of simultaneity bias. We should try to include it, although when the subsample of fertilizer users is small, the parameter estimates become highly unstable, so don't be surprised by bizarre estimated fertilizer responses. We have manure purchase data and we know if people apply manure from their own animals. These data are probably best used as simply a dummy variable indicating manure application, although feel free to experiment with using a combination of herd size and purchases, if you have sufficient degrees of freedom to play with in a given site. Plot size is determined at the start of the season and can reasonably be taken as exogenous to output realizations for a season. Same for education and experience. Chemical applications appear rare in our sample and are likely, given financial liquidity constraints, to be responsive to shocks. With the exception of some cash crops for which inputs are provided on credit as an advance against the crop, smallholders tend to spray when they've got a problem, not preemptively. This makes chemicals usage endogenous, in particular negatively correlated with output, all else held equal. We are probably best served by simply omitting chemicals entirely.

Allowing for production function shifts

A central part of our theory of poverty traps is the possibility of discrete jumps in productivity for those who can invest in key, lumpy inputs or who can incur the fixed/sunk costs of shifting to a different (and presumably, superior) production method. We therefore want to allow for this in our estimation of production functions.

The key implication is that we need to allow for different marginal products of land and/or labor across distinct production regimes. This is most easily done through interaction terms (e.g., with respect to fertilizer) or through nesting completely different production technologies, e.g., due to mechanization or use of different cultivation methods (e.g., SRI in the case of rice in Madagascar) using switching regressions estimation methods. It doesn't seem that we have sufficient cases of such significant shifts in production technologies in our data to justify switching regressions estimation. So we should probably stick just to interaction terms.

Another form of production function shift that some of us have discussed repeatedly in thinking through the structure of CLASSES arises due to the timing of labor activities.

Farmers who are late in performing key tasks – e.g., due to having to work for wages on others' farms – commonly suffer significant yield losses as a result. Where we have data on specific timing of activities, we can therefore use this. But since our data are all recall over a season or more, I doubt we have sufficiently accurate timing data to capture this effect econometrically. Please let us know if you think it feasible in a site's data set.

Multicollinearity and precision

The CLASSES model, like any simulation tool, depends on reasonably precise parameter estimates if it is to replicate observed behavior well and generate useful out-of-sample predictions of behavioral patterns. Multicollinearity – a high correlation among independent variables – reflects limited independent covariance between individual regressors, making it difficult to isolate the effects of one variable on output, holding the others constant. Since our sample sizes are small in each site, multicollinearity is a concern. It will be important, therefore, to compute correlation matrices among all prospective regressors in the production functions prior to estimation and to consider carefully whether to include both of any pair of variables that are relatively highly correlated with one another (e.g., r>0.50).

One input to which this concern commonly applies is seed. Seed is arguably the most important input to any annual crop. Because seeding densities typically do not vary dramatically within a population, seed quantities tend to be strongly, positively correlated with land area under cultivation. Moreover, seed quantities are notoriously difficult for smallholders to report accurately since they commonly store seed from the previous harvest for planting the next season or year and don't weigh it. Those who purchase seed typically have a more accurate recall of seed quantity applied. But few of our farmers buy seed. As but one example, seed application quantities appear highly unreliable in the Madzuu data. In general, we will not want to include seed quantity because (i) the data are especially prone to errors-in-variables problems, and (ii) they tend to be highly correlated with land area.

Functional form

As Chambers' seminal book (*Applied Production Analysis: A Dual Approach*, Cambridge University Press, 1988) emphasizes, choosing a functional form is more art than science. The basic principle is that one wants to restrict the empirical results as little as possible since classical statistical tests are only valid for inference under the maintained hypothesis that the general model is correct. The goal in our production function estimation is to come up with a reasonably accurate numerical representation of the production relationships in play in our sites. These production relationships can be described by (m+1)(m+2)/2 distinct parameters: 1 output level for a given input vector, m different marginal physical products, and the m(m+1)/2 elements of a symmetric Hessian matrix describing second-order effects. To the extent that data availability permits, we want to estimate production functions using functional forms that permit there to be (m+1)(m+2)/2 free parameters to estimate. Flexible functional forms of this sort are commonly known as *generalized quadratic* because they follow the general form

$$\mathrm{h(y)} = \beta_0 + \sum_{i=1}^{m} \beta_i g(x_i) + \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m} \beta_{ij} g(x_i) g(x_j) \qquad (9)$$

where h(•) and g(•) are arbitrary functions, typically linear, square root or natural logarithm, although they could be any function that fits the data well. The key here is to balance flexibility against conserving degrees of freedom so as to get sufficiently precise parameter estimates. This may imply dropping terms, although keep in mind that our objective is not inference but rather precise specification of the underlying production technology. Statistical significant is not our first-order concern in this exercise.

In general, we probably want to start by letting h(•) and g(•) be simple linear functions of the variables (i.e., h(y)=y and g(x)=x). One reason for this is that many of our input variables will have zero valued observations because some (even most) farmers don't use, for example, inorganic fertilizer or animal traction. This necessitates the use of terms that are linear or polynomial in the variable itself. The common practice of taking natural logarithms (e.g., a log-log model associated with a Cobb-Douglas production function or a translog flexible functional form) is problematic in the face of zero-valued observations since ln(0) is not well defined. Where regularity conditions (if you don't know what these are, ignore the sentence) favor a log-based form such as the translog, then the best practice is to replace zero-valued observations with 1/10th (or some similarly very small fraction) of the minimum positive observation in the data set. In general, however, we want to try to estimate production functions using a generalized quadratic that relies on transforms of variables that are well-defined for all values observed in our data, zeroes included.

Data Organization and ID Codes

One final observation on the organization of the data to be used in production function estimation for the BASIS CRSP project. For several of our sites, data are recorded in numerous separate files and will need to be combined into larger data files before running regressions. That will only be possible if all files contain the necessary identifying variables. For data that have been collected at household level, we need the household ID to be included, while for data collected at plot level, we need both household ID and plot ID to be included. The names of these variables must have the same spelling in each file in order to merge them easily. The key is that every data file needs to have the relevant ID code attached to each observation: plot ID codes for plot-level data and household ID codes for household-level data.

Conclusion

We hope the preceding comments are helpful as you set about estimating multiple output production functions from the data from each of our BASIS CRSP project sites. This has been a hasty, casual treatment of a reasonably complex subject. Please do not hesitate to ask questions or raise concerns as appropriate. One of the purposes of this memo is to stimulate discussion among the team. The main objective, however, is to equip country teams to get to work on estimating the multi-output production functions that characterize the systems under study. We hope this indeed facilitates your work, we welcome your comments and questions on these methods, and we look forward to working with you on this over the coming months.