Model Selection in Stochastic Frontier Analysis: Maize Production in Kenya

Yanyan Liu [*]

Department of Agricultural Economics and Department of Economics
Michigan State University
205 Cook Hall, East Lansing, MI 48823

Selected Paper prepared for presentation at the American Agricultural Economics Association Annual Meeting, Long Beach, California, July 23-26, 2006

[*] I would like to thank Peter Schmidt and Bob Myers for their suggestions and guidance. Thanks also go to Thomas Jayne for generously providing me the data, Elliot Mghenyi, Kirimi Sindi, Tavneet Suri, Zhiying Xu, and especially Margaret Beaver for helping me understand the data. All remaining errors and omissions are my own.

1 Introduction

Stochastic production frontier analysis has been widely used to study technical efficiency in various settings since its introduction by Aigner et al. (1977), and Meeusen and van den Broeck (1977). The approach has two components: a stochastic production frontier serving as a benchmark against which firm efficiency is measured, and a one-sided error term which has an independent and identical distribution across observations and captures technical inefficiency across production units. Recent studies have generalized the one-sided error term to allow its distribution to be heterogeneous by associating various features of the distribution with firm characteristics (see Battese and Coelli 1995; Caudill et al. 1995; Wang 2002; Wang 2003).

Allowing inefficiency to depend on firm characteristics enables researchers to examine the determinants of inefficiency, and to suggest policy interventions to improve efficiency. However, many policy suggestions in the previous literature have been limited for at least two reasons. First, little is known about how to choose among competing models of efficiency, or the implications of model choice on estimation results. Second, past studies on production efficiency have mostly focused on the *directions* of the influence of the exogenous factors on efficiency level while the *magnitudes* of the partial effects have often been overlooked. This is surprising given that the magnitudes of the effects of the explanatory variables on dependent variables are often the focal point in other regression analyses.

In this paper, we make three contributions to the stochastic frontier literature. First, we show how to estimate the quantitative magnitude of the partial effects of exogenous factors on output levels and how to put standard errors on these partial effects. We also propose an R^2-type measure to summarize the overall explanatory power of the exogenous factors on inefficiency. Second, we examine the effects of model selection on the empirical results. We find that different models can lead to rather different magnitudes of the partial effects of the exogenous factors. Others have also found that different models give different parameter estimates, which is on the face of it an unsurprising finding. Third, we show how a recently developed model selection procedure (Alvarez, Amsler, Orea and Schmidt, 2006, hereafter AAOS) can be used to choose among the competing models, and we use bootstrapping to provide evidence on the power of this procedure. The model selection procedure gives an unambiguous choice of best model. This is important because, if different models give different results and we cannot distinguish statistically among the models, we do not know which set of results to believe; whereas, if we can pick a clearly best model, it does not matter whether other models give different results.

Our empirical analysis is on maize production in Kenya. The problem of hunger in Kenya remains widespread. Ranked 159th out of 177 countries in the world in terms of GDP per capita, about 59 percent of the population in Kenya earned less than 2 dollars per day in 2002.[1] Kenya's economy heavily depends on agriculture with 75 percent of Kenyans making their living from farming. Maize is the primary staple food and most farmers are engaged in maize production in Kenya. In recent years, total maize output has not kept pace with the growing population and demand, largely due to falling land productivity: average national maize yields have fallen from over 2 tons per hectare in the early 1980's to about 1.6 tons per hectare recently (Nyoro et al. 2004).

In order to alleviate poverty and achieve food security in Kenya and other Sub-Saharan countries, it is important to identify and quantify the factors that hinder farm efficiency in maize production. Taking advantage of a detailed household survey in Kenya, we investigate determinants of productivity and inefficiency using stochastic frontier analysis. The variables used to explain inefficiency in our analysis are related to education background of the household, rural infrastructure, land tenure, credit constraints faced by the household,

[1] Human Development Report 2004 by United Nations Develop Programme (UNDP).

and farm size. These explanatory factors go well beyond those used to study production inefficiency in other studies of agriculture (see Kumbhahar et al. 1991; Huang and Kalirajan 1997; Alvarez and Arias 2004).

In the remainder of the paper, we first review the stochastic frontier production model commonly used in the literature. In section 3, we describe our data and variables used in the empirical analysis. Section 4 presents the estimation results from different model specifications. In section 5, we carry out specification tests to choose our final model. A novel detail is that we use the bootstrap to examine the reliability of these specification tests in choosing the correct model. Section 6 is an analysis of technical efficiency in maize production in Kenya based on the final model. Conclusions then follow.

2 A Stochastic Production Frontier Model

We now present a basic stochastic frontier production function. We examine the partial effects of exogenous factors and propose a measurement similar to R^2 to summarize the explanatory power of the exogenous factors on inefficiency levels. After that, we review several existing specifications of the one-sided error term in stochastic frontier analysis.

2.1 Basic Setup

The basic setup and notation follow Wang and Schmidt (2002) and AAOS. Fields are indexed by $i = 1, \ldots, N$. Let y_i be the log output; x_i be a vector of inputs; and z_i be a vector of exogenous variables that exert an influence on farm efficiency. Let y_i^* be the unobserved frontier which is denoted as

$$y_i^* = x_i'\beta + v_i, \tag{1}$$

where v_i is distributed as $N(0, \sigma_v^2)$ and is independent of x_i and z_i. The actual output level y_i equals y_i^* less a one-sided error, u_i, whose distribution depends on z_i. The model is written as

$$y_i = x_i'\beta + v_i - u_i(z_i, \theta), \quad u_i(z_i, \theta) \geq 0, \tag{2}$$

where θ is a vector of parameters. It is assumed that u_i and v_i are independent of each other, that conditional on z_i, u_i is independent of x_i, and that v_i is independent of x_i and z_i. The frontier function itself and the inefficiency part are generally estimated in one step using maximum likelihood estimation (MLE) to achieve both efficiency and consistency.[2]

Indexing exogenous factors with $k = 1, \ldots, K$, we take expectations conditional on x_i and z_i, and then take partial derivatives with respect to z_{ik} on both sides of equation (2), to get

$$\partial[E(y_i|x_i, z_i)]/\partial z_{ik} = \partial[E(-u_i|x_i, z_i)]/\partial z_{ik}. \tag{3}$$

Here, $\partial[E(-u_i|x_i, z_i)]/\partial z_{ik}$ can be interpreted as the partial effect of z_{ik} on efficiency $-u_i$, and can also be interpreted as the partial effect on y_i. Because y_i is the log output, $\partial[E(-u_i|x_i, z_i)]/\partial z_{ik}$ is the semi-elasticity of output with respect to the exogenous factors, i.e., the percentage change in expected output change when z_{ik} increases by one unit. Similarly, we have

$$\partial[V(y_i|x_i, z_i)]/\partial z_{ik} = \partial[V(u_i|x_i, z_i)]/\partial z_{ik}. \tag{4}$$

So $\partial[V(u_i|x_i, z_i)]/\partial z_{ik}$ is the partial effect of z_{ik} on the variance of both the inefficiency term u_i and log output. It can be interpreted as an estimator of the partial effect of z_{ik} on production uncertainty.

[2] Some studies use a two-step procedure where the frontier function is estimated first, and then the inefficiency term is regressed on exogenous variables in the second step. This procedure, however, suffers bias for two reasons. The first and more obvious reason is the possible correlation between the input variables in the frontier function and the variables in the inefficiency term. The second reason is that the inefficiency term from the first step is measured with error and the error is correlated with the exogenous factors. See Wang and Schmidt (2002) for an extensive discussion and evidence from Monte Carlo experiments.

Wilson et al. (2001) are among the first to search for an estimator of the partial effects of exogenous factors on technical efficiency. They suggest $\partial E[\exp(-u_i)|z_i, \epsilon_i]/\partial z_{ik}$. However, we regard $E[\exp(-u_i)|z_i, \epsilon_i]$ as an *estimate* of inefficiency, and we are interested instead in the effect of z_i on inefficiency or average inefficiency. That is, we are interested in the effect of z_i on $\exp[-u_i(z_i, \theta)]$ or $E[\exp(-u_i(z_i, \theta))]$, which motivates the expression given in equation (3) above. The measures $\partial[E(u_i|x_i, z_i)]/\partial z_{ik}$ and $\partial[V(u_i|x_i, z_i)]/\partial z_{ik}$ were first proposed and used in Wang (2002) and Wang (2003), but for different purposes than they are used here. Here, we interpret $\partial[E(-u_i|x_i, z_i)]/\partial z_{ik}$ as the semi-elasticity of output with respect to exogenous factors so that not only its sign but also its magnitude are of economic interest. We also provide formulas to compute valid standard errors for $\partial[E(-u_i|x_i, z_i)]/\partial z_{ik}$ and $\partial[V(u_i|x_i, z_i)]/\partial z_{ik}$ using the delta method in several model specifications as described below.

It will often be useful to measure how well the vector of exogenous factors, z, explains inefficiency, u. Surprisingly, this has not be addressed in the previous literature. We suggest a statistic R_z^2, to summarize the explanatory power of z. The variance of the inefficiency term u can be decomposed as

$$V(u) = V_z[E(u|z)] + E_z[V(u|z)]. \tag{5}$$

The fraction of variation in u that is explained by z is $\frac{V_z[E(u|z)]}{V(u)}$. Thus a natural measure of explanatory power would be

$$R_z^2 = \frac{\sum_{i=1}^n \left[\widehat{E(u_i|z_i)} - \frac{1}{n}\sum_{i=1}^n \widehat{E(u_i|z_i)}\right]^2}{\sum_{i=1}^n \left[\widehat{E(u_i|z_i)} - \frac{1}{n}\sum_{i=1}^n \widehat{E(u_i|z_i)}\right]^2 + \sum_{i=1}^n \widehat{V(u_i|z_i)}}. \tag{6}$$

Similarly to R^2 in an ordinary least squares regression, R_z^2 can be called the goodness of fit of the efficiency component, and it can be interpreted as the fraction of the sample variation in u that is explained by z.

2.2 Alternative Model Specifications

In the original specification of stochastic frontier functions, Aigner et al. (1977) and Meeusen and van den Broeck (1977) assumed an identical and independent half-normal distribution for the one-sided error terms u_i. Subsequent studies have generalized the model to allow for heterogeneity in the distribution of the inefficiency term. We will consider models in which the distribution of u_i is truncated normal. Kumbhakar et al. (1991), Huang and Liu (1994) and Battese and Coelli (1995) allow the mean of the pre-truncation normal distribution to depend on a set of exogenous factors. Reifschneider and Stevenson (1991), Caudill and Ford (1993), Caudill et al. (1995) and Hadri (1999) allow exogenous factors to affect the variance of the pre-truncation normal distribution. Wang (2003) allows both the mean and the variance to depend on exogenous factors.

Regardless of whether we parameterize the mean or the variance of the pre-truncated normal, both the mean and the variance of the truncated normal will depend on the exogenous factors. These are sometimes called models of heteroscedasticity, but the fact that the mean also changes makes this choice of words potentially misleading. Whereas heteroscedasticity affects only the efficiency of estimation in the usual linear model, in a stochastic frontier model with heterogeneity, failure to model the exogenous factors leads to biased estimation of the production frontier model and the level of technical inefficiency, hence leading to poor policy conclusions (e.g. Caudill and Ford 1993, Caudill et al. 1995, Hadri 1999, Wang 2003).

With different specifications available to model heterogeneity, it is unclear which should be chosen in an empirical analysis. The choices made in many past studies seem to be somewhat arbitrary. However, a carefully specified model might help to increase estimation efficiency and remove sources of bias and inconsistency (Wang 2003). Moreover, there has been little investigation on the difference in estimation results from various specifications which allow for heterogeneity. In order to deal with the model specification problem, researchers usually do sensitivity analysis using competing models. But if the competing models give very different results, it is difficult to pick one and discard the others. Wang (2003) treats this problem by specifying a flexible model. However, a more flexible model incorporates more parameters, which impose a higher computational burden and reduce degrees of freedom. Given that large samples are typically difficult to obtain in stochastic frontier models, some relevant parameters may be estimated imprecisely in flexible model forms.

AAOS suggests a procedure that helps to specify a proper model for the one-sided error term. First, assume the general model of inefficiency (Wang 2003) in which u_i is distributed as $N(\mu_i, \sigma_i^2)^+$, with $\mu_i = \mu \cdot \exp(z_i'\delta)$ and $\sigma_i = \sigma_u \cdot \exp(z_i'\gamma)$. This general model nests several simpler models, many of which have been used in previous studies. In particular, the following six models are special cases of the general model.

1. Scaled Stevenson model: Let $\delta = \gamma$. Then the distribution of u_i becomes $\exp(z_i'\delta) \cdot N(\mu, \sigma_u^2)^+$, which is used in Wang and Schmidt (2002) and discussed by Simar et al. (1994).

2. KGMHLBC model: Let $\gamma = 0$. Then the distribution of u_i becomes $N(\mu{\cdot}\exp(z_i'\delta), \sigma_u^2)^+$, which has been considered in Kumbhakar et al. (1991), Huang and Liu (1994), and Battese and Coelli (1995).

3. RSCFG-μ model: Let $\delta = 0$. Then the distribution of u_i becomes $N(\mu, \sigma_u^2{\cdot}\exp(2z_i'\gamma))^+$.

4. RSCFG model: Let $\mu = 0$. Then the distribution of u_i becomes $\exp(z_i'\gamma) \cdot N(0, \sigma_u^2)^+$, which is considered in Reifschneider and Stevenson (1991), Caudill and Ford (1993), and Caudill et al. (1995).

5. Stevenson model: Let $\delta = \gamma = 0$. Then the distribution of u_i becomes $N(\mu, \sigma_u^2)^+$, which is the model of Stevenson (1980).

6. ALS model: Let $\mu = \gamma = 0$. Then the distribution of u_i becomes $N(0, \sigma_u^2)^+$, which is the model of Aigner et al. (1977).

Among the six models, the scaled Stevenson, KGMHLBC and RSCFG-μ models have the same number of parameters. The RSCFG model is nested by the scaled Stevenson model and the RSCFG-μ model. Notice that the Stevenson model and the ALS model do not contain any variables (z_i) that influence the distribution of inefficiency. AAOS show how to use likelihood ratio (LR) tests, LM tests and Wald tests to test the above restrictions, and hence to choose a plausible model for inefficiency.

3 Data

In this section, we first discuss our data sources for the analysis of maize production in Kenya. We then describe the variables used in the frontier production function and the inefficiency term.

3.1 Data Source

The data are from a rural household survey of about 1100 households that planted maize in the main season of 2003-2004 in Kenya.[3] The survey was designed and implemented under the Tegemeo Agricultural Monitoring and Policy Analysis Project, a collaboration among Tegemeo Institute of Egerton University, Michigan State University, and the Kenya Agricultural Research Institute. Field level data are available and some households planted maize in more than one field. The survey includes not only detailed field production information but also rich demographic and infrastructure characteristics of each household. The production data for each field include size of field, yield, labor input associated with each type of planting activity, fertilizer application and seed usage. The demographic infor-

[3] See Suri (2005) for a study of the adoption decisions of hybrid seed by maize producers in Kenya using the same data set.

mation for each household includes the age, gender and education level of each household member; how far a household is from a bus stop, a motorable road, a telephone booth, mobile phone service, and extension service; whether a household member has non-farming income; whether a household receives loans; how much land a household owns, and land tenure. Rainfall and soil quality data are also available at the village level.

3.2 Variables in the Production Frontier

In the production frontier part of the model, the output variable is maize yield per acre, and the input variables are applied fertilizer nutrients, labor, maize seeds and machine usage. Since both the output and inputs are in per acre terms, land is not explicitly included as an input. Most of the maize fields are inter-crop fields where more than one type of crop is planted in the same season. Because most inputs (land, fertilizer and labor) are at the field level and cannot be separately allocated to maize production only, we generate an output index for inter-crop fields using:

$$Y_i = \left(\sum_j Y_{ij} P_j \right) / P_1, \tag{7}$$

where Y_i is the output index, P_j is the market price of crop j, Y_{ij} is the yield of crop j in field i, and crop 1 is maize. The fields with more than three types of crops are deleted because we want to focus on the fields where maize is the major crop.[4] Only pre-harvest labor input is included because harvesting and post-harvest activities have little effect, if any, on yield. The unit of labor is person-hours. One person-hour of labor from children younger than 16 is transformed to 0.6 person-hours of adult labor. Nitrogen, the most important nutrient in maize growth, is computed from fertilizer application data according to the quantity and composition of each type of fertilizer used.[5] The maize seeds can be grouped into hybrid seeds and local seeds. Hybrid seeds are more productive and more

[4] 637 out of the total 1718 fields are dropped.

[5] More than 20 types of fertilizers were applied.

expensive than local ones. Ideally we would want to generate a maize seed quantity index for inter-crop fields as in the case of maize yield, but we do not have good seed price data. We use a dummy variable MONO as an indicator for mono-crop fields. Tractor usage in land preparation is the only machine used for pre-harvest activities. This is captured by a dummy variable TRACTOR with 1 indicating that a tractor was used and 0 otherwise.

Besides inputs, some environmental variables are included on the right hand side of the frontier production function. Failure to control for environmental variables may cause a correlation between yield and some inputs (for example, if a farmer makes input decisions based on some soil properties that also affect maize yield) and therefore may bias estimates of the production frontier and inefficiency level (Sherlund et al. 2002). In order to control for environmental conditions, we include eight dummy variables indicating different zones. Farms in the same zone share similar terrain and climate conditions. We also include three village level variables: DRAINAGE, DRAINAGE^2 and STRESS. DRAINAGE captures the drainage property of the soil. It is a categorical variable ranging from one to 10 where one indicates the least and 10 the highest drainage. DRAINAGE^2 is the square of DRAINAGE. We include a quadratic term because yield increases in DRAINAGE at lower drainage levels and decreases at higher drainage levels. Rainfall is a very important factor in maize production in Kenya, because all of the maize fields are rain-fed fields, and drought is the usual cause of yield loss. We use a variable STRESS to capture the moisture stress in maize growth. STRESS is computed as the total fraction of 20-day periods with less than 40 millimeters of rain during the 2003-2004 main season. This is a better measure for moisture condition than total rainfall in that the total rainfall does not reflect the distribution of rainfall over time which is very important in maize growth.

Observations with missing values are discarded. Out of concern for large measurement errors, we also drop any observation that satisfies one of the following conditions: 1) yield lower than 65 kg per acre or higher than 4580 kg per acre, 2) seed usage less than two kg per acre or more than 20 kg per acre, and 3) labor input less than 40 person-hours per acre or more than 2200 person-hours per acre. After these adjustment, there are 815 fields (observations) remaining.[6] The 815 fields were managed by 660 households. Table 1 summarizes the descriptive statistics for the variables included in the frontier production function (excluding zone dummies).

[INSERT TABLE 1 ABOUT HERE]

3.3 Exogenous Factors Affecting Efficiency

Previous studies have identified numerous factors that limit farm productivity and efficiency. Education is arguably an important factor that affects productivity and efficiency. Kumbhakar et al. (1989) suggest that education increases the productivity of labor and land on Utah dairy farms while Kumbharkar et al. (1991) also show that education affects production efficiency. Huang and Kalirajan (1997) find that average household education level is positively correlated with technical efficiency levels for both maize and rice production in China.

Physical and social infrastructure, such as road conditions, access to telephone and mobile phone service, access to extension service, etc., have also been mentioned for their role in rural development and farm productivity. Jacoby (2000) examines the benefits of rural roads to Nepal farms and suggests that providing road access to markets would confer substantial benefits through higher farm profits. Karanja et al. (1998) show that distance to the nearest motorable road and access to extension services have positive effects on maize productivity in Kenya. More developed infrastructure helps farmers to obtain more information and thus may improve technical efficiency.

[6] That is, 266 fields are dropped due to missing or unreasonable values.

Land tenure is another element that affects farm performance. Secure tenure may induce more investment (such as soil conservation) and increase farm productivity in the long run. Place and Hazell (1993) suggest tenure to be important to investment and productivity in Rwanda. Puig-Junoy and Argiles (2000) show that farms with a large proportion of rented land have low efficiency in Spain.

Financial constraints, such as limited access to credit, might affect farm input decisions and efficiency. Ali and Flinn (1989) show that credit non-availability is positively and significantly related to profit inefficiency for rice producers in Pakistan. Parikh et al. (1995) find that farmers with larger loans are more cost efficient in Pakistan. The effect of financial constraints on technical efficiency seems to be unexamined. This effect may exist because, besides the quantities of input used, the timing of input usage is also important in affecting yields. The farms that face financial constraints may not be able to optimize production.

The inverse relationship between farm productivity and farm size has been a long-standing empirical puzzle in development economics since Sen (1962) (see Benjamin 1995; Barrett 1996; Lamb 2003). The empirical results have been mixed on the relationship between efficiency and farm size. Kumbahakar et al. (1991) show that large farms are relatively more efficient both technically and allocatively. Ahmad and Bravo-Ureta (1995) find a negative correlation between herd size and technical efficiency, while Alvarez and Arias (2004) find a positive relationship between technical efficiency and size of Spanish Dairy farms. Huang and Kalirajan (1997) show that the size of household arable land is positively related to technical efficiency in maize, rice and wheat production in China. Parikh et al. (1995) find that cost inefficiency increases with farm size. Hazarika and Alwang (2003) show that cost inefficiency in tobacco production is negatively related to tobacco plot size but unrelated to total farm size in Malawi.

The household survey discussed above allows us to investigate all of the above exogenous factors simultaneously. We group the exogenous variables into five categories: 1) socio-economic variables including the highest education level among the household members (EDUHIGH) and gender of household head (FEMHEAD);[7] 2) infrastructure variables including how far a house is from the nearest bus stop (DISTBUS), from telephone or mobile-phone service (DISTPHONE) and from extension service (DISTEXTN);[8] 3) land tenure, which is a dummy variable (OWNED) with 1 indicating that the field is owned by the household and 0 that the field is rented; 4) credit constraints including two variables: CRDCSTR and RNFINC, where CRDCSTR is a dummy variable with 1 indicating the household has unsuccessfully pursued credits and 0 otherwise. RNFINC is the proportion of household members that have non-farming income; and 5) size variables including farm size (TTACRES) and field size (ACRES). Table 2 summarizes the notation and descriptive statistics for these exogenous factors.

[INSERT TABLE 2 ABOUT HERE]

4 Estimation Results From Competing Models

In this section, we report the estimation results for various model specifications. We start with the flexible translog functional form for the frontier production function, and we interact moisture stress and the dummy of hybrid maize seed with inputs out of the concern that they may affect the output elasticities.

[7] EDUHIGH can capture the effects of eduction on efficiency for a household better than the average education level or the education level of the household head, in that the one who receives the highest education can help the household head and the other household members in making production decisions.

[8] We use DISTBUS instead of how far a household is from a motorable road, because only a very small proportion of the households in Kenya own motorable transportation tools (like tractors), and bus and bicycles are the major transportation tools there.

There are 29 parameters in the frontier production function. If we use the AAOS general model, there are 23 parameters in the error term ϵ_i. The total dimension of the parameter space is 52. Even for the simpler models, such as the scaled Stevenson model, the KGMHLBC model, and the RSCFG model, the dimension is still large (42 parameters). To maximize a likelihood with such a high dimension can be computationally difficult given the complexity and non-regularity of our likelihood function. We do not want to reduce the dimension of the exogenous factors under investigation, neither do we want to sacrifice the flexibility of the efficiency model. Instead of taking a less flexible frontier production function, such as Cobb-Douglas, we follow a three-step procedure to drop some unnecessary (statistically insignificant) variables. The details of this procedure are given in the appendix of Liu (2006). The following analysis is based on a reduced set of explanatory variables in the frontier production function and a reduced set of exogenous factors.

Table 3 reports the estimation results of the efficiency component for all of the five models that allow efficiency to depend on farm characteristics. Both the LR test and Wald test reject the null hypothesis that all the exogenous factors have zero effect at the 1% significance level in each of the five models.

[INSERT TABLE 3 ABOUT HERE]

All the five models yield similar results for production frontier and efficiency estimates, consistent with previous studies (e.g. Caudill et al. 1995). Table 4 reports the estimation results of the production frontiers.[9] The parameter estimates are very similar in the different models. The Battese and Coelli efficiency estimates are computed for each observation in all the models. Their correlations among alternative models are reported in table 5. The lowest correlation is 0.97.

[9] The production frontiers are estimated together with their efficiency components, though reported separately.

[INSERT TABLE 4 ABOUT HERE]

[INSERT TABLE 5 ABOUT HERE]

The coefficients of the exogenous factors reported in table 3 are not very interesting by themselves, because they are the parameters of the pre-truncated distribution of the inefficiency term u_i. By themselves, these parameters do not tell us how the exogenous factors affect the distribution of u_i which is truncated. In order to quantify the effects of exogenous factors, we compute $\partial[E(-u_i|x_i, z_i)]/\partial z_i$ and $\partial[V(u_i|x_i, z_i)]/\partial z_i$ for each observation. The formulas for $\partial[E(-u_i|x_i, z_i)]/\partial z_i$ and its standard error for the general model are provided in the appendix; while for $\partial[V(u_i|x_i, z_i)]/\partial z_i$, the formulas are provided in the appendix of Liu (2006). To obtain the formulas for the nested models, we only need to impose the corresponding restrictions on the parameters.[10]

The partial effects of the exogenous factors evaluated at the sample mean and their standard errors are reported in table 6. For each of the exogenous factors, the signs of the partial effects are the same for all the models. However, different models give quantitatively different values for the partial effects. For example, the partial effects of TTACRES on the conditional mean of $-u$ range from 0.0023 to 0.0072, and these differences are large relative to the standard errors of the estimates.

[INSERT TABLE 6 ABOUT HERE]

Table 7 reports the average partial effects of EDUHIGH on $E(-u_i|x_i, z_i)$ for the observations within each of the four quartiles of the efficiency levels estimated in the KGMHLBC model. The KGMHLBC model shows an increasing trend of the partial effect of education on efficiency levels from low to high quartiles, while the scaled Stevenson model, RSCFG-μ model and RSCFG model suggest a decreasing trend.[11] According to the KGMHLBC

[10] Wang (2002) gives the expression, for these derivatives but not for the standard errors.

[11] We observe similar patterns for the other exogenous factors. These results are not reported in order to save space.

model, we would conclude that the households with lower efficiency levels would not benefit as much as the ones with higher efficiency levels from further investment in education. However, an opposite conclusion would follow if we use the scaled Stevenson model, the RSCFG-μ model or the RSCFG model.

[INSERT TABLE 7 ABOUT HERE]

Table 8 reports the correlations of partial effects of EDUHIGH on $E(-u_i|x_i, z_i)$ among alternative models. Most correlations are very low and some are even negative.[12] This further confirms our conclusion that different models yield rather different partial effects. Therefore, if we are only interested in the signs of the yield semi-elasticities with respect to exogenous factors, model specification is not important. However, if we are interested in the magnitudes of the yield semi-elasticities, it is important to choose the appropriate model specification.

[INSERT TABLE 8 ABOUT HERE]

5 Model Selection

In this section, we apply the procedure proposed by AAOS to select an appropriate model for our empirical application. A bootstrap analysis then follows to evaluate the performance of the model selection procedure.

5.1 Empirical Model Selection

We start with the general model, and then use LR tests to find simpler models that the data do not reject. We choose LR tests over Wald and LM tests because the LR statistics

[12] We observe similar patterns for the other exogenous factors. These results are not reported in order to save space.

are more stable numerically than the Wald and LM statistics.

Estimation of the general model yields a log-likelihood value of -616.30. Table 9 reports the log-likelihood values for the six restricted models nested in the general model. Taking the general model as the unrestricted model, we then test the restrictions that would reduce the general model to simpler specifications. Based on LR tests (test statistics with Chi-squared critical values are listed in table 9), we obtain the following results:

- We can reject the scaled Stevenson model ($\delta = \gamma$), RSCFG-μ model ($\delta = 0$), and RSCFG model ($\mu = 0$) at the 5% significance level.
- We fail to reject the KGMHLBC model ($\gamma = 0$) at any reasonable significance level.
- We can reject the Stevenson model ($\delta = \gamma = 0$) and ALS model ($\mu = \gamma = 0$) at any reasonable significance level.

Because both the Stevenson model and ALS model are rejected, we conclude that the exogenous factors do affect efficiency. Among RSCFG, RSCFG-μ, and scaled Stevenson models, the RSCFG model is preferred because we fail to reject the RSCFG model at any reasonable significance level using the RSCFG-μ model or the scaled Stevenson model as the unrestricted model. Moreover, among all the models, the KGMHLBC model is most preferred because it is the only one that we can accept at any reasonable significance level. Therefore, we select the KGMHLBC model as our final model.

[INSERT TABLE 9 ABOUT HERE]

5.2 A Bootstrap Evaluation

The model selection procedure proposed by AAOS leads to one clearly preferred model, the KGMHLBC model, among the set of competing models. It is encouraging to obtain an unambiguous outcome. However, it is also relevant to ask about the reliability of the model selection criterion, which is a question of the size and power properties of the LR tests. We investigate this question using the bootstrap. That is, we generate data via the bootstrap assuming that the KGMHLBC model is correct, and then we see how reliably the model selection procedure picks the KGMHLBC model. So far as we are aware this is a novel suggestion. It is useful because we are using the bootstrap to evaluate the probability with which the actual model selection procedure will pick the correct model.

The KGMHLBC model is written as

$$y_i = x_i'\beta + v_i - u_i, \quad \text{where} \quad u_i \sim N[\mu \cdot \exp(z_i'\delta), \sigma_u^2]^+ \quad \text{and} \quad v_i \sim N(0, \sigma_v^2). \tag{8}$$

We take the following steps to conduct the parametric bootstrap:

1. Using the actual sample data $\{(y_i, x_i, z_i)\}_{i=1}^n$, we estimate the vector of parameters, $\hat{\theta} = \{\hat{\beta}, \hat{\delta}, \hat{\mu}, \hat{\sigma_u^2}, \hat{\sigma_v^2}\}$, in the KGMLBC model using MLE. These results were given above.

2. Then we generate data sets based on the parameter estimates from step 1. For $i = 1, \ldots, n$, draw u_i^* from $N[\hat{\mu} \cdot \exp(z_i'\hat{\delta}), \hat{\sigma_u^2}]^+$ and draw v_i^* from $N(0, \hat{\sigma_v^2})$, and then compute $y^* = x_i'\hat{\beta} + v_i^* - u_i^*$.

3. Based on the pseudo-data $\{y_i^*, x_i, z_i\}_{i=1}^n$ generated in step 2, we estimate all of the seven models using MLE. We obtain the log-likelihood value (ll) and parameter estimates $(\hat{\theta})$ in each of the models, denoted as $\zeta^* = \{(ll_j^*, \hat{\theta}_j^*)\}_{j=1}^J$, where j indexes the different models.

4. Repeat steps 2 and 3 B times to obtain $\mathscr{B} = \{\zeta_b^*\}_{b=1}^B$.[13]

We use the log-likelihood statistics in $\mathscr{B}$ to conduct the specification tests for each pseudo-data set as in section 5. We take the general model as the unrestricted model and conduct LR tests at the 5% significance level. The results are:

[13] We set $B = 1000$.

- We reject the true model in 5.7% of the pseudo-data sets, the scaled Stevenson model in 75% of the pseudo-data sets, the RSCFG-μ in 78% of the pseudo-data sets, and the RSCFG in 75% of the pseudo-data sets.

- We reject both the Stevenson model and the ALS model in 99.9% of the pseudo-data sets. That is, in only one of the 1000 data sets, we would wrongly conclude that the set of exogenous factors do not affect efficiency.

- We accept the true model and reject all of the other models in 66.0% of the pseudo-data sets. Only in 0.4% of the data sets, we reject the true model and accept an alternative one at the same time.

- In 28.4% of the pseudo-data sets, we simultaneously accept the true model and at least one of the alternative models. And we reject all of the models simultaneously in 5.3% of the data sets.

We view these results as quite favorable. If the KGMHLBC model is correct, the model selection procedure will reject it with small probability (6%), and will pick it unambiguously with relatively high probability (66%).

The bootstrap results can be used to generate confidence intervals for any of our original estimates. These confidence intervals may be more accurate in finite samples than those generated by first order asymptotic approximations such as the delta method. For example, we can use the parameter estimates of the KGMHLBC model in $\mathscr{B}$ to compute the partial effects for every observation in each pseudo-data set. Confidence intervals then follow directly from the set of $\mathscr{B}$ estimates. For example, with $B = 1000$, a 90% confidence interval for a parameter ranges from the 50th to the 950th largest values of the bootstrap estimates of that parameter. This is called the "percentile bootstrap". Table 10 reports 90% percentile bootstrap confidence intervals for the partial effects in the KGMHLBC model, evaluated at the sample mean. For purposes of comparison, it also gives the 90%

confidence intervals based on the delta method (i.e. using the standard errors as given in Table 6). The confidence intervals given by bootstrap and the delta method are not very different. This confirms the reliability of the delta method.

[INSERT TABLE 10 ABOUT HERE]

6 Post-Estimation Analysis

This analysis is based on the results of our selected model, the KGMHLBC model. Table 11 reports output elasticity estimates for local seed users and hybrid seed users (with the standard errors in the parentheses). The estimates are evaluated at the sample means.[14] The sum of the output elasticities with respect to nitrogen fertilizer, labor, and seed quantity is less than 1 (0.80 for local seed users and 0.74 for hybrid seed users). This does not mean the technology is decreasing returns to scale because we are holding land input constant by using yield per acre. Results show that yield is more responsive to nitrogen fertilizer application and seed quantity but less responsive to labor for hybrid seed users than for local seed users.

[INSERT TABLE 11 ABOUT HERE]

Figure 1 plots the density of the Battese and Coelli technical efficiency estimates. The minimum efficiency level is 18% and the maximum is 98%. The mean of technical efficiency is 71%, while the mode is around 80%. The distribution is left skewed.

[INSERT FIGURE 1 ABOUT HERE]

Goodness of fit statistic for the efficiency component, R_z^2, is 0.1035, indicating that 10.35% of the sample variation in efficiency can be explained by the exogenous factors. In table

[14] The means of FERTILIZER, LABOR, and SEED are computed after taking logarithms.

6, the school years of highest educated household member (EDUHIGH), ratio of household members who have non-farming income (RNFINC) and total acres of farm land (TTACRES) all had positive partial effects on the mean and negative effects on the variance of efficiency. Household head being female (FEMHEAD), distance to the nearest bus-stop (DISTBUS) and land being owned by the household (OWNED) all have negative effects on the mean and positive effects on the variance of efficiency. Therefore, an average household tends to have a higher efficiency level and a lower uncertainty on efficiency if it is characterized with a higher education level, more off-farm income, or larger farm size. Alternatively, it tends to have a lower efficiency level and higher uncertainty of efficiency if it has a female head, or is far from a bus-stop. These results are consistent with a prior reasoning. The effects of education, credit constraints, farm size and infrastructure on efficiency have been discussed extensively in the previous literature. Females are subject to social discrimination in Kenya. There are usually two situations in which a female can become the head of the household. One is that she is a single mother, and the other is that her husband is dead. Households headed by females are less efficient because females do not have the same inheritance rights as males in rural Kenya. A widow cannot obtain the full property of the land left by her husband, and has to give away a certain proportion of the harvest to her husband's brothers. This reduces the incentive to work hard. A surprising result is that farmers tend to be more efficient in rented fields than in their own fields. There are possible two reasons: 1) a fixed rent has to be paid at planting time, which provides more incentives for farmers who work in a rented field than in their own fields; 2) farmers rent fields that they know are productive. To the extent the second reason is a factor, the variable OWNED might capture the unobserved land quality not included as a covariate in our production frontier.

As explained earlier, not only the directions but the values of the partial effects on $E(-u_i|x_i, z_i)$ are of economic interest. According to the KGMHLBC model (see table 10), one more school year would increase yield per acre by 0.52 percent for an average household, ceteris paribus. One kilometer closer to publice transportation would increase yield per acre by 3.7 percent. An increase of one acre in farm size would raise yield per acre by 0.23 percent. If the proportion of household members who receive off-farm income increases by 10 percent, yield per acre would increase by 1.3 percent. However, a household with a female head tends to be 14 percent less efficient than a household with a male head, and farmers tend to be 17 percent more efficient working in rented fields than in their own fields.

Based on our estimation results for the efficiency component, investments in education and infrastructure help improve technical efficiency. Extension services can perhaps make up for reduced efficiency due to insufficient school education received by farmers. However, we found that the distance to the office of an extension service is insignificant. This suggests that the government should work on improving the quality of the extension service rather than setting up more offices. The result that larger farms are more efficient can provide some guidance for land reallocation. Better access to credit would also improve efficiency. Households with female head need special help to improve efficiency of production.

7 Conclusion

Poverty reduction in Africa has proved to be an immense challenge. This paper identifies factors that limit technical efficiency in maize production in Kenya and quantifies partial effects of these factors on the output level. We simultaneously examine five categories of exogenous factors: socio-economic variables, farm size, land tenure, credit constraints and infrastructure.

In our stochastic frontier analysis, we find that different stochastic frontier models predict similar efficiency levels and the same directions for partial effects of exogenous factors at the sample mean. However, the magnitudes of these partial effects for individual farms are rather different across model specifications. This finding calls for more attention to model selection in empirical stochastic frontier analysis. To choose among competing models, we employ the specification tests recently proposed by Alvarez, Amsler, Orea, and Schmidt (2006). In our application, these tests yield an unambiguous choice of best model, and an analysis of the model choice procedure using the bootstrap indicates that the model choice procedure is reasonably reliable.

In the paper we also propose an R^2-type measure that indicates the explanatory power of the exogenous factors that affect inefficiency. In our application we find that our exogenous factors explain approximately 10% of the variation in efficiency levels.

Appendix

Partial Effects of Exogenous Factors and Standard Errors

The mean of u_i conditional on x_i and z_i is:

$$E(-u_i|x_i, z_i) = -\sigma_i(R_1 + R_2) \tag{9}$$

where

$$\begin{aligned} R_1 &= \mu_i/\sigma_i, && (10) \\ R_2 &= \phi(R_1) \cdot [\Phi(R_1)]^{-1}. && (11) \end{aligned}$$

Assume there are K exogenous factors (K_1 continuous variables and $K_2 = K - K_1$ dummy variables). We deal with the continuous variables first. Let z_i^c be the K_1 dimensional vector of the continuous variables. We derive the partial effects of z_i^c on the mean efficiency as

$$\partial E(-u_i|x_i, z_i)/\partial z_i^c = \gamma^c\sigma_i(R_1R_3 - R_2) - \delta^c\sigma_i R_1(1 + R_3) \tag{12}$$

where δ^c and γ^c are the coefficient vectors of z_i^c,

$$R_3 = -R_2^2 - R_2R_1. \tag{13}$$

Next we derive the variances of the partial effects of z_i^c. Let $\theta' = (\delta' \ \gamma')$, and $g(\theta) = \partial[E(-u_i|x_i, z_i)]/\partial z_i^c$, where $g(\theta)$ is $K_1 \times 1$ dimensional vector. Following the delta method,

$$\sqrt{n}[g(\hat{\theta}) - g(\theta)] \longrightarrow N\left[0, \left(\frac{\partial g(\theta)}{\partial \theta'}\right)\Omega\left(\frac{\partial g(\theta)}{\partial \theta'}\right)'\right], \tag{14}$$

We derive $\partial g(\theta)/\partial\delta'$ and $\partial g(\theta)/\partial\gamma'$ as

$$\partial g(\theta)/\partial\delta' = -\sigma_i(\gamma^c z_i' + D)R_1(1 + R_3) - \sigma_i(\delta^c - \gamma^c)z_i'R_4, \tag{15}$$

$$\partial g(\theta)/\partial\gamma' = \sigma_i\gamma^c z_i'(R_1R_3 - R_2 - R_4) + \sigma_i D(R_1R_3 - R_2) + \sigma_i\delta^c z_i'R_4 \tag{16}$$

where $D = [I_{K_1} \ 0_{K_1 \times K_2}]$ is a $K_1 \times K$ dimensional matrix,

$$R_4 = R_1(1 + R_3) - R_1^2(R_2 + R_1R_3 + 2R_2R_3) \tag{17}$$

$\frac{\partial g(\theta)}{\partial \theta'} = \left[\frac{g(\theta)}{\delta'} \; \frac{g(\theta)}{\gamma'} \right]$ is a $K_1 \times 2K$ dimensional matrix, which depend on the model parameters δ and γ. We can get the estimate of $\frac{\partial g(\theta)}{\partial \theta'}$ by substituting the estimates of δ and γ into the above formulas. The variances of the partial effects can be estimated by substituting the estimate of $\frac{\partial g(\theta)}{\partial \theta'}$ as well as the estimate of the variance-covariance matrix of $\hat{\theta}$ into the formula (14).

Next we compute partial effects of dummy variables. Let z_{ik} be the dummy of concern. The partial effect of z_{ik} on $E(-u_i|x_i, z_i)$ is

$$\begin{aligned} d(\theta) &= E(-u_i|x_i, z_i, z_{ik} = 1) - E(-u_i|x_i, z_i, z_{ik} = 0) \\ &= [-\sigma_i(R_1 + R_2)]|_{z_{ik}=1} - [-\sigma_i(R_1 + R_2)]|_{z_{ik}=0} \qquad (18) \end{aligned}$$

Similarly, following the delta method, we have

$$\sqrt{n}[d(\hat{\theta}) - d(\theta)] \longrightarrow N \left[0, \left(\frac{\partial d(\theta)}{\partial \theta'} \right) \Omega \left(\frac{\partial d(\theta)}{\partial \theta'} \right)' \right] \qquad (19)$$

We then have $\partial d(\theta)/\partial \delta'$ and $\partial d(\theta)/\partial \gamma'$ as follows

$$\partial d(\theta)/\partial \delta' = [-\sigma_i R_1(R_1 + R_3) z_i']|_{z_{ik}=1} - [-\sigma_i R_1(R_1 + R_3) z_i']|_{z_{ik}=0} \qquad (20)$$

$$\partial d(\theta)/\partial \gamma' = [-\sigma_i (R_2 - R_1 R_3) z_i']|_{z_{ik}=1} - [-\sigma_i (R_2 - R_1 R_3) z_i']|_{z_{ik}=0} \qquad (21)$$

$\frac{\partial d(\theta)}{\partial \theta'} = \left[\frac{d(\theta)}{\delta'} \; \frac{d(\theta)}{\gamma'} \right]$ is a $1 \times 2K$ dimensional matrix. The variances of the partial effects for z_{ik} can be estimated similarly as for the continuous variables described earlier.

ACKNOWLEDGEMENTS

The author would like to thank Peter Schmidt and Robert Myers for their suggestions and guidance. Thanks also go to Thomas Jayne for generously providing me the data, Elliot Mghenyi, Kirimi Sindi, Tavneet Suri, Zhiying Xu, and especially Margaret Beaver for helping me understand the data. All remaining errors and omissions are my own.

References

[1] Ahmad M, Bravo-Ureta BE. 1995. An Econometric Decomposition of Dairy Output Growth. *American Journal of Agricultural Economics* 77: 914-921.

[2] Ali M, and Flinn JC. 1989. Profit Efficiency Among Basmati Rice Producers in Pakistan Punjab. *American Journal of Agricultural Economics* 71: 303-310.

[3] Alvarez A, Arias C. 2004. Technical Efficiency and Farm Size: a Conditional Analysis. *Agricultural Economics* 30: 241-250.

[4] Alvarez A, Amsler C, Orea L, Schmidt P. 2006. Interpreting and Testing the Scaling Property in Models Where Inefficiency Depends on Firm Characteristics. *Journal of Productivity Analysis* forthcoming.

[5] Aigner DJ, Lovell CAK, Schmidt P. 1977. Formulation and Estimation of Stochastic Frontier Production Functions. *Journal of Econometrics* 6: 21-37.

[6] Barrett C. 1996. On Price Risk and the Inverse Farm Size - Productivity Relationship. *Journal of Development Economics* 51: 193-216.

[7] Battese GE, Coelli TJ. 1995. Frontier Production Functions, Technical Efficiency and Panel Data: With Applications to Paddy Farmers in India. *Journal of Productivity Analysis* 3: 153-169.

[8] Benjamin D. 1995. Can Unobserved Land Quality Explain The Inverse Productivity Relationship? *Journal of Development Economics* 46: 51-84.

[9] Caudill SB, Ford JM. 1993. Biases in Frontier Estimation Due to Heteroskedasticity. *Economics Letters* 41: 17-20.

[10] Caudill SB, Ford JM, Gropper DM. 1995. Frontier Estimation and Firm Specific Inefficiency Measures in the Presence of Heteroskedasticity. *Journal of Business and Economic Statistics* 13: 105-111.

[11] Hadri K. 1999. Estimation of a Doubly Heteroscedastic Stochastic Frontier Cost Function. *Journal of Business and Economic Statistics* 17: 359-363.

[12] Hazarika G, Alwang J. 2003. Access to Credit, Plot Size and Cost Inefficiency Among Smallholder Tobacco Cultivators in Malawi. *Agricultural Economics* 29: 99-109.

[13] Huang CJ, Liu JT. 1994. Estimation of a Non-Neutral Stochastic Frontier Production Function. *Journal of Productivity Analysis* 5: 171-180.

[14] Huang Y, Kalirajan K. 1997. Potential of China's grain Production: Evidence from the Household Data. *Agricultural Economics* 70: 474-475.

[15] Jacoby H. 2000. Access to Markets and the Benefits of Rural Roads. *Economic Journal* 110: 713-737.

[16] Karanja D, Jayne T, Strasberg P. 1998. Maize Productivity And Impact of Market Liberalization in Kenya. Michigan State University International Development Working Paper.

[17] Kumbhakar S, Biswas B, Bailey D. 1989. A Study of Economic Efficiency of Utah Dairy Farmers: A System Approach. *The Review of Economics and Statistics* 71: 595-604.

[18] Kumbhakar S, Ghosh S, McGuckin J. 1991. A Generalized Production Frontier Approach for Estimating Determinants of Inefficiency in US Dairy Farms. *Journal of Business and Economic Statistics* 9: 279-286.

[19] Lamb R. 2003. Inverse Productivity: Land Quality, Labor Markets, and Measurement Error. *Journal of Development Economics* 71: 71-95.

[20] Liu Y. 2006. Papers on Agricultural Insurance and Farm Productivity. Michigan State University PhD Dissertation Chapter 2.

[21] Meeusen W, van den Broeck J. 1977. Efficiency Estimation from Cobb-Douglas Production Functions with Composed Error. *International Economic Review* 18: 435-44.

[22] Nyoro J, Kirimi L, Jayne T. 2004. Competitiveness of Kenyan and Ugandan Maize Production: Challenges for the Future. Michigan State University International Development Working Paper.

[23] Parikh A, Ali F, Shah MK. 1995. Measurement of Economic Efficiency in Pakistani Agriculture. *American Journal of Agricultural Economics* 77: 675-685.

[24] Place F, Hazell P. 1993. Productivity Effects of Indigenous Land Tenure Systems in Sub-Saharan Africa. *American Journal of Agricultural Economics* 75: 10-19.

[25] Puig-Junoy J, Argiles J. 2000. Measuring and Explaining Farm Inefficiency in a panel Data Set of Mixed Farms. Pompeu Fabra University Working Paper.

[26] Reifschneider D, Stevenson R. 1991. Systematic Departures from the Frontier: A Framework for the Analysis of Firm Inefficiency. *International Economic Review* 32: 715-723.

[27] Sen A. 1962. An Aspect of Indian Agriculture. *Economics Weekly* Februray: 243-246.

[28] Sherlund S, Barrett C, Adesina A. 2002. Smallholder Technical Efficiency Controlling for Environmental Production Conditions. *Journal of Development Economics* 69: 85-101.

[29] Simar L, Lovell CAK, van den Eeckaut P. 1994. Stochastic Frontiers Incorporating Exogenous Influences on Efficiency. Discussion Papers No. 9403. Institut de Statistique, Universite Catholique de Louvain.

[30] Stevenson RE. 1980. Likelihood Functions for Generalized Stochastic Frontier Estimation. *Journal of Econometrics* 13: 57-66.

[31] Suri T. 2005. Selection and Comparative Advantage in Technology Adoption. Yale University Job Market Paper.

[32] Wang HJ. 2002. Heteroscedasticity and Non-Monotonic Efficiency Effects of a Stochastic Frontier Model. *Journal of Productivity Analysis* 18: 241-253.

[33] Wang HJ. 2003. A Stochastic Frontier Analysis of Financing Constraints on Investment: The Case of Financial Liberalization in Taiwan. *Journal of Business and Economic Statistics* 21: 406-419.

[34] Wang HJ, Schmidt P. 2002. One-Step and Two-Step Estimation of the Effects of Exogenous Variables on Technical Efficiency Levels. *Journal of Productivity Analysis* 18: 129-144.

[35] Wilson P, Hadley D, Asby C. 2001. The Influence of Management Characteristics on the Technical Efficiency of Wheat Farmers in Eastern England. *Agricultural Economics* 24: 329-338.

Figure 1: Kernel density of Battese and Coelli technical efficiency estimates

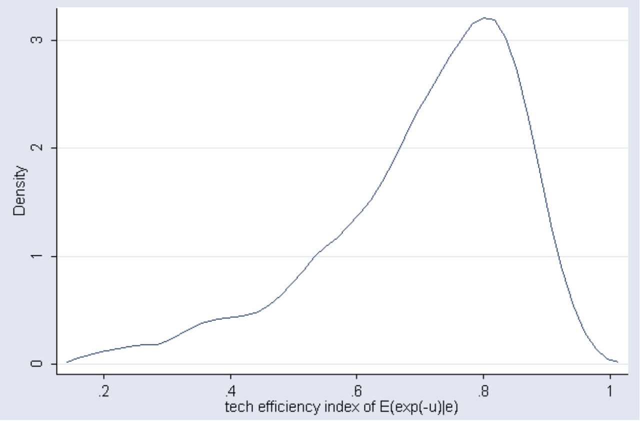

Table 1: Descriptive statistics for the variables in the production frontier

Variable	Notation	Mean	Std. Dev.	Min	Max
YIELD	Maize yield index (kg/acre)	1071	726	69	4410
LABOR	Pre-harvest labor input (person-hour/acre)	344	271	40	2160
FERTILIZER	Nitrogen fertilizer application (kg/acre)	11	12	0	63
SEED	Maize seed quantity (kg/acre)	8.5	3.3	2.5	18.8
TRACTOR	If tractor used in land preparation (1=yes, 0=no)	0.28	0.45	0	1
MONO	If mono-crop field (1=yes, 0=no)	0.11	0.31	0	1
HYBRID	If hybrid seed (1=yes, 0=no)	0.72	0.45	0	1
STRESS	Moisture stress (0-1)	0.14	0.21	0	1
DRAINAGE	Drainage of soil (categorical 1-10)	7.2	2.1	1	10

Table 2: Descriptive statistics for the exogenous variables in the efficiency model

Variable	Notation	Mean	Std Dev	Min	Max
EDUHIGH	# school years for the highest educated member	12	5.5	0	24
FEMHEAD	If the household head is female (1=yes, 0=no)	0.19	0.39	0	1
DISTBUS	Distance to the nearest bus-stop (km)	2.4	2.4	0	20
DISTPHONE	Distance to the nearest phone service (km)	0.78	1.6	0	15
DISTEXTN	Distance to the nearest extension service (km)	5.2	4.5	0	33
OWNED	If the field owned by the household (1=yes, 0=no)	0.86	0.35	0	1
CRDCSTR	If pursued credits and was rejected (1=yes, 0=no)	0.08	0.27	0	1
RNFINC	Percentage of members that have non-farming income	0.20	0.19	0	1
TTACRES	Total acres of land owned by the household	7.46	10.9	0.13	110
ACRES	Acres of the field	1.46	2.01	0.03	27

Table 3: Estimates for the efficiency components in alternative models

LYIELD	General	Scaled Stevenson	KGMHLBC	RSCFG-μ	RSCFG
	Variables in function μ_i				
μ	-4.1(6.9)	-0.30(0.36)	-1.45(0.72)	-0.75(0.40)	0
EDUHIGH	0.034(0.049)	-0.018(0.0068)	0.053(0.024)		
FEMHEAD	-5.3(41)	0.22(0.093)	-2.3(2.0)		
DISTBUS	-0.36(0.16)	0.048(0.016)	-0.31(0.14)		
OWNED	-1.4(1.0)	0.35(0.11)	-1.3(0.41)		
RNFINC	0.82(1.2)	-0.36(0.19)	1.4(0.73)		
TTACRES	0.0018(0.045)	-0.013(0.003)	0.024(0.012)		
	Variables in function σ_i^2				
σ_u^2	2.7(5.9)	0.42(0.13)	0.59(0.14)	0.54(0.12)	0.34(0.11)
EDUHIGH	-0.0063(0.015)	-0.018(0.0068)		-0.014(0.0048)	-0.032(0.014)
FEMHEAD	-0.22(0.28)	0.22(0.093)		0.18(0.072)	0.41(0.17)
DISTBUS	-0.014(0.044)	0.048(0.016)		0.040(0.012)	0.087(0.030)
OWNED	-0.061(0.46)	0.35(0.11)		0.28(.073)	0.63(0.22)
RNFINC	-0.14(0.36)	-0.36(0.19)		-0.29(0.15)	-0.63(0.38)
TTACRES	-0.012(0.013)	-0.013(0.003)		-0.011(0.0015)	-0.020(0.014)
# observations	815	815	815	815	815
Log-likelihood	-616.30	-623.63	-618.71	-623.42	-623.70
LR statistic	56.84	34.54	50.62	38.36	37.98
Wald statistic	26.80	18.28	29.74	77.69	27.17
1% critical value	26.22	16.81	16.81	16.81	16.81

Table 4: Estimates for the production frontier in alternative models

LYIELD	General	Scaled Stevenson	KGMHLBC	RSCFG-μ	RSCFG
LFERTILIZER	0.15 (0.020)	0.15 (0.020)	0.15 (0.020)	0.15 (0.020)	0.15 (0.020)
LLABOR	0.33 (0.050)	0.33 (0.052)	0.33 (0.049)	0.33 (0.052)	0.33 (0.052)
LSEED	0.33 (0.048)	0.32 (0.050)	0.33 (0.048)	0.32 (0.050)	0.32 (0.050)
LFERTILIZER^2	0.025 (0.004)	0.026 (0.004)	0.026 (0.004)	0.026 (0.004)	0.026 (0.004)
LFERTILIZER×HYBRID	-0.062 (0.016)	-0.063 (0.016)	-0.063 (0.016)	-0.063 (0.016)	-0.063 (0.016)
LLABOR×HYBRID	-0.16 (0.059)	-0.15 (0.061)	-0.16 (0.059)	-0.16 (0.061)	-0.15 (0.060)
LLABOR×STRESS	-0.23 (0.14)	-0.29 (0.14)	-0.26 (0.14)	-0.29 (0.14)	-0.29 (0.14)
LSEED×STRESS	-0.29 (0.17)	-0.28 (0.19)	-0.29 (0.17)	-0.27 (0.20)	-0.29 (0.19)
HYBRID	0.19 (0.063)	0.20 (0.059)	0.20 (0.063)	0.20 (0.059)	0.20 (0.059)
STRESS	-0.38 (0.18)	-0.36 (0.18)	-0.39 (0.18)	-0.36 (0.18)	-0.37 (0.18)
MONO	-0.22 (0.059)	-0.21 (0.060)	-0.23 (0.058)	-0.21 (0.060)	-0.21 (0.60)
DRAINAGE	0.15 (0.056)	0.13 (0.056)	0.15 (0.055)	0.13 (0.057)	0.13 (0.056)
DRAINAGE^2	-0.012 (0.005)	-0.001 (0.005)	-0.011 (0.005)	-0.001 (0.005)	-0.001 (0.005)
TRACTOR	0.15 (0.056)	0.15 (0.051)	.15 (0.057)	0.14 (0.050)	0.15 (0.051)
Zone Dummies	Omitted	Omitted	Omitted	Omitted	Omitted
σ_v^2	0.16 (0.023)	0.14 (0.023)	0.15 (0.020)	0.15 (0.022)	0.13 (0.021)

Note: LYIELD is log YIELD. LFERTILIZER, LLABOR and LSEED are defined similarly.

Table 5: Correlation of efficiency estimates among alternative models

	General	Scaled Stevenson	KGMHLBC	RSCFG-μ	RSCFG
General	1				
Scaled Stevenson	0.9793	1			
KGMHLBC	0.9910	0.9848	1		
RSCFG-μ	0.9839	0.9986	0.9843	1	
RSCFG	0.9700	0.9970	0.9833	0.9917	1

Table 6: Partial effects of exogenous factors, evaluated at the sample mean

	General	Scaled Stevenson	KGMHLBC	RSCFG-μ	RSCFG
	Partial effects on $E(-u_i \mid x_i, z_i)$				
EDUHIGH	.0080(.0044)	.0079(.0012)	.0052(.0044)	.0080(.00081)	.0081(.0029)
FEMHEAD	-.12(.11)	-.10(.051)	-.14(.058)	-.11(.049)	-.11(.052)
DISTBUS	-.037(.025)	-.021(.0038)	-.037(.016)	-.022(.0028)	-.022(.0083)
OWNED	-.19(.074)	-.14(.047)	-.17(.052)	-.14(.042)	-.14(.058)
RNFINC	.19(.12)	.16(.039)	.13(.11)	.17(.028)	.16(.090)
TTACRES	.0075(.0021)	.0058(.00067)	.0023(.0015)	.0061(.00040)	.0049(.0023)
	Partial effects on $V(u_i \mid x_i, z_i)$				
EDUHIGH	-.0042(.0020)	-.0045(.0015)	-.0024(.0020)	-.0044(.0012)	-.0045(.0016)
FEMHEAD	.035(.058)	.064(.037)	.066(.026)	.063(.034)	.065(.038)
DISTBUS	.016(.013)	.012(.0055)	.017(.0072)	.012(.0049)	.012(.0057)
OWNED	.083(.040)	.070(.029)	.078(.021)	.068(.026)	.071(.035)
RNFINC	-.097(.062)	-.091(.048)	-0.061(.050)	-.091(.043)	-.088(.051)
TTACRES	-.0046(.0016)	-.0033(.0011)	-.0011(.00070)	-.0033(.00083)	-.0028(.0014)

Table 7: Average partial effects of EDUHIGH on $E(-u_i|x_i, z_i)$, for the observations within each of the four quartiles based on efficiency levels predicted in KGMHLBC model

	General	Scaled Stevenson	KGMHLBC	RSCFG-μ	RSCFG
0-25% percentile	0.0067	0.0092	0.0039	0.0092	0.0092
25-50% percentile	0.0074	0.0085	0.0052	0.0085	0.0085
50-75% percentile	0.0078	0.0080	0.0059	0.0081	0.0081
75-100% percentile	0.0079	0.0069	0.0072	0.0070	0.0071

Table 8: Correlation of partial effects of EDUHIGH on $E(-u_i|x_i, z_i)$ among alternative models

	General	Scaled Stevenson	KGMHLBC	RSCFG-μ	RSCFG
General	1				
Scaled Stevenson	-0.3910	1			
KGMLBC	0.7811	-0.7899	1		
RSCFG-μ	-0.3716	0.9991	-0.7861	1	
RSCFG	-0.4140	0.9882	-0.8047	0.9970	1

Table 9: Results of the specification tests for model selection, taking the general model as the unrestricted model

	Scaled Stevenson	KGMHLBC	RSCFG-μ	RSCFG	Stevenson	ALS
log-likelihood	-623.63	-618.71	-623.42	-623.70	-641.44	-642.04
LR statistics	14.66	4.82	14.24	14.80	50.28	51.48
# restrictions	6	6	6	7	12	13
1% critical value	16.81	16.81	16.81	18.48	26.22	27.69
5% critical value	12.59	12.59	12.59	14.07	21.03	22.36
10% critical value	10.64	10.64	10.64	12.02	18.55	19.81

The value of log-likelihood for the general model is -616.30.

Table 10: Partial effect of the exogenous factors on $E(-u_i|x_i, z_i)$ and their 90% confidence intervals based on bootstrap and the delta method in the KGMHLBC model, evaluated at the sample mean

	EDUHIGH	FEMHEAD	DISTBUS	OWNED	RNFINC	TTACRES
	.0052	-.14	-.037	-.19	.13	.0023
Bootstrap	(.00047, .011)	(-.22, -.048)	(-.058, -.0078)	(-.28, -.035)	(-.011, .30)	(.00011, .0053)
Delta method	(-.0020, .012)	(-.24, -.045)	(-.063, -.011)	(-.26, -.084)	(-.051, .31)	(-.0017, .0048)

Table 11: Output elasticity with respect to inputs for local seed users and hybrid seed users, evaluated at the sample means

Inputs	Local seed users	Hybrid seed users
FERTILIZER	0.209 (.00076)	0.224 (.0011)
LABOR	0.300 (.0027)	0.177 (.0063)
SEED	0.293 (.0032)	0.336 (.0026)