DOI 10.1007/s11123-008-0111-9

Model selection in stochastic frontier analysis with an application to maize production in Kenya

Yanyan Liu · Robert Myers

Published online: 23 September 2008

Abstract This paper shows how to compute the standard errors for partial effects of exogenous firm characteristics influencing firm inefficiency under a range of popular stochastic frontier model specifications. We also develop an R^2-type measure to summarize the overall explanatory power of the exogenous factors on firm inefficiency. The paper also applies a recently developed model selection procedure to choose among alternative stochastic frontier specifications using data from household maize production in Kenya. The magnitude of estimated partial effects of exogenous household characteristics on inefficiency turns out to be very sensitive to model specification, and the model selection procedure leads to an unambiguous choice of best model. We propose a bootstrapping procedure to evaluate the size and power of the model selection procedure. The empirical application also provides further evidence on how household characteristics influence technical inefficiency in maize production in developing countries.

Keywords Stochastic frontier model · Model selection · Bootstrapping · Maize production in Kenya

JEL Classification C52

Stochastic production frontier analysis has been widely used to study technical inefficiency 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 inefficiency is measured, and a one-sided error term which captures technical inefficiency. In early applications the one-sided error was assumed to be identically and independently distributed across firms, but more recent studies have allowed its distribution to be heterogeneous and depend on various firm characteristics (see Battese and Coelli 1995; Caudill et al. 1995; Wang 2002, 2003).

Allowing inefficiency to depend on firm characteristics allows researchers to identify the determinants of inefficiency and to suggest possible policy or behavioral responses which might improve efficiency. However, this approach has been hampered by two problems. First, existing studies have mostly focused on the *directions* of the influence of the firm characteristics on technical inefficiency while generally overlooking the *magnitudes* of the partial effects. This makes it difficult to determine which type of policy intervention will have the largest impact on inefficiency. This problem is somewhat surprising given that the magnitudes of the effects of explanatory variables on dependent variables are often the focal point in other types of regression analyses. Second, the relationship between firm characteristics and technical inefficiency is often sensitive to the model used to incorporate firm characteristics, and choosing between competing models is difficult (see Alvarez et al. 2006, hereafter AAOS). This model uncertainty makes policy recommendations quite tenuous.

In this paper we make four contributions to the stochastic frontier literature. First, we take the formulas for estimating the partial effects of exogenous firm characteristics on firm inefficiency that have been proposed by Wang (2002, 2003) and show how to put standard errors

Y. Liu (✉)
The World Bank, 1818 H Street, N.W., MC3-452, Washington, DC 20433, USA
e-mail: yliu3@worldbank.org

R. Myers
Michigan State University, East Lansing, MI, USA

around these point estimates. The standard errors are computed using the delta method and will be useful in assessing the precision of estimated partial effects. Second, we propose an R^2-type measure to summarize the overall explanatory power of the exogenous factors on inefficiency. To date, there has been no way of assessing the overall power of firm characteristics to explain variations in inefficiency across firms. Our R^2-type measure provides an easily computed means of doing this. Third, we propose a bootstrapping procedure to evaluate the power of the recently developed model selection procedure suggested by AAOS to choose among competing models of the influence of firm characteristics on inefficiency. This bootstrapping procedure should prove useful in various applications. Fourth and finally, we apply our procedures to an empirical application of stochastic frontier analysis of maize production in Kenya. In the application we apply the model selection procedures of AAOS and use bootstrapping to evaluate the power of the procedure. We then compute point estimates of partial effects of farm characteristics on inefficiency and their standard errors. We also use our R^2-type measure to evaluate the joint explanatory power of the farm characteristics. We find that while alternative models of the relationship between farm household characteristics and technical inefficiency in maize production in Kenya tend to provide the same direction of the influence of household characteristics, the magnitudes of the partial effects on firm inefficiency are quite sensitive to model selection.

In the remainder of the paper, we first review the standard stochastic frontier production model and then extend it to provide: (i) standard errors around point estimates of the effects of firm characteristics on technical inefficiency; and (ii) an R^2-type measure of the overall explanatory power of firm characteristics on inefficiency. Next we describe our data and variables used in the empirical application to Kenyan maize production, followed by estimation results from alternative model specifications. Results for the magnitude of partial effects of farm household characteristics on inefficiency are quite sensitive to model specification. Then we carry out AAOS specification tests to choose a final model and use our bootstrapping procedure to examine the reliability of these specification tests in choosing the correct model. The final sections contain an analysis of technical inefficiency in maize production in Kenya based on the final model chosen, and some concluding comments.

1 Stochastic production frontier models

The basic setup and notation follow Wang and Schmidt (2002) and AAOS. Firms are indexed by $i = 1,\ldots,N$. Let y_i be log output; x_i be a vector of inputs; and z_i be a vector of exogenous variables that exert influence on firm inefficiency. Let y_i^* be the unobserved frontier which is modeled 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, and β is a parameter vector. The actual log output level y_i equals y_i^* less a one-sided error, u_i, whose distribution depends on z_i. The full 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 one another and that u_i is independent of x_i (conditional on z_i). The model is usually implemented by assuming u_i is distributed as $N(\mu_i, \sigma_i^2)^+$ with various specifications (discussed below) used to model μ_i and σ_i. The frontier function and the inefficiency part are generally estimated in one step using maximum likelihood estimation (MLE) to achieve both efficiency and consistency.[1]

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 Eq. 2,[2] to get

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

The term $\partial[E(-u_i|x_i, z_i)]/\partial z_{ik}$ can be interpreted as the partial effect of z_{ik} on efficiency and from (3) is also the partial effect on y_i. Because y_i is log output, $\partial[E(-u_i|x_i, z_i)]/\partial z_{ik}$ is the semi-elasticity of output (efficiency) with respect to the exogenous factors (i.e., the percentage change in expected output when z_{ik} increases by one unit). Similarly, taking conditional variances 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 y_i. It can be interpreted as an estimator of the partial effect of z_{ik} on production uncertainty.

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 proposed and used in Wang (2002, 2003), but a means for computing their standard errors was not

[1] 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 is biased 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.

[2] Here we assume there is no overlap between x and z, i.e., no variable appears in both the frontier component (x) and the inefficiency component (z). It is straightforward to generalize the following equations to allow for overlap between x and z.

provided. In the Appendix we provide formulas for computing estimates of $\partial[E(-u_i|x_i, z_i)]/\partial z_{ik}$ and $\partial[V(u_i|x_i, z_i)]/\partial z_{ik}$, along with their standard errors using the delta method for several popular model specifications for μ_i and σ_i.

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

$$V(u_i) = V_z[E(u_i|z_i)] + E_z[V(u_i|z_i)], \tag{5}$$

where $V_z[E(u_i|z_i)]$ is the variance of the conditional mean function over the distribution of z_i, and $E_z[V(u_i|z_i)]$ is the expected variance around the conditional mean of u_i. The fraction of variation in u_i that is explained by z_i is $V_z[E(u_i|z_i)]/V(u_i)$. Thus a natural measure of explanatory power over the sample would be

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

where $\hat{E}$ and $\hat{V}$ indicate sample estimates of the mean and variance of u_i conditional on z_i. Letting $R_1 = \mu_i/\sigma_i$, $R_2 = \phi(R_1)[\Phi(R_1)]^{-1}$, and $R_3 = -R_2^2 - R_1R_2$, where $\phi(\cdot)$ and $\Phi(\cdot)$ are the density and cumulative density functions for the standard normal, then the mean and variance of u_i conditional on z_i can be expressed as

$$E(u_i|x_i, z_i) = \sigma_i \cdot (R_1 + R_2) \tag{7}$$

$$V(u_i|x_i, z_i) = \sigma_i^2 \cdot (1 + R_3). \tag{8}$$

So all that remains to compute R_z^2 is to estimate $\hat{\mu}_i$ and $\hat{\sigma}_i$ for a specific model specification (see the model specification section below) and then substitute these estimators for the population values μ_i and σ_i in (7) and (8) to get the sample estimates of $\hat{E}(u_i|z_i)$ and $\hat{V}(u_i|z_i)$.

Similar 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.

1.1 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 while maintaining the assumption of half normality. Kumbhakar et al. (1991), Huang and Liu (1994), and Battese and Coelli (1995) allow the mean of the pre-truncated normal distribution of u_i to depend on a set of exogenous factors. Reifschneider and Stevenson (1991), Caudill and Ford (1993), Caudill et al. (1995) and Hadri (1999) allow the variance of the pre-truncated normal distribution of u_i to depend on the exogenous factors. Wang (2003) allows both the mean and the variance of the pre-truncated distribution of u_i to depend on exogenous factors.

Regardless of whether we allow the mean, the variance, or both the mean and the variance of the *pre-truncated* normal to depend on exogenous factors, both the mean and the variance of the *truncated* half normal will always depend on the exogenous factors. These are sometimes called models of heteroscedasticity, but the fact that the mean also changes makes this terminology potentially misleading. Whereas heteroscedasticity affects only the efficiency of estimation in a standard linear model, in a stochastic frontier model with heterogeneity in the distribution of the inefficiency term, failure to model the exogenous factors appropriately leads to biased estimation of the production frontier model and of the level of technical inefficiency, hence leading to poor policy conclusions (see 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 used in particular applied settings. 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 potential bias and inconsistency (Wang 2003). Moreover, there has been little investigation of how the choice of model specification influences the estimation results. 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 that nests most of the usual model specifications for μ_i and σ_i. However, a more flexible model has more parameters, which imposes a higher computational burden and reduces degrees of freedom. Given that large samples are typically difficult to obtain in stochastic frontier estimation, some relevant parameters may be estimated imprecisely in flexible model specifications. Even when large samples are available, finding an appropriate parsimonious model can still improve performance and a more flexible model specification may not always be preferred.

AAOS suggest a procedure for selecting a 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, as outline in AAOS.

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).
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. Also 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.

2 Empirical application

The empirical application is to maize production in Kenya using detailed household survey data. The problem of hunger in Kenya remains widespread and its economy depends heavily on agriculture with 75% of Kenyans making their living from farming. Maize is the primary staple food and most farmers are engaged in maize production. In recent years, total maize output has not kept pace with growing population and demand, largely due to falling land productivity: average national maize yields have fallen from over 2 tons/ha in the early 1980s to about 1.6 tons/ha recently (Nyoro et al. 2004). The technical efficiency level of Kenya maize production is therefore an important economic and policy issue.

2.1 Data

The data are from a rural household survey of about 1,100 households planting 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. Figure 1 is a map of Kenya with the round dots representing sampled villages. These villages were chosen randomly from each of eight predetermined agro-economic zones and then households were sampled randomly from each selected village.

Field level data are available for each sampled household 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 the field, yield, labor input, fertilizer application, and seed usage. The demographic information 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-farm 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.

2.2 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:

$$Y_i = \frac{\sum_j Y_{ij} P_j}{P_1}, \tag{9}$$

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. 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 (LABOR) 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

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

[4] Six hundred and thirty-seven out of the total 1,718 fields are dropped.

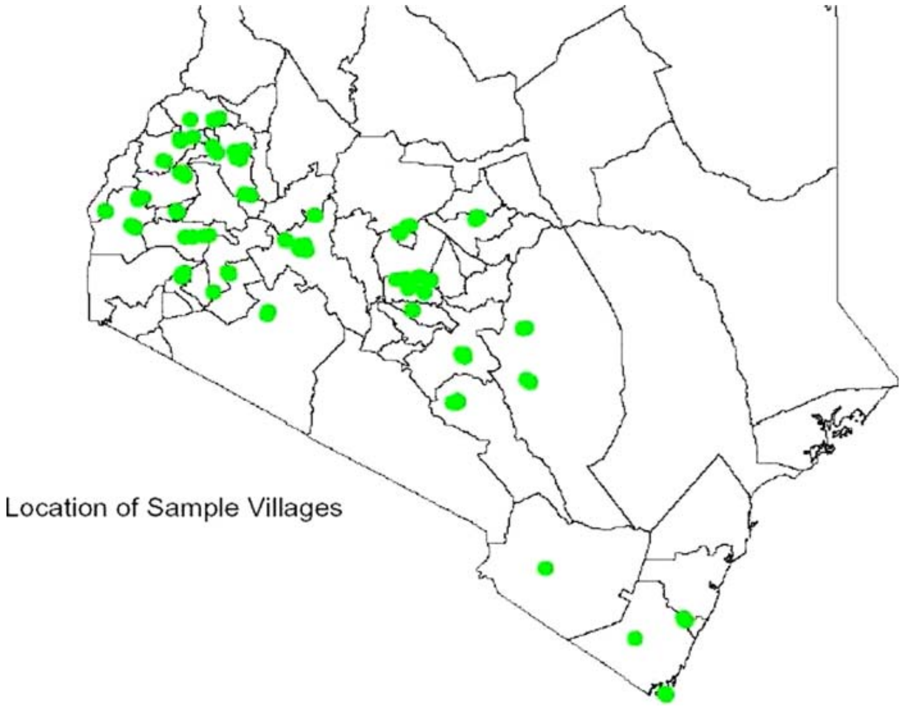

Fig. 1 Location of sample villages in Kenya (*source*: Suri 2005)

labor. Nitrogen (FERTILIZER), 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] Maize seeds can be separated into hybrid seeds and local seeds. All fields used either hybrid seeds or local seeds (no combinations in the same field). These seed inputs are captured by two variables, SEED measures the amount of (hybrid or local) seed per acre applied to the field, and HYBRID is a dummy variable measuring one for hybrid seeded fields and zero otherwise. We also use a dummy variable MONO as an indicator for mono-crop fields because these might be expected to have systematically different yields than multi-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 one indicating that a tractor was used and zero otherwise.

Environmental variables are also included on the right hand side of the frontier production function. Failure to control for environmental variables may cause a correlation between some inputs and unobserved factors in the error term (for example, if a farmer makes input decisions based on 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 seven dummy variables indicating the different agro-economic 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 ten where one indicates the least and ten the highest drainage. DRAINAGE^2 is the square of DRAINAGE. We include a quadratic term because yield is expected to increase in DRAINAGE at lower drainage levels and decrease at higher levels. Rainfall is a very important factor in maize production in Kenya because all of the maize fields are rain-fed and drought is the usual cause of yield loss. We use the 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 mm of rain during the 2003–2004 main season. This is a better measure for moisture conditions than total rainfall because total rainfall does not reflect the distribution of rainfall over time, which is very important in maize growth.

Any observations with missing values were discarded. Because of potential 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 4,580 kg per acre, (2) seed usage less than 2 kg per acre or more than 20 kg per acre, and (3) labor input less than 40 person-hours per acre or more than 2,200 person-hours per acre. After these filters were applied, there are 815 fields

[5] More than 20 types of fertilizers were applied. While some of these use nitrogen, phosphorous, and other nutrients in various proportions, nitrogen is usually the major nutrient deficiency and many of the major fertilizers use nitrogen and phosphorous in fixed proportions. Therefore, the level of applied nitrogen should give a reasonably accurate measure of the impact of fertilizer on yields.

Table 1 Descriptive statistics for the variables in the production frontier

Variable	Notation	Mean	SD	Min	Max
YIELD	Maize yield index (kg/acre)	1,071	726	69	4,410
LABOR	Pre-harvest labor input (person-hour/acre)	344	271	40	2,160
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

(observations) remaining. 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).

2.3 Exogenous factors affecting efficiency

Previous studies have identified numerous factors that may limit farm productivity and efficiency. Education is arguably an important factor and Kumbhakar et al. (1989) find that education increases the productivity of labor and land on Utah dairy farms while Kumbhakar 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. Here we measure education with EDUHIGH, the highest level of education among all household members.[6] We also investigate gender effects by including a dummy variable for female-headed households (FEMHEAD).

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 Nepalese 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. Here we use three infrastructure variables to account for these effects on efficiency—DISTBUS, distance of the house from the nearest bus stop;[7] DISTPHONE, distance of the house from the nearest telephone or mobile phone service; and DISTEXTN, distance from the nearest extension service office.

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 land tenure is 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. Here we use a dummy variable (OWNED) with one indicating that the field is owned by the household and zero indicating the field is rented.

Financial constraints, such as limited access to credit, might also 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 effects of financial constraints on technical efficiency seem to be unexamined to date but may be important because the timing of input usage can be an important factor influencing yields. When farmers face financial constraints, they may resort to relatives or friends for loans or try to obtain in-kind inputs through governments or other input subsidy programs. These extra efforts may prevent them from applying inputs at the right times to optimize productivity. We attempt to capture this effect using CRDCSTR (a dummy variable with one indicating the household has unsuccessfully pursued credit and zero otherwise), and RNFINC (the proportion of household members that have non-farm income).

The 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). Empirical results on the

[6] EDUHIGH may capture the effects of education 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.

[7] 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.

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

Variable	Notation	Mean	SD	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

relationship between efficiency and farm size have been mixed. 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. Here we include farm size (TTACRES) and field size (ACRES) as measures of the size effect. Descriptive statistics for the household survey data used to define the exogenous factors affecting efficiency are summarized in Table 2.

3 Estimation results from competing models

In this section, we report the estimation results under alternative model specifications for the inefficiency component of the model. We use a flexible translog functional form for FERTILIZER, LABOR, and SEED in the frontier production function. We also interact the dummy variable for hybrid maize (HYBRID), and the variable for moisture stress (STRESS), with FERTILIZER, LABOR, and SEED because there may be important interactions between these variables. In order to obtain more precise estimation and simplify computation, we drop jointly insignificant variables based on LR tests and the 10% significance level.[8] The dropped variables are the second order effects for LABOR and SEED, all interaction effects among FERTILIZER, LABOR, and SEED, interaction effects for SEED and HYBRID, and FERTILIZER and STRESS, two zone dummy variables, DISTPHONE, DISTEXTN, CRDCSTR, and ACRES. Tables 3 and 4 report the MLE estimation results for alternative specifications for the inefficiency part, after dropping the jointly insignificant variables.

The parameter estimates for the frontier part of the model are very similar across alternative models for the inefficiency component (see Table 3). Furthermore, Both the LR test and Wald test reject the null hypothesis that all the exogenous factors have zero effect on inefficiency at the 1% significance level in each of the five models (see Table 4). Hence, it seems clear that the exogenous factors have a statistically significant effect irrespective of the model specification employed to model inefficiency. The Battese and Coelli efficiency estimates are computed for each observation in all the models and their correlations across alternative models are reported in Table 5. The lowest correlation is 0.97. Therefore, all five models yield similar results for the production frontier and for the rankings of inefficiency among households, consistent with previous studies (e.g. Caudill et al. 1995).

The goodness of fit statistics for the inefficiency component, R_z^2, are reported at the bottom of Table 4 for the alternative model specifications. For example, the value of R_z^2 for the KGMHLBC model is 0.1035, indicating that 10.35% of the sample variation in inefficiency can be explained by the exogenous factors. The remaining 89.65% of the sample variation is due to other unobserved factors (such as managerial skill). Not surprisingly, the general model provides the best fit at 12.75%.

The coefficients of the exogenous factors reported in Table 4 are not very interesting by themselves because they are the parameters of the pre-truncated distribution of the inefficiency term u_i. So these parameters do not tell us how the exogenous factors affect the distribution of u_i. 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 computing these measures and their standard errors for the general model are provided

[8] The LR test results are available from the authors on request.

Table 3 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)
Constant & Zone Dummies	not reported				
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. Standard errors are in parentheses

Table 4 Estimates for the inefficiency 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)	0	0
FEMHEAD	−5.3 (41)	0.22 (0.093)	−2.3 (2.0)	0	0
DISTBUS	−0.36 (0.16)	0.048 (0.016)	−0.31 (0.14)	0	0
OWNED	−1.4 (1.0)	0.35 (0.11)	−1.3 (0.41)	0	0
RNFINC	0.82 (1.2)	−0.36 (0.19)	1.4 (0.73)	0	0
TTACRES	0.0018 (0.045)	−0.013 (0.003)	0.024 (0.012)	0	0
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	−0.014 (0.0048)	−0.032 (0.014)
FEMHEAD	−0.22 (0.28)	0.22 (0.093)	0	0.18 (0.072)	0.41 (0.17)
DISTBUS	−0.014 (0.044)	0.048 (0.016)	0	0.040 (0.012)	0.087 (0.030)
OWNED	−0.061 (0.46)	0.35 (0.11)	0	0.28 (.073)	0.63 (0.22)
RNFINC	−0.14 (0.36)	−0.36 (0.19)	0	−0.29 (0.15)	−0.63 (0.38)
TTACRES	−0.012 (0.013)	−0.013 (0.003)	0	−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
R_z^2	0.1275	0.0848	0.1035	0.0936	0.0773

Note: Standard errors are in parentheses. The LR and Wald statistics test the null hypothesis that the exogeneous factors have no joint influence on inefficiency

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 $E(-u_i\|x_i, z_i)$					
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 $V(u_i\|x_i, z_i)$					
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)

Note: Standard errors are in parentheses

in the Appendix. To obtain the formulas for the nested models, we only need to impose the corresponding restrictions on the parameters.[9]

The partial effects of the exogenous factors evaluated at the sample mean are reported in Table 6 along with their standard errors. 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. So conclusions about the semi-elasticity of output with respect to farm size may differ by a factor of more than 100%, depending on which inefficiency model is used.

Table 7 reports the average partial effects of EDUHIGH on $E(-u_i|x_i, z_i)$ for alternative model specifications over observations within each of the four quartiles of the efficiency levels.[10] 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. So using the KGMHLBC model we would conclude that the households with lower efficiency levels would not benefit as much from increased education as the ones with higher efficiency levels. However, an opposite conclusion would follow if we use the scaled Stevenson model, the RSCFG-μ model or the RSCFG model.[11]

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 (see footnote 11). This further confirms 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 critical. However, if we are interested in the magnitudes of the yield semi-elasticities, it is important to choose the appropriate model specification.

[9] Wang (2002) gives the expressions for these derivatives but not for the standard errors.

[10] The quartiles were computed using the KGMHLBC model.

[11] Similar patterns are observed for the other exogenous factors but these results are not reported to conserve space.

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

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

4 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.

4.1 Empirical model selection

We start with the general model, and then use LR tests to find simpler models that the data do not reject. 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. LR test statistics with Chi-squared critical values are listed in Table 9 and provide the following results:

- We 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 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.

4.2 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. However, it is important 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 approach has not been used previously in the literature. 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.

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 specification tests for model selection

	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)

The KGMHLBC model is written as

$$y_i = x_i'\beta + v_i - u_i, \quad \text{where} \quad u_i \sim N[\mu \cdot \exp(z_i'\delta), \sigma_u^2]^+ \quad \text{and} \quad v_i \sim N(0, \sigma_v^2). \tag{10}$$

We take the following steps to conduct the parametric bootstrap:

1. Using the actual sample data $\{(y_i, x_i, z_i)\}_{i=1}^n$, estimate the KGMLBC model using MLE to get $\hat{\theta} = \{\hat{\beta}, \hat{\delta}, \hat{\mu}, \hat{\sigma}_u^2, \hat{\sigma}_v^2\}$. These results are provided in Tables 3 and 4.
2. Next generate pseudo-data sets based on the parameter estimates from step 1. That is, for $i = 1,\ldots,n$, draw u_i^* from $N[\hat{\mu} \cdot \exp(z_i'\hat{\delta}), \hat{\sigma}_u^2]^+$, v_i^* from $N(0, \hat{\sigma}_v^2)$, and then compute $y_i^* = x_i'\beta + v_i^* - u_i^*$.
3. Based on the pseudo-data $\{y_i^*, x_i, z_i\}_{i=1}^n$ generated in step 2, estimate all seven inefficiency models using MLE. Take the log-likelihood value (ll^*) and parameter estimates ($\hat{\theta}^*$) in each of the models, denoted as $\zeta^* = \{(ll_j^*, \hat{\theta}_j^*)\}_{j=1}^7$, where j indexes the different models.
4. Repeat steps 2 and 3 1,000 times to obtain $\mathcal{B} = \{\zeta_b^*\}_{b=1}^{1000}$.

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

- 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 1,000 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. We reject the true model and accept an alternative one at the same time in only 0.4% of the data sets.
- 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.

These results suggest that the AAOS model selection criteria do a good job of discriminating between models. 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 also 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 $\mathcal{B}$ to compute the partial effects for every observation in each pseudo-data set. Confidence intervals then follow directly from the set of $\mathcal{B}$ estimates. For example, given 1,000 pseudo-data sets 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 computed as in the appendix and reported 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.

5 Post-estimation analysis

Post-estimation analysis is based on the results of our selected KGMHLBC model. Table 11 reports output elasticity estimates for local seed users and hybrid seed users

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)

Note: Standard errors are in parentheses

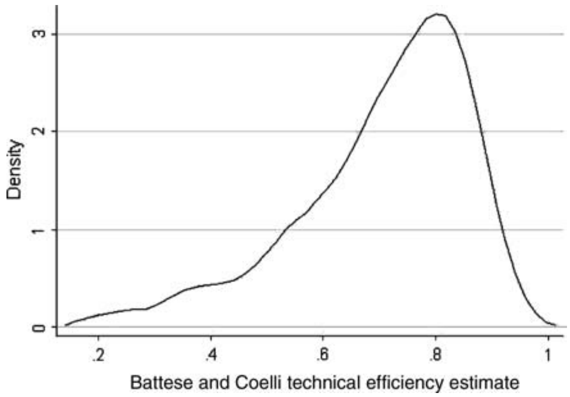

Fig. 2 Kernel density estimate based on Battese and Coelli technical efficiency estimates

calculated at their respective sample means with their standard errors in parentheses.[12] The sum of the output elasticities with respect to FERTILIZER, LABOR, and SEED is less than 1 (0.80 for local seed users and 0.74 for hybrid seed users). However, this is expected and does not mean the technology is decreasing returns to scale because we are holding land constant (production is measured as yield per acre). Results show that output elasticities with respect to FERTILIZER and SEED are higher for hybrid seed users than local seed users, but the output elasticity with respect to LABOR is higher for local seed users.

Figure 2 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.

The statistic R_z^2 suggests that about 10% of the sample variation in inefficiency can be explained by the set of exogenous factors (see bottom of Table 4). From Table 6, EDUHIGH, RNFINC and TTACRES all have positive partial effects on the mean and negative effects on the variance of efficiency. FEMHEAD, DISTBUS, and 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 has 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 mostly consistent with a priori reasoning and the previous literature. The effects of education, credit constraints, farm size and infrastructure on efficiency have been discussed extensively in the previous literature. The effect of female head could be due to the fact that females are subject to social discrimination in Kenya. There are generally two situations in which a female can become the head of a household. One is that she is a single mother, and the other is that her husband is dead. Females do not have the same inheritance rights as males in rural Kenya. A widow cannot obtain full rights to the land left by her husband and has to give away a certain proportion of the harvest to her husband's brothers. This may reduce the incentive to work intensively.

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 the 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 a little over half a percent for an average household, ceteris paribus. Being one kilometer closer to public transportation would increase yield per acre by 3.7%. An increase of one acre in farm size would raise yield per acre by less than one-third of a percent. If the proportion of household members who receive off-farm income increases by 10%, yield per acre would increase by 1.3%. However, using the same amount and the same quality of inputs, a household with a female head tends to produce 14% less maize than a household with a male head, and farmers tend to produce 17% more maize working in rented fields than in their own fields.

6 Conclusion

This paper makes four contributions to the stochastic frontier literature. First, we provide formulas to compute the standard errors for partial effects of exogenous firm characteristics on output levels and inefficiency for alternative model specifications. Second, we develop an R^2-type measure that summarizes the overall explanatory power of the exogenous factors that affect inefficiency. Third, we propose a bootstrapping procedure to evaluate the power of the recently developed model selection procedure suggested by AAOS to choose among competing models of the influence of firm characteristics on inefficiency. To our knowledge, bootstrapping has not been used previously to examine the size and power of model selection criteria. Fourth and finally, we apply our procedures and the AAOS model selection approach to an empirical

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

application of stochastic frontier analysis of maize production in Kenya.

The application is to Kenyan maize production and we find that different specifications provide similar efficiency rankings of households and predict the same directions for partial effects of exogenous factors. However, the magnitudes of these estimated partial effects are rather different across model specifications. This finding calls for more attention to model selection in empirical stochastic frontier analysis. The specification tests recently proposed by Alvarez et al. (2006) yield an unambiguous choice of best model using the Kenyan maize data. After evaluating the size and power of the test procedures with our bootstrap analysis we then use the preferred model to identify factors that limit technical efficiency in maize production in Kenya, and quantify their partial effects on maize yields. We examine the effects of education, female head of household, distance from a bus stop, land owned or rented, extent of off-farm income, and farm size on the level of efficiency. Approximately 10% of the variation in efficiency levels is accounted for by these household characteristics, and while education, non-farm income, and farm size increase technical efficiency, female-headed households, distance from a bus stop, and land being owned rather than rented all decrease it.

Acknowledgments The authors would like to thank Peter Schmidt for his suggestions and comments. Thanks also go to Thomas Jayne for generously providing us the data, Elliot Mghenyi, Kirimi Sindi, Tavneet Suri, and Margaret Beaver for comments and suggestions. All remaining errors or omissions are our own.

Appendix

Estimating partial effects of exogenous factors and their standard errors for the general model

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 and variance of efficiency via differentiation as

$$\partial E(-u_i|x_i, z_i)/\partial z_i^c = \gamma^c \sigma_i (R_1 R_3 - R_2) - \delta^c \sigma_i R_1 (1 + R_3) \tag{11}$$

$$\partial V(u_i|x_i, z_i)/\partial z_i^c = 2\gamma^c \sigma_i^2 (1 + R_3 + R_4) - \delta^c \sigma_i^2 R_4, \tag{12}$$

where $\mu_i = \mu \cdot \exp(z_i^{'} \delta)$, $\sigma_i = \sigma_u \cdot \exp(z_i^{'} \gamma)$, δ^c and γ^c are the coefficient vectors associated with z_i^c, R_1, R_2, and R_3 are as defined in the text, and $R_4 = R_1(R_2 + R_1 R_3 + 2R_2 R_3)$.

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

$$\sqrt{n}[g(\hat{\theta}) - g(\theta)] \to N\left[0, \left(\frac{\partial g(\theta)}{\partial \theta'}\right) \Omega \left(\frac{\partial g(\theta)}{\partial \theta'}\right)'\right], \tag{13}$$

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

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

$$\frac{\partial g(\theta)}{\partial \delta'} = -\sigma_i (\gamma^c z_i' + D) R_1 (1 + R_3) - \sigma_i (\delta^c - \gamma^c) z_i' R_5, \tag{15}$$

$$\begin{aligned}\frac{\partial g(\theta)}{\partial \gamma'} = {} & \sigma_i \gamma^c z_i' (-R_1 - R_2 - R_1 R_4) + \sigma_i D (R_1 R_3 - R_2) \\ & + \sigma_i \delta^c z_i' R_5,\end{aligned} \tag{16}$$

$$\frac{\partial h(\theta)}{\partial \delta'} = \sigma_i^2 \left[\gamma^c z_i' (R_6 - 2R_4) - \delta^c z_i' R_6 - R_4 D\right], \tag{17}$$

$$\begin{aligned}\frac{\partial h(\theta)}{\partial \gamma'} = {} & \sigma_i^2 \left[\gamma^c z_i' (4 + 4R_3 + 4R_4 - R_6) + \delta^c z_i' (R_6 - 2R_4)\right. \\ & \left. + 2(1 + R_3 + R_4) D\right],\end{aligned} \tag{18}$$

where $D = [I_{K1} \quad 0_{K1 \times K2}]$ is a $K_1 \times K$ dimensional matrix, and

$$R_5 = R_1(1 + R_3) - R_1 R_4, \tag{19}$$

$$R_6 = R_4 + R_1(2R_1 R_3 + 2R_1 R_3^2 - R_1 R_4 - 2R_2 R_4). \tag{20}$$

$\frac{\partial g(\theta)}{\partial \theta'} = \left[\frac{g(\theta)}{\delta'} \quad \frac{g(\theta)}{\gamma'}\right]$ and $\frac{\partial h(\theta)}{\partial \theta'} = \left[\frac{h(\theta)}{\delta'} \quad \frac{h(\theta)}{\gamma'}\right]$ are $K_1 \times 2K$ dimensional matrices, which depend on the model parameters δ and γ. We can get the estimates of $\frac{\partial g(\theta)}{\partial \theta'}$ and $\frac{\partial h(\theta)}{\partial \theta'}$ 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 formulas (13) and (14).

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

$$\begin{aligned} d(\theta) &= E(-u_i|x_i, z_i, z_{ik} = 1) - E(-u_i|x_i, z_i, z_{ik} = 0) \\ &= [-\sigma_i(R_1 + R_2)]|_{z_{ik}=1} - [-\sigma_i(R_1 + R_2)]|_{z_{ik}=0} \end{aligned} \tag{21}$$

$$\begin{aligned} r(\theta) &= V(u_i|x_i, z_i, z_{ik} = 1) - V(u_i|x_i, z_i, z_{ik} = 0) \\ &= [\sigma_i^2(1 + R_3)]|_{z_{ik}=1} - [\sigma_i^2(1 + R_3)]|_{z_{ik}=0} \end{aligned} \tag{22}$$

Similarly, following the delta method, we have

$$\sqrt{n}[d(\hat{\theta}) - d(\theta)] \to N\left[0, \left(\frac{\partial d(\theta)}{\partial \theta'}\right) \Omega \left(\frac{\partial d(\theta)}{\partial \theta'}\right)'\right] \tag{23}$$

$$\sqrt{n}[r(\hat{\theta}) - r(\theta)] \to N\left[0, \left(\frac{\partial r(\theta)}{\partial \theta'}\right) \Omega \left(\frac{\partial r(\theta)}{\partial \theta'}\right)'\right] \tag{24}$$

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

$$\begin{aligned}\partial d(\theta)/\partial\delta' = {} & [-\sigma_i R_1(R_1+R_3)z_i']|_{z_{ik}=1} \\ & - [-\sigma_i R_1(R_1+R_3)z_i']|_{z_{ik}=0}\end{aligned} \tag{25}$$

$$\begin{aligned}\partial d(\theta)/\partial\gamma' = {} & [-\sigma_i(R_2-R_1R_3)z_i']|_{z_{ik}=1} \\ & - [-\sigma_i(R_2-R_1R_3)z_i']|_{z_{ik}=0}\end{aligned} \tag{26}$$

$$\partial r(\theta)/\partial\delta' = [-\sigma_i^2 R_4 z_i']|_{z_{ik}=1} - [-\sigma_i^2 R_4 z_i']|_{z_{ik}=0} \tag{27}$$

$$\begin{aligned}\partial r(\theta)/\partial\gamma' = {} & [(2+2R_3+R_4)\sigma_i^2 z_i']|_{z_{ik}=1} \\ & - [(2+2R_3+R_4)\sigma_i^2 z_i']|_{z_{ik}=0}\end{aligned} \tag{28}$$

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

References

Ahmad M, Bravo-Ureta BE (1995) An econometric decomposition of dairy output growth. Am J Agric Econ 77:914–921

Ali M, Flinn JC (1989) Profit efficiency among basmati rice producers in Pakistan Punjab. Am J Agric Econ 71:303–310

Alvarez A, Arias C (2004) Technical efficiency and farm size: a conditional analysis. Agric Econ 30:241–250

Alvarez A, Amsler C, Orea L, Schmidt P (2006) Interpreting and testing the scaling property in models where inefficiency depends on firm characteristics. J Productiv Anal 25: 201–212

Aigner DJ, Lovell CAK, Schmidt P (1977) Formulation and estimation of stochastic frontier production functions. J Econometrics 6:21–37

Barrett C (1996) On price risk and the inverse farm size-productivity relationship. J Dev Econ 51:193–216

Battese GE, Coelli TJ (1995) Frontier production functions, technical efficiency and panel data: with applications to paddy farmers in India. J Productiv Anal 3:153–169

Benjamin D (1995) Can unobserved land quality explain the inverse productivity relationship? J Dev Econ 46:51–84

Caudill SB, Ford JM (1993) Biases in frontier estimation due to Heteroskedasticity. Econ Lett 41:17–20

Caudill SB, Ford JM, Gropper DM (1995) Frontier estimation and firm specific inefficiency measures in the presence of Heteroskedasticity. J Bus Econ Statist 13:105–111

Hadri K (1999) Estimation of a doubly heteroscedastic stochastic frontier cost function. J Bus Econ Statist 17:359–363

Hazarika G, Alwang J (2003) Access to credit, plot size and cost inefficiency among smallholder tobacco cultivators in Malawi. Agric Econ 29:99–109

Huang CJ, Liu JT (1994) Estimation of a non-neutral stochastic frontier production function. J Productiv Anal 5:171–180

Huang Y, Kalirajan K (1997) Potential of China's grain production: evidence from the household data. Agric Econ 70:474–475

Jacoby H (2000) Access to markets and the benefits of rural roads. Econ J 110:713–737

Karanja D, Jayne T, Strasberg P (1998) Maize productivity and impact of market liberalization in Kenya. Michigan State University International Development, Working Paper

Kumbhakar S, Biswas B, Bailey D (1989) A study of economic efficiency of Utah dairy farmers: a system approach. Rev Econ Statist 71:595–604

Kumbhakar S, Ghosh S, McGuckin J (1991) A generalized production frontier approach for estimating determinants of inefficiency in US dairy farms. J Bus Econ Statist 9:279–286

Lamb R (2003) Inverse productivity: land quality, labor markets, and measurement error. J Dev Econ 71:71–95

Meeusen W, van den Broeck J (1977) Efficiency estimation from Cobb-Douglas production functions with composed error. Int Econ Rev 18:435–44

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

Parikh A, Ali F, Shah MK (1995) Measurement of economic efficiency in Pakistani agriculture. Am J Agric Econ 77:675–685

Place F, Hazell P (1993) Productivity effects of indigenous land tenure systems in sub-Saharan Africa. Am J Agric Econ 75:10–19

Puig-Junoy J, Argiles J (2000) Measuring and explaining farm inefficiency in a panel data set of mixed farms. Pompeu Fabra University, Working Paper

Reifschneider D, Stevenson R (1991) Systematic departures from the frontier: a framework for the analysis of firm inefficiency. Int Econ Rev 32:715–723

Sen A (1962) An aspect of Indian agriculture. Economics Weekly (February):243–246

Sherlund S, Barrett C, Adesina A (2002) Smallholder technical efficiency controlling for environmental production conditions. J Dev Econ 69:85–101

Stevenson RE (1980) Likelihood functions for generalized stochastic frontier estimation. J Econometrics 13:57–66

Suri T (2005) Selection and comparative advantage in technology adoption. Yale University, Job Market Paper

Wang HJ (2002) Heteroscedasticity and non-monotonic efficiency effects of a stochastic frontier model. J Productiv Anal 18:241–253

Wang HJ (2003) A stochastic frontier analysis of financing constraints on investment: the case of financial liberalization in Taiwan. J Bus Econ Statist 21:406–419

Wang HJ, Schmidt P (2002) One-step and two-step estimation of the effects of exogenous variables on technical efficiency levels. J Productiv Anal 18:129–144