CHAPTER 3

PRODUCTION EFFICIENCY: CONCEPTS, APPROACHES TO MEASUREMENT, AND EMPIRICAL APPLICATIONS

3.1 Introduction

Efficiency is considered to be one of the most important issues in the production process. Efficiency is measured by comparing the actually attained or realized value of the objective function against what is attainable at the frontier. The resource constraint makes increasing efficiency one of the important goals of any individual and society since efficiency improvement is one of the important sources of growth. Thus, efficiency has policy implications both at the micro and macro economic levels.

The analysis of production and resource use in the farming sector has started to occupy an important place in agricultural policy frameworks that seek to increase domestic production by encouraging optimal resource utilization. Increasing technical and allocative efficiency is an important factor of productivity growth and is more appropriate in developing countries like Ethiopia where resources are scarce and raising production through improved efficiency does not generally require increasing the resource base or developing new technology. The importance of measuring and analyzing the level of efficiency of firms cannot thus be overemphasized. The analysis of technical and allocative efficiency under current technological change in agriculture will help policy makers to formulate adequate and appropriate extension services, pricing, marketing, credit, input distribution and land distribution policies.

3.2 Components of Production Efficiency

In microeconomic theory of the firm, production (or economic) efficiency is decomposed into technical and allocative efficiency. A producer is said to be technically efficient if production occurs on the boundary of the producer's production possibilities set, and technically inefficient if production occurs on the interior of the production possibilities set. That is, technical efficiency is the extent to which the maximum possible output is achieved from a given combination of inputs (Ellis, 1988). On the other hand, a producer is said to be allocatively efficient if production occurs in a region of the production possibilities set that satisfies the producer's behavioral objective.

Farrell (1957) distinguished between technical and allocative efficiency in production through the use of a frontier production function. Technical efficiency is the ability to produce a given level of output with a minimum quantity of inputs under certain technology. Allocative efficiency refers to the ability of using inputs in optimal proportions for given factor prices (i.e., where the ratio of marginal products for each pair of inputs is equal to the ratio of market prices). Economic efficiency is the product of technical and allocative efficiency. An economically efficient input-output combination would be on both the frontier function and the expansion path. Alternatively, economic efficiency can be defined as the ability of a production organization or any other entity to produce a given output at minimum cost. If a firm has achieved both technically efficient and allocatively efficient levels of production, it is economically efficient and new investment streams may be critical for any new development.

3.3 Production Technology and Sources of Output Growth

The specification of a production technology forms the basis for the conceptualization and measurement of efficiency. The determination of a benchmark for efficiency analysis depends on certain assumptions to be made about the behavior of the firm. The behavior of the production entity or the firm can be described either by the production function, cost function, profit function, or demand and supply functions. A rational decision maker will always attempt to maximize the gains or minimize the losses, which are defined by the respective maximization or minimization functions. There are different alternative economic theories of peasant household behavior, which assume that peasant households maximize one or more household objectives. In this study the behavior of the smallholders will be analyzed in terms of the production function approach.

The production technology that transforms inputs $X \in \mathrm{R}^n$ into net outputs $Y \in \mathrm{R}^m$ is modeled by the production function $Y = f(X)$, where $f(X)$ specifies the maximum output obtainable from the input vector X. This function represents the maximum attainable output for a given set of inputs. In other words, it represents a locus of efficient input-output combinations. The production technology is thus a mathematical relation on *f*, which transforms inputs into outputs and it gives the set of all technologically feasible input-output vectors. Production economics focuses not only on how resources are allocated but also on developing an efficient method of allocating resources to attain growth or economic development. Hence the analysis of economic development can also be approached through the theory of production economics.

Production in general may be increased in different ways. First, production may be increased through increased use of inputs, termed as horizontal expansion. In order for producers to use more inputs, either output prices must increase or the input prices must fall or both. This source of economic growth has little applicability in the present economic and social environment in Ethiopia, which is faced with resource limitations. Output can also be increased by improving efficiency usually referred to as the improvement approach. This approach requires the improvement of conditions or the removal of some existing institutional constraints to increase output using the existing technology. The other major source of growth is the transformation approach, which is characterized by a shift or an improvement in farm technology such as the use of technical packages, including improved seeds, fertilizers, credit, and chemicals that shift the production function outwards. Output per unit of input will be increased by changing the parameters of the production function. When new types of inputs of production are introduced into the production process, the production surface or the production horizon is changed.

3.4 The Efficiency Hypothesis

In the economic literature on efficiency, an important and often controversial subject is what is called the efficiency hypothesis. The notion that traditional farmers are 'poor but efficient' in their static environment has often been a view that drew the attention of several economists. The efficiency hypothesis, which was advanced by Schultz (1964) states that farm families in the developing countries are 'poor but efficient'. He explicitly stated that there are comparatively few significant inefficiencies in the allocation of the factors of production in traditional agriculture.

According to this hypothesis, since peasants are efficient within the constraints of existing technology, then only a change in the technology will bring about an increase in output. This hypothesis had influenced the perception of economists for a long time and its policy implications had remained to be of central importance in resource allocation. Accordingly, new investments and technological inputs from outside have been increasingly emphasized rather than extension and education efforts. Even in situations where the efficiency hypothesis did not apply, development policy makers have been overlooking opportunities for relatively inexpensive gains in production and concentrating only on expensive options such as investment in developing new technologies.

Conceptually the 'poor but efficient' hypothesis is related to a situation where external conditions are steady and not to situations which leave the farmer in a continuous disequilibrium. But farmers' environment is in a continuous motion, which necessitates an alteration in the technological, economic, and ecological conditions. The ever-growing degradation of tropical soils as well as the high man-land ratios under population pressure are best indicators for the disturbances of traditional farming systems. Different measures adopted by farmers to adjust to the rapidly changing environment create possibilities for substantial differences in efficiency. Farmers also find themselves in disequilibrium because of the continuously generated and diffused new technological innovations as well as by the continuous changes in input and output prices (Ali and Chaudhry, 1990). Accordingly, Schultz (1964) excluded very explicitly those who experienced a significant alteration of technological, economic or ecological conditions to which they had no time to fully adjust. Yet such alterations have meanwhile become typical for a great number of agricultural locations. With the rapidly changing environment, farmers attempt, more or less deliberately, to adjust their land use system, agronomic practices (such as soil preparation, planting date, plant density, fertilization, number of weeding, cultivation, and sowing date) and even the household economy and off-farm activities. Such adjustments require skills, awareness of risks and evaluation of current gains against fulfillment of future expectations and resource protection. This opens up possibilities for substantial inter-farm differences with respect to the path chosen and the technical efficiency of factor reallocation.

In addition to the conceptual arguments, the 'poor but efficient' hypothesis has also been a subject for a number of empirical investigations. For instance, after reviewing previous studies, Shapiro (1983) rejected this hypothesis and did his own empirical investigation of

Tanzanian cotton farmers and showed that output could be increased by 51 percent. This could be brought about if all farmers achieved those levels of technical efficiency that were in fact achieved by the best farmers in the sample using the inputs and technologies that the less efficient ones used. Studies conducted in the Philippines also showed that there were 25 to 50 percent inefficiencies in rice production (Lingrad, Castillo and Jayasuriya, 1983; Dawson and Lingrad, 1989). Ali and Flinn (1989) also found that the profit of rice farmers in Pakistan could be increased by 28 percent by improving their efficiency. A study in Punjab indicated that the income of farmers could be raised by 13 to 30 percent using the current technology (Ali and Chaudhry, 1990).

So the universal validity of this hypothesis is questionable in an environment that is no longer static and is characterized by substantial changes of technology, economy, and environment. It is also virtually impossible to meet the assumptions of facing the same production technology and prices for inputs and outputs and accept the profit-maximizing behavior of peasants. Rejecting the 'poor but efficient' hypothesis does not, however, necessarily imply that the theory does not have any contribution. At least it has been successful in placing peasant economics rationality on the agenda.

The Schultzian hypothesis was the point of departure for taking much more seriously the logic of peasant farm systems in order to discover the underlying logic of peasant farm practices instead of dismissing them as backward, lazy and irrational. The theory of profit maximization, which was the basis for the 'poor but efficient' hypothesis, is only one of the theories advanced to explain peasant household behavior. Several other alternative economic theories of peasant household behavior have been presented in the literature. The risk-averse peasant model (Ellis, 1988), the Chayanov model of utility maximization (Chayanov, 1966) and the new household models (Singh et al., 1986) are other major peasant household models frequently discussed in the literature.

The profit maximizing theory assumes that peasants are profit maximizing economic agents and are thus efficient producers. On the other hand, the risk-averse peasant theory argues that poor small farmers are necessarily risk-averse and they attempt to increase family security rather than maximize profit. The Chayanovian peasant model sets up a theory of the peasant household, which contains both consumption and production components and is based upon two basic assumptions: the absence of labor market and the flexible access to land. The new household economic models, which are similar to the Chayanovian model, relax some of the assumptions while at the same time maintain the integration between consumption and production. They drop the non-existence of the labor market and the unlimited supply of land assumptions.

No theory can be said to fully explain all aspects of peasant production systems and each may have relevance in explaining different aspects of the peasant economy. On the other hand, the theories are not distinct in all respects and none of the peasant household models make the study of technical efficiency inappropriate. Ellis (1988) pointed out that none of the theories assume or predict that peasant farmers are uniformly technically efficient in the sense that they all operate on the same 'best' production function. The simple conclusion to draw from this is that varying technical efficiency amongst peasant farms is always worth investigating irrespective of the microeconomic theory of the farm household.

3.5 Efficiency under New Technology

Because modern agricultural technology is recognized to be an important tool for increasing agricultural production, policy makers have paid attention mainly to the choice of technology, and to the adoption of such chosen technology by farmers (Kalirajan, 1991; Ali and Chaudhry, 1990). Following the neoclassical Hirschman's model of economic development, policy makers in developing countries have followed the method of providing various incentive measures to induce farmers to achieve a high rate of adoption of the chosen modern technology. Contrary to the expectation, the field-level performances of many new technologies have been shown not to be as suggested by the Hirschman's model of development. In this context, Schumpeterian theory of development provides an explanation. It stresses the fact that technological progress depends not only on the choice of technology but also on the appropriate application of any technology (Kalirajan, 1991).

With the introduction of a new input (e.g., a new variety), farmers may experience initial inefficiency as they learn about the new input. This inefficiency may include technical inefficiency as farmers acquire skills in applying the input and allocative errors as they adjust the level of use of the new input to their own specific circumstances (Ghatak and Ingersent, 1984; Ali and Byerlee, 1991; Xu and Jeffrey, 1998). This is especially true if the environmental variables have strong interaction with the new inputs. If the introduction of a new input is a one-time change to the system, farmers will eventually adjust to a reasonably efficient use of the input through learning by doing. In practice, agriculture in developing countries has undergone profound changes in both the technical and economic environments. Changes in the technical environment are often accompanied by changes in the economic environment. The development of better transportation and marketing infrastructure encourages crop specialization. At the same time, input-output price relationships are subject to sharp changes, especially with the policy reforms in many developing countries, which have gradually eliminated subsidies on critical inputs such as fertilizers. The combination of an evolving technical and economic environment means that the equilibrium required for economic efficiency is a constantly moving target (Ali and Byerlee, 1991).

The complexity of decision making in a dynamic environment is compounded by several other sources of complexity in a modernizing agriculture. These sources of complexity are caused by the following factors. (1) A wide array of purchased inputs which can potentially be applied. (2) Strong interaction between some purchased inputs and environmental variables (e.g., between fertilizer and soil type or rainfall). (3) Interaction between the purchased inputs and the time and method of application of the inputs leading to high variability in output. (4) Interaction between management of preceding and succeeding crops in a multiple cropping sequence (Ghatak and Ingersent, 1984; Ali and Byerlee, 1991; Ellis, 1988). In a dynamic agriculture where decision making is a complex process, it is hypothesized that in the short run the managerial skills and information available to farmers may be more important in causing inefficiencies than other institutional factors.

3.6 Causes of Economic Inefficiency

The early interest in economic efficiency centered on the question of whether small farmers of the Third World were economically rational and price responsive. This question is no longer seriously debated. Rather, economic efficiency should be viewed only as a standard by which to judge resource productivity against its potential. As such, interest now centers on *system inefficiencies* that cause resource productivity to fall below its potential. Technical inefficiency due to inappropriate timing and method of using an input is likely to reflect inadequate information and technical skills on the part of farmers. However, factors external to farmers such as untimely input supply may also be important in some cases (Ellis, 1988; Ghatak and Ingersent, 1984; Ali and Byerlee, 1991).

Allocative errors may also reflect inadequate information and skills, but other factors such as risk aversion, capital constraints, and institutional constraints (e.g., tenancy) influence allocative efficiency. Moreover, interdependence of production and consumption decisions in farm households and failures in input markets are also expected to play an important role especially in determining optimum use of resources. Many of these factors, such as input market failures, are exogenous to the farmer. Even the failure to use the most efficient technique of production due to inadequate information suggests that the cost to the individual farmer of acquiring better information is greater than the benefits because of failure in information markets (Ellis, 1988; Ali and Byerlee, 1991). Therefore, the presence of inefficiency in resource use at the farm level is not inconsistent with the rationality of small farmers.

3.7 Approaches to Efficiency Measurement

The measurement of production efficiency has been highly recognized as an important exercise in view of its relevance for policy makers in showing whether it is possible to increase output by simply increasing the efficiency of the firm without substantial additional resources. The methodologies for examining the production efficiency of farmers can generally be grouped into four different broad categories: the average factor productivity estimates; the linear programming approach; the production function approach; and the profit function methodology. The simplest measure of efficiency is the partial or average productivity index. This approach is an unsatisfactory measure since it ignores the presence of other factors, which affect average or marginal productivity and considers only one input at a time.

A simple comparison of total factor productivity is not a satisfactory efficiency indicator because farm households differ with respect to factor proportions, subsistence needs, and off-farm income opportunities, all of which have an impact on the revenue obtainable from a given resource endowment (de Haen and Runge-Metzger, 1989). The attempts to overcome this shortcoming led to the development of total factor productivity indexes in which a weighted average of inputs was compared with average output. The profit function approach is also seriously criticized because of its assumption of profit maximization as the given objective in the allocation process (Ellis, 1988). Farmers' objectives may not necessarily be that of profit maximization. Utility maximization or minimizing risk could be important factors influencing farmers' decision making.

The conventional production function approach is the most widely used measure in the analysis of production efficiency of farmers. The traditional approach is to estimate an average production function by a statistical technique such as least squares. Average production functions have received far more attention for the simple statistical reason that the mean of the error terms is zero. This is, however, not consistent with the definition of the production function.

Thus finding a measure of technical efficiency that is consistent with the definition of production function has been a major concern for many researchers. The production technology is represented by the transformation (production) function that defines the maximum attainable outputs from different combinations of inputs. Alternately, if considered from an input orientation side, it describes the minimum amount of inputs required to achieve a given output level. In other words, the production function describes a boundary or a frontier.

Given the definition of a production function, interest has then centered more on specifying and locating the production frontier. Alternative production models have often been proposed and the frontier model is one of these models and there seems to be a consensus in the recent literature on production function estimation that the production frontier rather than the average production function corresponds to the theoretical notions of the production function. Farrell (1957) had been the pioneer who introduced the frontier measure of efficiency, which reflects actual firm performances, and can include all relevant factors of production and is consistent with the textbook definition of the production function.

The frontier production function approach has some obvious advantages over the traditional methodologies and its use has therefore become widespread. The primary advantage of the method is that it is more closely related to the theoretical definition of a production function, which relates to the maximum output attainable from a given set of inputs. The second advantage of the method lies in the fact that estimates of technical efficiency of a firm in the sample may be obtained by comparing the observed output with the predicted (or attainable) output. Deviations from the frontier have acceptable interpretations as measures of the inefficiency of economic units. This approach provides a benchmark against which one can measure the relative efficiency of a firm. The production frontier is, however, unknown and it has to be empirically constructed from observed data in order to compare the position of a firm or a farm relative to the frontier. Several methods have been developed for the empirical measurement of frontier models. The different methods that are developed to estimate the frontier production function can be categorized based on certain major criteria (Assefa, 1995). First, based on the way the frontier is specified, the frontier may be specified as a parametric function or as a non-parametric function. Second, based on the way the frontier is estimated, the frontier may be estimated either through programming techniques or through the explicit use of statistical procedures. Third, based on the way the deviations from the frontier are interpreted, deviations may be interpreted simply as inefficiencies or they could be treated as mixtures of inefficiency and statistical noise.

3.7.1 Deterministic Frontiers

3.7.1.1 Non-parametric Programming

Farrell's (1957) original work formed the basis of the non-parametric programming method with subsequent extensions of his work by Charnes et al. (1978) and Färe et al. (1985) giving rise to what is often referred to as Data Envelopment Analysis (DEA). In this approach, technical efficiency is defined as the minimum input for any particular combination of outputs. Farrell's original approach of computing the efficiency frontier as a convex hull in the input coefficient space was generalized to multiple outputs. This was reformulated into calculating the individual input saving efficiency measures by solving a linear programming (LP) problem for each unit by Charnes et al. (1978) under the constant returns to scale assumption. Färe et al. (1985), Banker et al. (1984), and Byrnes et al. (1984) extended this approach to the case of variable returns to scale and developed corresponding efficiency measures.

DEA is a nonparametric approach to distance function estimation (Färe et al., 1994). The method involves the use of linear programming to construct a piecewise linear envelopment frontier over the data points such that all observed points lie on or below the production frontier. Let X be a $K \times N$ matrix of inputs, which is constructed by placing the input vectors, x_i, of all N firms side by side, and Y denotes the $M \times N$ output matrix which is formed in an analogous manner.

The output oriented variable returns to scale DEA frontier is defined by the solution to N linear programs of the form

$$\begin{aligned} &\min_{\theta,\lambda} \quad \theta \\ &\text{subject to} \quad -y_i/\theta + Y\lambda \geq 0 \\ &\qquad x_i + X\lambda \geq 0 \\ &\qquad N1'\lambda = 1 \\ &\qquad \lambda \geq 0, \end{aligned} \tag{3.1}$$

where $N1$ is an $N \times 1$ vector of 1s, λ is an $N \times 1$ vector of weights, and θ is the output distance measure. We note that $0 \leq \theta \leq 1$ and that $1/\theta$ is the proportional expansion in outputs that could be achieved the ith firm, with input quantities held constant.

In a similar manner, the input-orientated variable returns to scale DEA frontier is defined by the solution to N linear programs of the form

$$\begin{aligned} &\max_{\rho,\lambda} \rho \\ &\text{subject to} \quad -y_i + Y\lambda \geq 0 \\ &\qquad x_i/\rho - X\lambda \geq 0 \\ &\qquad N1'\lambda = 1 \\ &\qquad \lambda \geq 0, \end{aligned} \tag{3.2}$$

where ρ is the input distance measure. We note that $1 \leq \rho \leq \infty$ and that $1/\rho$ is the proportional reduction in inputs that could be achieved by the ith firm, with output quantities held constant.

The technical efficiency measure under constant returns to scale, also called the 'overall' technical efficiency measure, is obtained by solving N linear programs of the form

$$\begin{aligned} &\min_{\theta_i^{\mathrm{CRS}}} \theta_i^{\mathrm{CRS}} \\ &\text{subject to } -Y\lambda + y_i \le 0 \\ &\qquad\qquad \theta_i^{\mathrm{CRS}} x_i - X\lambda \ge 0 \\ &\qquad\qquad \lambda \ge 0 \end{aligned} \tag{3.3}$$

where θ_i^{CRS} is a technical efficiency measure of the ith firm under constant returns to scale and $0 \le \theta_i^{\mathrm{CRS}} \le 1$. The output and input oriented models will estimate exactly the same frontier surface and, therefore, by definition, identify the same set of firms as being efficient. The efficiency measures may, however, differ between the input and output orientations. Under the assumption of constant returns to scale, the estimated frontier and the efficiency measures remain unaffected by the choice of orientation (Coelli and Perelman, 1999).

Farrell (1957) used an input-oriented approach to illustrate the measurement of efficiency. He used a simple example involving firms which use two inputs, X_1 and X_2, to produce a single output Y, under the assumption of constant returns to scale. The constant returns to scale assumption allows representing the technology using a unit isoquant. Farrell discussed the extension of his method so as to accommodate more than two inputs. Knowledge of the unit isoquant of the fully efficient firm, represented by TT′ in Figure 3.1, permits the measurement of technical efficiency. If a given firm uses quantities of inputs, defined by the point K, to produce a unit of output, the technical inefficiency of that firm could be represented by the distance YK, which is the amount by which all inputs could be proportionally reduced without a reduction in output. This is usually expressed in percentage terms by the ratio YK/OK, which represents the percentage by which all inputs could be reduced.

The technical efficiency (TE) of a firm operating at K is measured by the ratio TEk = OY/OK, which is equal to one minus YK/OK. TEk will take a value between zero and one, and hence provides an indicator of the degree of technical inefficiency of the firm. A value of one indicates the firm is fully technically efficient. For example, the point Y is technically efficient because it lies on the efficient isoquant.

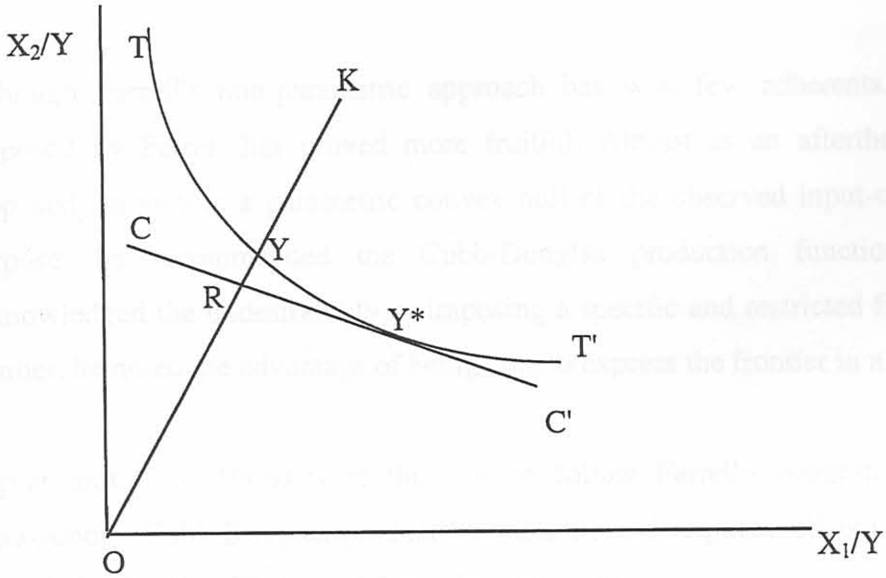

Figure 3.1: Farrell's Measure of Technical and Allocative Efficiencies

Farrell has also demonstrated that the unit isoquant provides a set of standards for measuring allocative efficiency. The isocost line CC′ gives the minimum cost of producing one unit of output given relative input prices. The allocative efficiency (AE) of the firm operating at K is defined to be the ratio, AEk = OR/OY, since the distance RY represents the reduction in production costs that would occur if production were to occur at the allocatively (and technically) efficient point Y*, instead of at the technically efficient but allocatively inefficient point Y. The total economic efficiency (EE) is defined to be the ratio, EEk = OR/OK, where the distance RK can also be interpreted in terms of a cost reduction. Thus, the product of technical and allocative efficiency provides the overall economic efficiency measure.

On the whole, the principal advantage of the non-parametric approach to technical efficiency measurement is that no functional form is imposed on the data. The principal disadvantage is that the frontier is computed from a supporting subset of observations from the sample and is therefore particularly susceptible to extreme observations and measurement errors. A second disadvantage of the approach is that the process of resource allocation to achieve better output is never explicitly used in the model. A third disadvantage of the approach is that estimated functions have no statistical properties, and hence the estimated production frontier has no statistical properties to be evaluated upon.

3.7.1.2 Parametric Programming

Although Farrell's non-parametric approach has won few adherents, a second approach proposed by Farrell has proved more fruitful. Almost as an afterthought, Farrell (1957) proposed computing a parametric convex hull of the observed input-output ratios. For this purpose, he recommended the Cobb-Douglas production function. Although Farrell acknowledged the undesirability of imposing a specific and restricted functional form on the frontier, he noted the advantage of being able to express the frontier in a mathematical form.

Aigner and Chu (1968) were the first to follow Farrell's suggestion. They specified a homogenous Cobb-Douglas production frontier, and required all observations to be on or beneath the frontier. Their model may be written as

$$\ln Y_i = \ln f(X_i;\beta) - u_i, \tag{3.4}$$

where Y_i is the output of the i^{th} firm, X_i is the input of the i^{th} firm and u_i is a one-sided disturbance term. The one-sided error term forces $Y \leq f(X)$. The elements of the parameter vector β may be estimated either by linear programming (i.e., minimizing the sum of the absolute values of the residuals, subject to the constraint that each residual be non-positive) or by quadratic programming (i.e., minimizing the sum of squared residuals, subject to the same constraint). The authors suggested that minimization of the sum of absolute deviations, $\sum_{i=1}^{n}\left|Y_i - f(X_i;\beta)\right|$, subject to $Y \leq f(X_i;\beta)$, which is a linear programming problem if $f(X_i;\beta)$ is linear in β. This is equivalent to minimization of the one-sided error term, u_i. Alternatively, they suggested minimization of the sum of squared deviations, $\sum_{i=1}^{n}\left[Y_i - f(X_i;\beta)\right]^2$, subject to the same constraint, which is a quadratic programming problem if $f(X_i;\beta)$ is linear in β. Although Aigner and Chu (1968) did not do so, the technical efficiency of each observation can be computed directly from the vector of residuals, since u_i represents technical inefficiency.

The principal advantage of the parametric deterministic approach vis-à-vis the non-parametric approach is the ability to characterize frontier technology in a simple mathematical form. However, the mathematical form may be too simple. The parametric approach imposes a structure on the frontier that may be unwarranted. The restrictive homogenous Cobb-Douglas specification has been relaxed by Forsund and Jansen (1977) and Forsund and Hajlmarsson (1979), among others. The parametric approach often imposes limitations on the number of observations that can be technically efficient. In the homogenous Cobb-Douglas case, for example, when the linear programming algorithm is used, there will, in general, be only as many technically efficient observations as there are parameters to be estimated. As was the case with the non-parametric frontier, the estimated frontier is supported by a subset of the data and is therefore extremely sensitive to outliers. One possibility suggested by Aigner and Chu (1968) and implemented by Timmer (1971) was essential just to discard a few observations. This has led to the development of the so-called probabilistic frontiers, which are estimated by the type of mathematical programming techniques discussed above, except that some specified proportion of the observations is allowed to lie above the frontier. The selection of this proportion is essentially arbitrary, lacking any explicit economic or statistical justification. If the rate of change of the estimates with respect to succeeding deletions of observations diminishes rapidly, this suggestion will be useful.

A final problem with this approach is that the estimates which it produces have no statistical properties. That is, mathematical programming procedures produce estimates without standard errors, t-ratios, and so forth. Basically this is because no assumptions are made about the regressors or the disturbance term in equation (3.4), and without some statistical assumptions inferential results cannot be obtained.

3.7.1.3 Statistical Frontier

The shortcomings of the programming approaches have led to the further development of the deterministic statistical frontiers. The statistical frontier models are similar to the deterministic programming frontier model. The deterministic statistical model involves statistical techniques and assumptions to be made about statistical properties.

The model of the previous section can be made amenable to statistical analysis by introducing some assumptions. Note that the model in equation (3.4) can be written as

$$Y = f(X)e^{-u} \tag{3.5}$$

or

$$\ln Y = \ln f(X) - u, \tag{3.6}$$

where $u \geq 0$ and thus $0 \leq e^{-u} \leq 1$, and where ln $f(X)$ is linear in the Cobb-Douglas case presented in equation (3.4). The question that must be asked is what to assume about X and u. The answer that has been given most often is to assume that the observations on u are independently and identically distributed, and that X is exogenous. Any number of distributions for u could be specified. Aigner and Chu (1968) did not explicitly assume such a model, though it seems clear that it was assumed implicitly. Afriat (1972) was the first to explicitly propose this model. He proposed a two-parameter beta distribution for $\exp(-u)$, and proposed that the model to be estimated by the maximum likelihood method. This amounts to a gamma distribution for u, as considered further by Richmond (1974). On the other hand, Schmidt (1976) has shown that if u is exponential, then Aigner and Chu's linear programming procedure is maximum likelihood, while their quadratic programming procedure is maximum likelihood if u is half-normal. It should be stressed that the choice of a distribution for u is important because the maximum likelihood estimates depend on it in a fundamental way - different assumed distributions lead to different estimates. This is a problem because there do not appear to be good a priori arguments for any particular distribution.

A further problem with maximum likelihood in the frontier setting is that the range of the dependent variable (output) depends on the parameters to be estimated, as pointed out by Schmidt (1976). This is because Y ≤ $f(X)$ and $f(X)$ involve the parameters to be estimated. This violates one of the regularity conditions invoked to prove the general theorem that maximum likelihood estimators are consistent and asymptotically efficient. As a result, the statistical properties of the maximum likelihood estimators needed to be reconsidered. This is done by Greene (1980) who showed that the usual desirable asymptotic properties of maximum likelihood estimators still hold if the density of u is zero at $u = 0$ and the derivative of the density of u with respect to its parameters approaches zero as u approaches zero. As noted by Greene (1980), the gamma density satisfies this criterion and is thus potentially useful here. However, it is a little troubling that one's assumption about the distribution of technical inefficiency should be governed by statistical convenience.

There is also an alternative method of estimation, first noted by Richmond (1974), based on the ordinary least squares results, which is called corrected OLS (COLS). Suppose equation (3.6) is linear (Cobb-Douglas). Then, in the first step, OLS is used to obtain consistent and unbiased estimates of the slope parameters and a consistent but biased estimate of the intercept parameter. In the second step the biased OLS intercept β_0 is shifted up ("corrected") to ensure that the estimated frontier bounds the data from above. The COLS intercept is estimated consistently by

$$\hat{\beta}_0^* = \hat{\beta}_0 + \max(\hat{u}_i), \tag{3.7}$$

where the $\hat{u}_i$ are the OLS residuals. The OLS residuals are corrected in the opposite direction, and so

$$-\hat{u}_i^* = \hat{u}_i - \max(\hat{u}_i). \tag{3.8}$$

The COLS residuals $\hat{u}_i^*$ are nonnegative, with at least one being zero, and can be used to provide consistent estimates of the technical efficiency of each producer by means of $\mathrm{TE}_i = Exp(-\hat{u}_i^*)$.

The COLS technique is easy to implement, and generates an estimated production frontier that lies on or above the data. However, this simplicity comes at a cost: The estimated production frontier is parallel to (in natural logarithms of the variables) to the OLS regression, since only the OLS intercept is corrected. This implies that the structure of "best practice" production technology is the same as the structure of the "central tendency" production technology (Kumbhakar and Lovell, 2000). This is an undesirably restrictive property of the COLS procedure, since the structure of best practice production technology ought to be permitted to differ from that of production technology down in the middle of the data, where producers are less efficient than best practice producers.

3.7.2 The Stochastic Frontier Production Function

The stochastic frontier production model represents an improvement over the traditional average production function and over the deterministic functions, which use mathematical programming to construct production frontiers. The notion of a deterministic frontier shared by all firms ignores the possibility that a firm's performance may be affected by factors entirely outside its control such as bad weather and input supply breakdowns as well as by factors under its control (i.e., technical inefficiency). To lump up the effects of exogenous shocks, both favorable and unfavorable, together with the effects of measurement errors and inefficiency into a single one-sided error term, and to label the mixture inefficiency is a problem with the deterministic frontiers.

According to Forsund et al. (1980) this conclusion is reinforced if one considers also the statistical noise that every empirical relationship contains. The standard interpretation is that, first, there may be measurement errors on the dependent variables. Second, the equation may not be completely specified, with the omitted variables individually unimportant. Both of these arguments hold just as well for production functions as for any other kind of equation, and it is dubious at best not to distinguish this noise from inefficiency, or to assume that noise is one-sided. These agreements lie behind the stochastic frontier (also called composed error) model developed independently by Aigner et al. (1977) and Meeusen and van den Broeck (1977). The essential idea behind the stochastic frontier model is that the error term is composed of two parts. A symmetric component permits random variation of the frontier across firms, and captures the effects of measurement error, other statistical noise, and random shocks outside the firm's control. A one-sided component captures the effects of inefficiency relative to the stochastic frontier.

The stochastic production function considered is defined as (Battese, 1992)

$$Y_i = f(X_i;\beta)\exp(\varepsilon_i), \tag{3.9}$$

where Y_i is total output of the i^{th} firm; $f(X_i;\beta)$ is a suitable function of the inputs vector X_i; β is a vector of unknown parameters; and ε_i is a random variable whose distributional properties are defined below.

The residual random variable, ε_i, in the production function of equation (3.9) is defined by

$$\varepsilon_i = v_i - u_i, \tag{3.10}$$

where v_i's are assumed to be independently and identically distributed as a normal random variable with mean zero and variance σ^2_v [i.e., $v_i \sim N(0, \sigma^2_v)$], and independent of the u_i's, which are assumed to be non-negative truncations of the normal distribution with mean, μ, and variance, σ_u^2 , [i.e., $u_i \sim N|(\mu, \sigma_u^2)|$], and μ, σ_u^2 and σ^2_v are unknown parameters to be estimated. The variance of ε is given by $\sigma^2 = \sigma_u^2 + \sigma^2_v$. The decomposition of the residual random variable, ε_i, in the production function (3.9), as specified in equation (3.10), is the decisive property which defines the stochastic frontier production function. The first term, v_i, is a random error which is assumed to be involved in the traditional linear regression allowing for the random variation of production across farms, and captures the effects of statistical noise, measurement errors, and the exogenous shocks beyond the control of the producing unit. The mean of this random error term is zero. The second term, u_i, is a non-negative firm effect variable, which is assumed to account for the existence of technical inefficiency of production of the i^{th} firm. The mean of the firm effect term is zero for a half-normally truncated distribution. If $u_i = 0$, production lies on the stochastic frontier and is technically efficient; if $u_i > 0$, production lies below the frontier and is inefficient. If the firm effect random term u_i is absent from the model, equation (3.9) becomes an average production function used in most econometric studies. Alternatively, if the random disturbance v_i is absent from equation (3.9), the model reduces to a deterministic frontier often estimated by linear programming techniques.

The economic logic behind equation (3.9) is that the production process is subject to two economically distinguishable random disturbances, with different characteristics. The non-negative firm effect u_i reflects the fact that each firm's output lies on or below its frontier. Any such deviation is the result of factors under the firm's control, such as technical and economic inefficiency, the will and effort of the producer and employees. The condition that $u_i \geq 0$ forces that all observations lie on or beneath the stochastic production frontier. The economic meaning of the one sided u_i component is that each firm's production must lie either on or below the production frontier. Any downward deviation from the frontier is due to technical inefficiency for the firm. If these inefficiencies could be eliminated, the firm would produce on the frontier. The error term u_i then represents technical inefficiency in the production process. But the frontier itself can vary randomly across firms, or over time for the same firm. According to this interpretation the frontier is stochastic, with random disturbance $-\infty \leq v_i \leq \infty$ being the result of favorable as well as unfavorable external events such as luck, climate, and topography.

The other important issue in stochastic frontier models is the assumption about u_i. Any number of one-sided distributions exist, which could plausibly be assumed to represent the distribution of the shortfall of output from the frontier. Aigner et al. (1977) considered half-normal and exponential distributions, while Meeusen and van den Broeck (1977) considered exponential ones. Other possibilities include gamma (Richmond, 1974) and lognormal (Greene, 1980). In most empirical research, however, the error term u_i is usually assumed to follow one of the following three distributions (Lee, 1983; Schmidt and Lin, 1984; Bauer, 1990): (1) half-normal $u_i \sim |N(0, \sigma_u^2)|$; (2) truncated normal at zero (μ, σ_u^2); (3) exponential EXP (μ, σ_u^2), where EXP indicates exponential distribution. Exponential is identical to the half-normal case, except that the technical inefficiency term u_i is assumed to follow the one-parameter exponential distribution. The result is similar to that for the half-normal (Jondrow et al., 1982). Greene (1990), however, also offered a two-parameter gamma distribution model.

Because of the ease of estimation and interpretation and the fact that technical efficiencies are in most cases similar for each distribution, there is a tendency by researchers to use the half-normal and truncated normal distributions. In addition, standard tests for distribution selection are not available. According to Lee (1983) since there are no a priori arguments for the choice of a particular distribution, one needs to base the choice and evaluation on statistical means. Lee (1983) proposed a Lagrange-Multiplier test to assess different distributions for the inefficiency term.

With the specifications (3.9) and (3.10), a measure of each firm's technical efficiency can be defined as

$$\mathrm{TE}_i = \frac{Y_i}{f(X_i;\beta)\exp(v_i)}, \tag{3.11}$$

where TE_i is technical efficiency for the i^{th} firm. As v_i is unobservable, (3.11) is not estimable. In other words, measurement of firm-specific technical efficiency requires first the estimation of the non-negative error u_i, that is the decomposition of ε_i into two individual components, u_i and v_i. Jondrow et al. (1982) suggested a technique for this decomposition using the conditional distribution of u_i given the total disturbance ε_i.

Following Jondrow et al. (1982) and adopting Battese and Corra's (1977) parameterization, the firm specific technical efficiency estimate can be derived from the conditional distribution of u_i given ε_i. The technical efficiency of the i^{th} firm is then given by

$$\mathrm{E}(u_i/\varepsilon_i) = \frac{\sigma_u \sigma_v}{\sigma}\left[\frac{f(.)}{1-F(.)} - \frac{\varepsilon_i}{\sigma}\left(\frac{\gamma}{1-\gamma}\right)^{1/2}\right] \tag{3.12}$$

where ε_i are estimated residuals for each farmer and $f(.)$ and F(.) are the values of the standard normal density function and standard normal distribution function, respectively, evaluated at $\frac{\varepsilon_i}{\sigma}\left(\frac{\gamma}{1-\gamma}\right)^{1/2}$. The maximum likelihood estimation of equation (3.9) yields estimators for β and γ where $\gamma = \frac{\sigma_u^2}{\sigma^2}$ and $\sigma^2 = \sigma_u^2 + \sigma_v^2$. γ explains the total variation of output from the frontier which can be attributed to technical inefficiency and lies between zero and one. The estimates of u_i and v_i can be obtained after replacing ε_i, σ and γ by their estimates. Hence, individual technical efficiency can be measured as $\mathrm{TE}_i = \exp(-\mathrm{E}(u_i/\varepsilon_i))$ which represents the level of technical efficiency of the i^{th} firm relative to the frontier firm.

More recent developments in frontier methodology include multi-equation models based on production, cost or profit function specifications. Coelli (1995) provides a review of these and other recent extensions of the stochastic frontier approach that take advantage of panel data structures. A major advantage of panel data models is that there is no longer need to assume that inefficiency is independent of the regressors. In addition, these models do not restrict the efficiency term to follow a specific distribution for the inefficiency term while making these restrictions testable propositions.

3.7.3 Stochastic Frontier Efficiency Decomposition

All the models discussed so far are only appropriate for measuring technical efficiency per se. The measurement of technical, allocative, and economic efficiency can only be handled, in a stochastic frontier framework, through the efficiency decomposition technique. The stochastic efficiency decomposition methodology was proposed by Bravo-Ureta and Rieger (1991), which was an extension of the model introduced by Kopp and Diewert (1982) to decompose cost efficiency into technical and allocative efficiency measures. Stochastic efficiency decomposition is generally based on the duality between production and cost functions.

Bravo-Ureta and Rieger (1991) utilize the level of output of each firm adjusted for statistical noise, observed input ratios, and the parameters of the stochastic frontier production function (SFPF) to decompose overall efficiency into technical and allocative efficiency. The parameters of the SFPF are actually used to derive the parameters of the dual cost function. Let the SFPF be redefined in its original form (e.g., Aigner et al., 1977) as

$$Y_i = f(X_i;\beta) + v_i - u_i. \tag{3.13}$$

If v_i is now subtracted from both sides of equation (3.13), we obtain

$$Y_i^* = f(X_i;\beta) - u_i = Y_i - v_i, \tag{3.14}$$

where Y_i^* is the i^{th} firm's observed output adjusted for the statistical noise captured by v_i, $f(\cdot)$ is the deterministic frontier output, and u and v are, respectively, the inefficiency and random components of overall deviations from the frontier. Adjusted output Y^* is used to derive the technically efficient input vector, X^t. The technically efficient input vector for the i^{th} firm, X_i^t, is derived by simultaneously solving equation (3.14) and the observed input ratios $\frac{x_1}{x_i} = k_i (i > 1)$ where k_i is equal to the observed ratio of the two inputs in the production of Y_i^*. The technically efficient input vectors form the basis for deriving the technical efficiency measures by taking ratios of the vector norms of the efficient and observed input quantities while the adjusted output is used to derive allocative and economic efficiencies employing the dual cost frontier function that is analytically derived from the SFPF.

Assuming that the production function in equation (3.9) is self-dual (e.g., Cobb-Douglas), the dual cost frontier can be derived algebraically and written in a general form as

$$C_i = h(W_i, Y_i^*; \delta), \tag{3.15}$$

where C_i is the minimum cost of the k^{th} firm associated with output Y_i^*, W_i is a vector of input prices for the i^{th} firm, and δ is a vector of parameters to be estimated. The economically efficient input vector for the i^{th} firm, X_i^e, is derived by applying Shephard's Lemma and substituting the firm's input prices and adjusted output level into the resulting system of input demand equations

$$\frac{\partial C_i}{\partial W_n} = X_n^e(W_i, Y_i^*; \theta), \tag{3.16}$$

where θ is a vector of parameters, $n = 1,2,.....,N$ inputs. The observed, technically efficient, and economically efficient costs of production of the i^{th} firm are equal to $W_i' X_i$, $W_i' X_i^t$, $W_i' X_i^e$, respectively. These cost measures are used to compute technical (TE) and economic efficiency (EE) indices for the i^{th} firm as

$$\mathrm{TE}_i = \frac{W_i' X_i^t}{W_i' X_i} \tag{3.17}$$

and

$$\mathrm{EE}_i = \frac{W_i' X_i^e}{W_i' X_i}. \tag{3.18}$$

Following Farrell (1957), the allocative efficiency (AE) index can be derived from equations (3.17) and (3.18) as

$$AE_i = \frac{W_i' X_i^e}{W_i' X_i^t}. \tag{3.19}$$

3.8 Empirical Applications of Production Frontiers

Frontier production function models have been applied in a considerable number of empirical studies both in agriculture and non-agricultural sectors since the pioneering work of Farrell (1957). Farrell's notion of an efficient unit isoquant provided a standard based on best results in practice, from which to gauge the efficiency of any sample observation. Recently, there have been increasing concerns about assessing the relative technical efficiency/inefficiency of firms, or the ability of firms to produce maximum output with a given set of inputs and technology.

A number of empirical works since Farrell's (1957) seminal paper used deterministic frontier production functions to analyze technical efficiency (e.g., Timmer, 1971; Russell and Young, 1983; Shapiro, 1983; Mijindadi and Norman, 1984; de Haen and Runge-Metzger, 1989; Ekayanake and Jayasuriya, 1987; Ali and Chaudhry, 1990; Saito, 1994; Getachew, 1995; Llewelyn and Williams, 1996). While the results of most of these studies indicated the existence of substantial inefficiencies of production, some of these studies actually found no evidence of significant inefficiencies among farmers and suggested that opportunities for inexpensive production gains through efficiency improvement are minimal. Analyses of technical efficiency of farmers in Tanzania (Shapiro, 1983), Punjab (Ali and Chaudhry, 1990), Ghana (de Haen and Runge-Metzger, 1989), Ethiopia (Getachew, 1995), Nigeria (Mijindadi and Norman, 1983), Kenya (Saito, 1994), England (Russell and Young, 1983), and Sri Lanka (Ekayanake and Jayasuriya, 1987) showed existence of considerable inefficiency ranging from 15 percent in Pakistan to 50 percent in Sri Lanka. On the other hand, studies of technical efficiency of US agriculture (Timmer, 1971) and East Java of Indonesia (Llewelyn and Williams, 1996) indicated very low levels of inefficiency, suggesting that the welfare losses are small and thus the expansion of agricultural output through improving production efficiency are limited.

Since the introduction of the stochastic frontier models (Aigner et al., 1977; Meeusen and van den Broeck, 1977), considerable applications of the stochastic frontier production, cost, and profit function models to both agricultural and non-agricultural production have been documented (e.g., Kalirajan and Shand, 1988; Parikh and Shah, 1994; Kumbhakar, 1994; Assefa, 1995; Parikh et al., 1995; Kumbhakar and Heshmati, 1995; Sharif and Dar, 1996; Ali, 1996; Getu et al., 1998; Seyoum et al., 1998). While only Parikh and Shah (1994) obtained insignificant technical inefficiencies, on average, of 4 percent among farmers in the North-West province of Pakistan, the rest of the studies obtained high levels of inefficiencies and hence proved the existence of considerable opportunities for output growth through efficiency improvement in other developing countries. Parikh and Shah (1994) identified lack of education, restricted credit and fragmented holdings as the causes of technical inefficiency and recommended policies which consolidate holdings, provide credit or educate farmers as these factors would tend to improve efficiency.

Kumbhakar (1994) estimated the technical efficiency of farms in West Bengal in India and obtained a mean technical inefficiency of 24 percent. While Kumbhakar and Heshmati (1995) analyzed the technical efficiency of Swedish dairy farms and found a mean technical inefficiency of 15 percent, Ali (1996) analyzed the technical efficiency of farmers in Nepal and obtained an average resource-use inefficiency of 25 percent. Ali (1996) also found poor land preparation, crop disease, off-farm work, and plot distance to have a negative impact on technical efficiency of wheat farmers. Assefa (1995) and Getu et al. (1998) used the Cobb-Douglas stochastic frontier model to analyze the technical efficiency of smallholder agriculture in central and eastern Ethiopia, respectively. While Getu et al. (1998) obtained higher average technical inefficiency of 32 percent, Assefa (1995) obtained average technical inefficiencies of 12 percent, confirming the regional variation in farmers' efficiency of production and also the higher technical efficiency among farmers in the relatively modern agricultural areas such as the central highlands. Assefa (1995) found that education, number of oxen, time of fertilizer delivery, farming experience, credit availability, distance from market center, farm size, and extension contact are important factors influencing technical efficiency.

Applications of the stochastic frontier models included that of comparison of the efficiencies of farmers across crop varieties. Sharif and Dar (1996), for example, studied the technical efficiency of farmers in the cultivation of traditional and high-yielding varieties of rice. They found that the overall yield variability in high-yielding varieties cultivation was due to technical inefficiency, while the opposite was true for traditional crops where random factors accounted for much of yield variability. Sharif and Dar (1996) also found that household characteristics such as educational level, growing experience, and farm size are associated with the technical efficiency differentials. Kalirajan and Shand (1988) analyzed the level and causes of technical efficiency of farmers in Southern India operating under rain-fed cultivation and obtained average inefficiencies of 35, 28, and 32 percent in the production of rice-1, corn, and rice-2, respectively. Kalirajan and Shand, in addition, examined the sources of technical efficiency differentials and found that while rice production efficiency is influenced by farming experience and extension officials' visits, corn production, on the other hand, appeared to be highly dependent on financial availability. On the basis of the analysis, Kalirajan and Shand concluded that a mere choice of high-yielding technology is not sufficient to increase the production of rice. What is important is the proper use or application of the technology.

The frontier models have also been used to evaluate the performance of different development programs. The effectiveness of a world bank-sponsored agricultural credit program as a tool for improving the agricultural productivity and income of the traditional farmers in Brazil was examined by assessing the technical and allocative efficiency of farmers who participated in the credit program vis-à-vis a comparable group of non participating farmers by Taylor et al. (1986). Since there was no significant difference between the technical and allocative efficiency of the two groups, the authors concluded that the agricultural credit program did not bring any significant impact on the efficiencies of the participating farmers. Seyoum et al. (1998) also investigated the technical efficiency of two samples of maize producers in eastern Ethiopia, one involving farmers within the SG project and the other involving farmers outside this program. The study used the Cobb-Douglas stochastic frontier production function in which the technical inefficiency effects are assumed to be functions of age, education, and extension. They found that farmers within the SG project are more technically efficient than farmers outside the project, relative to their respective technologies. Moreover, for farmers within the project, extension and education were found to have a positive and significant effect on technical efficiency while age has a negative influence. For farmers outside the project, extension and age were found to have a negative influence while education had no influence at all.

The stochastic efficiency decomposition technique has also been applied by a couple of authors to estimate the technical, allocative, and economic efficiency of farmers. For instance, Xu and Jeffrey (1998) obtained significantly lower technical, allocative, and economic efficiency indices for hybrid rice production in China as compared with conventional rice production across all the three regions studied. Singh et al. (2000) obtained lower technical, allocative, and economic efficiency for newly established Indian dairy processing plants after liberalization of the dairy industry compared to the old plants as they needed time to reach full operation, the right choice of products and other managerial skills required for higher performance.

Ali and Chaudhry (1990) estimated the mean technical, allocative, and economic efficiency measures for crop production in Pakistan at 84, 61, and 51 percent, respectively, while the corresponding measures for dairy farms in the USA were 83, 85, and 70 percent (Bravo-Ureta and Rieger, 1991). The average technical, allocative, and economic efficiency measures for crop-livestock farmers in Brazil were 17, 74, and 13 percent, respectively (Taylor et al., 1986), while the corresponding estimates for swine producers in Hawaii were 75.9, 80.3, and 60.3 percent, respectively (Sharma et al., 1999). Bravo-Ureta and Evenson (1994) obtained the three measures for cotton and cassava production. The average technical, allocative, and economic efficiency measures for cotton production were 58, 70, and 40 percent, respectively, while the corresponding figures for cassava were 59, 88, and 52 percent, respectively. Singh et al. (2000) also obtained average technical, allocative, and economic efficiency measures, respectively, of 86.7, 84.4, and 72 percent for Indian private dairy processing plants while the corresponding figures for the cooperative dairy processing plants were 87.4, 90.4, and 78.8 percent, showing that the new private dairy processing plants were less efficient than the old cooperative plants. All these studies indicated the existence of considerable potential within the farms to increase production through improved technical, allocative, and economic efficiency.

This study employs an extended stochastic frontier efficiency decomposition technique that accounts for scale effects that are a source of substantial bias in efficiency measures obtained from the Bravo-Ureta and Rieger method. The stochastic efficiency decomposition technique allows us to estimate the technical, allocative, and economic efficiencies of farmers using our cross-sectional data. The details of the analytical framework are given in Chapter 5 where it is applied to smallholder farmers in Eastern Ethiopia.