University of Connecticut

Department of Economics Working Paper Series

Choice of Inputs and Outputs for Production Analysis

by

Subhash C Ray
University of Connecticut

Working Paper 2021-05
March 2021

365 Fairfield Way, Unit 1063
Storrs, CT 06269-1063
Phone: (860) 486-3022
Fax: (860) 486-4463
http://www.econ.uconn.edu/

Choice of Inputs and Outputs for Production Analysis

Subhash C Ray

subhash.ray@uconn.edu

Abstract

Production creates value by transforming inputs into outputs. Classification of variables as inputs or outputs depends on the scope of decision making by the firm. Inputs enter the boundary of the firm from outside without any prior processing by the firm and once transformed into outputs exit its jurisdiction without any further processing. This chapter highlights the defining characteristics of inputs and outputs both for single stage and multi-stage production. The necessary conditions that must be met for a valid aggregation of several inputs into total expenditure are discussed both for nonparametric and parametric models of production. Several statistical tests of hypotheses related to aggregation of several inputs or exclusion of individual inputs in nonparametric models are discussed. The technology set of feasible input-output bundles invariably depends on many environmental or contextual variables that are outside the control of the producer. In parametric Stochastic Frontier Analysis they can be directly included as determinants of the mean or variance the technical efficiency factor causing shifts in the production frontier. In nonparametric Data Envelopment Analysis influence of such factors is measured through a second stage regression of efficiency scores on the contextual variables. The alternative approaches of a second stage least squares regression and a truncated regression are briefly discussed. The chapter ends with examples of input-output choice in several popular areas of application like manufacturing, banking, and health care.

Keywords: Multi-stage production; Input aggregation; Contextual variables; Second stage regression

JEL Classification Codes: D24, C44

The paper has benefitted from valuable comments by Subal Kumbhakar on an earlier draft.

Choice of Inputs and Outputs for Production Analysis

Subhash C Ray

Production is the act of transforming inputs into outputs creating value in the process. Enhancing human welfare is the ultimate goal of an economic system and welfare of an individual in the society is increased through consumption of goods and services. Thus, production of these goods and services is the means of improving welfare. As noted by Marx and Engels, men begin to "distinguish themselves from animals as soon as they begin to produce their means of subsistence, a step which is conditioned by their physical organization. By producing their means of subsistence men are indirectly producing their actual material life".[1] Over the different phases of civilization, human societies morphed from hunter-gatherer tribes, into farming communities, and ultimately into the modern industrial economies supported by artificial intelligence, machine learning and information technology. Despite all the technological progress, however, the essence of production remains conversion of inputs into outputs. In order to measure rates of technical progress or to evaluate productive performance of a firm, one must appropriately identify and measure inputs and outputs. The broad objective of this chapter is to highlight a number of important points that must be considered while classifying resources as inputs or outputs in any specific context.

The rest of the chapter unfolds as follows. Section 2 starts with some defining characteristics of inputs and outputs. This is followed by a discussion of multi-stage production, the distinction between fixed and variable inputs, the role of fixed inputs in short run cost minimization, and dealing with bad outputs or zero and negative outputs and inputs.

Section 3 considers the validity of input aggregation and use of total cost as a single aggregate input in both nonparametric Data Envelopment Analysis (DEA) and parametric Stochastic Frontier Analysis (SFA) models, statistical tests for aggregation and/or inclusion of inputs, and bootstrapping the estimated DEA efficiency scores. Various issues related to use of a 2nd stage regression to measure the marginal effects of non-discretionary contextual variables on efficiency measurement are discussed in section 4. Section 5 wraps up the chapter with examples

[1] Marx and Engels *The German Ideology*, 'Feuerbach' 1

of input-output selection for three frequently analyzed industries: manufacturing, banking, and health care. Section 6 is the conclusion.

2. Inputs and Outputs: Some Basic Features

In production economics, the firm is the decision-making agent with a clearly defined boundary or jurisdiction of its decisions. It acquires its inputs from outside and, once production is complete, ships out the finished output. In most (if not all) cases, such inputs have been subject to prior processing by some other firm before it enters into its jurisdiction. For example, in agriculture, the farmer uses fertilizers and pesticides as inputs. These fertilizers and other chemicals had to be produced by other firms before they were available as inputs to the farmer. However, the farmer did not participate in the production of these chemicals. At the other end, wheat produced by the farmer is ground into flour by a flour mill. But once the bag of wheat has been shipped out by the farm, it plays no role in any further processing of its product. Thus, an input enters the jurisdiction of the firm from outside without any prior processing whereas its output, once it exits the boundary of the firm, is not subject to any further processing by the firm.

In some cases, an input used by the firm may not be supplied from outside. A steel plant may acquire coal for its furnace from its own mines. But coal provided by the company's captive mine as fuel for the furnace is no different from coal purchased from some other supplier in the market and has not been excavated by the steel plant.

In classifying resources as inputs and outputs, the critical points to note are:

- An input exists even before production starts. The truckload of coal was already there even before the steel plant started production.
- An input is depleted in stock as production is carried out. The inventory of coal in the warehouse of the steel plant is reduced as more and more coal is used for steel making.
- An input has not been subject to any prior processing by the firm. The truckload of coal has not been excavated from the mine by the steelmaking plant.
- An output did not exist in its present form prior to production. Although the iron ore, coal, and other materials used by the steelmaker were in existence, there was no steel before production.
- An output increases in stock as production is carried out.

- Once the output is shipped out, it is not subject to any further processing by the firm.

In some cases, the output may not be physically differentiable from the inputs. Making a bowl of salad, for example, requires slicing tomatoes, cucumbers, and lettuce, chopping onions, and adding salad dressing. The bowl of salad is the output while the various ingredients and labor are the inputs. In this case the output does not look much different from the inputs and most of the ingredients (at least the vegetables) can be identified and physically separated. But the prepared bowl of salad is more valuable than the raw ingredients. The act of slicing, chopping, and mixing the ingredients is the production process that adds value to the ingredients.

An extreme example can be found from the postal service industry where the accumulated mail (consisting of letters, magazines, and packages) at the post office are inputs and the delivered mail at the different addresses are the outputs. In this case, there is no *physical transformation* of the inputs at all. However, the letter carrier's labor and the transportation service of the mailman's vehicle is used to deliver the mail from the post office to the recipient's addresses thereby creating value. Even though not physically altered in shape or size, the accumulated mail at the post office is turned into delivered mail at the recipient's address and becomes more valuable to the addressee than the undelivered mail.

In any empirical application, the relevant input and output variables need to be defined with care. There should be some obvious technical relationship between inputs and outputs. It is important to remember that definition of inputs and outputs will depend on the boundary of the firm.

2.1 Multi-stage Production

The process of producing the output typically involves various tasks - some to be performed simultaneously and others sequentially. Often, a number of tasks can be grouped together to define a specific stage of production. In many cases, one can identify multiple stages of production within the jurisdiction of the same firm. In such cases, the output from the earlier stage becomes an input for the next stage. For example, in furniture making lumber, labor, and tools are the inputs and the fabricated parts are the outputs in the fabrication shop. These parts along with additional labor and tools are the inputs in the assembly shop while the assembled but unpainted furniture are the output at this stage. These unfinished furniture, more labor, and paints are the inputs and the finished furniture are the outputs at the final stage. One can think of each division within the firm as a decision-making unit with the inputs and outputs appropriately defined to measure efficiency at each stage. Alternatively, the entire firm can be viewed as the decision-making unit and only the lumber, labor, tools used at the different stages, and the paints are treated as inputs while the finished furniture are the outputs. A strand in the DEA literature described as *network DEA* (but more appropriately a *multi-stage DEA*) seeks to ascribe the overall (in)efficiency of a firm to the different stages of production in a multi-stage production setting.

2.2 Classification of inputs: Variable and Fixed

In production economics a distinction is made between variable and fixed inputs. Variable inputs are those the quantities of which change with the quantity of output produced. These include raw materials, unskilled labor, and energy. By contrast, one does not see a change in plant size or the number of managers with every change in the output level. These are called fixed inputs. It is important to recognized that the fixed versus variable input designation is not a physical but an economic concept. When the output level increases, the firm can schedule production workers for additional hours and use more fuel and raw materials without any significant adjustment cost. Frequent changes in plant size (even when physically possible, by leasing extra facilities at short notice) would entail enormous amount of adjustment cost. Similarly, hiring and firing management personnel would result in large adjustment costs in the form of severance pay or hiring bonus. In such cases, the firm is better off by not altering the levels of these inputs. That is the reason why they are considered to be fixed. But this is only in the short run. In the long run, when the time comes to renew the lease on the premises or to renew the contract of management staff, the firm can change the levels of these inputs that would be appropriate for its planned level of output.

2.3 Fixed Inputs and Short Run Cost Function

For cost minimization in the short run, the firm will have to find the least expensive bundle of the variable inputs that can produce the targeted output when combined with the preset quantities of the fixed inputs. Note that because both the prices and the quantities of the fixed inputs are predetermined the level of fixed costs is already set. Cost minimization in the short run, therefore, requires minimization of the variable costs only.

Consider the production possibility set

$$T = \left\{(x, y) : x \in R_{+}^{n} \text{ can produce } y \in R_{+}^{m}\right\} \qquad (1)$$

where x is a vector of inputs and y is a vector of outputs.

Suppose that the firm's input bundle is partitioned as $x = (v; K)$ where v is the vector of variable inputs and K is the vector of fixed inputs. A corresponding partition of the input price vector is $q = (w, r)$. The firm has a target output vector y^0 and has the given bundle of fixed inputs K^0. The minimum variable cost would be

$$VC(w; K^0, y^0) = \min w'v : \left(v, K^0, y^0\right) \in T. \qquad (2)$$

In parametric econometric analysis, one can specify a suitable functional form of the variable cost and estimate the parameters of the model empirically. The regularity conditions that must be satisfied by any estimated variable cot functions are

(i) $(\frac{\partial VC}{\partial y}) > 0;$ (positive marginal costs)

(ii) $(\frac{\partial VC}{\partial w}) > 0;$ (positive conditional input demands implied by Shephard's lemma).

(iii) $(\frac{\partial VC}{\partial y})' y = VC;$ (linear homogeneity in the variable input prices) and

(iv) the matrix $(\frac{\partial^2 VC}{\partial w . \partial w'})$ is negative definite (concavity in the input prices).

Additionally, (v) $(\frac{\partial VC}{\partial K}) < 0.$ This last condition follows from the fact that because the marginal productivity of each fixed input is positive, any increase in a fixed input must be offset by a decrease in some variable input if the output is to remain constant. This causes the variable cost to decline. In fact, one can define the shadow price of the sth fixed input as

$$\rho_s = -\frac{\partial VC}{\partial K_s}. \qquad (3)$$

In the single fixed input case, one can perform a statistical test for the hypothesis that the shadow price (ρ) of the fixed input is equal to its market price (r). When multiple fixed inputs are involved one can apply the Mahalanobis D^2 test to compare the mean vectors of shadow prices and the market prices of the fixed inputs.

The corresponding DEA LP problem is

$$
\begin{aligned}
& VC_0^* = \min w'v \\
& s.t. \sum_{j=1}^{N} \lambda_j v^j \le v; \\
& \quad \sum_{j=1}^{N} \lambda_j K_j \le K^0; \\
& \quad \sum_{j=1}^{N} \lambda_j y^j \ge y^0; \\
& \quad \sum_{j=1}^{N} \lambda_j = 1; \\
& v \ge 0; \lambda_j \ge 0, (j = 1, 2, ..., N).
\end{aligned} \tag{4}
$$

The short run cost of the firm is $C(w, r, y^0, K^0) = VC(w, K^0, y^0) + r'K^0$. The shadow price of a fixed input is the negative of the optimal value of the dual variable associated with that particular fixed input. The marginal cost of any output is the dual variable associated with that output constraint.

It is important to clearly distinguish between fixed inputs and outputs. Output is clearly a 'cost driver' in the sense that an increase in the sense that an increase in the output quantity will result in an increase in the cost. For the fixed inputs, there are two opposing forces at work. Clearly, with prices of the fixed inputs given, any increase in the quantity of fixed inputs will increase the fixed cost. But, to the extent that the fixed inputs have substitutability with any of the variable inputs, an increase in a fixed input *holding the outputs unchanged* will lower the quantity of some variable input and hence the cost of variable inputs. Hence the short run *total cost* may rise or fall.

Sometimes, fixed inputs are confused with outputs because there are variable inputs required for the maintenance of these fixed inputs. One requires labor and other resources for maintenance of a furnace in a steel plant. In this case, an increase in the level of a fixed input is actually increasing the quantity and the cost associated with variable inputs. None the less, the furnace remains an input and not an output because it is not produced in the steel plant. For accurate costing, the maintenance cost of the fixed input (like the furnace) should be included in the user cost of the fixed input and would constitute a part of the fixed cost of the firm.

2.4 Bad Outputs: Zero or Negative Inputs and Outputs

Bad output

In some cases, production results in some unintended and undesirable *bad* outputs side by side with the desired or *good* outputs. Electric power generation also results in smoke emission from thermal plants. Cement production leads to toxic waste on the side. In production economics bad outputs have been treated in various ways – as inputs, joint products with the good outputs, and as by products of good output production. Of these, the input interpretation is the least satisfactory. There was no smoke before power generation started. The amount of smoke in the air is increased rather than reduced as more power is generated. Further, the smoke in the air is not used for producing anything by the power plant. Thus, it possesses none of three defining characteristics of an input listed above. As for the joint production vis-à-vis by product interpretation, the context should determine which one is more appropriate. In some cases, products are physically joint – like cowhide and beef or some flowers and allergenic pollens. There joint production is the obvious interpretation. Similarly, the same polluting input coal produces heat (for electricity) and carbon. The two outputs – power and carbon emission - are not physically joint but are ascribed to the same input. There by-production is more appropriate. In other cases, the conceptualized production technology depends on the analyst's perspective.[2] It is important to note in this context that good or bad characterization is related to an individual's preferences and is not a characteristic of the technology. For example, while smoke is generally considered a bad output, in rural India villager's often burn wet hay explicitly to create smoke to drive away mosquitoes. In that specific case, smoke is the intended *good* output.

Zero Input or Output

[2] For a detailed discussion of bad outputs, see the chapter by Murty and Russell in volume 1 of this handbook.

In some cases, one finds that the firm is not using some input at all. For example, in agriculture in less developed countries, farmers use both chemical fertilizers and farmyard manure. But often farms are found to use no manure at all. The same is true for the use of bullocks or motor-powered tillers. This implies that the one input can be substituted for the other. Soil nutrients can be provided either by manure or by chemical fertilizers. In such cases, the two inputs should be aggregated as 'Fertilizers and Manure'. The exact basis for aggregation may be their chemical composition or some other formula determined by experts. Similarly, different kinds of fuel (coal, oil, and natural gas) can be aggregated based upon their calorific values.

In some other cases, substitutability between pairs of inputs is less obvious. One can think of chemical weed killers and labor. One may spray chemicals to kill weeds or use labor to root the weeds out. In this situation, aggregation is neither possible nor recommended and one must treat any zero value of the chemical herbicide input as it is.

As for outputs, in multi-product firms, some outputs may not be produced at all. This is a perfectly rational decision of a profit or revenue maximizing firm. In fact, even when all firms produce all of the outputs, one would like to compare the combined cost of standalone production of individual outputs with the production cost of a diversified firm producing all of the outputs together to measure economies of scope. This is particularly a problem when outputs are measured in log (as in a Translog cost function). Often the recommended solution is to use a Box-Cox transformation instead of the logarithm of the output quantity to circumvent the problem of defining the log of zero.

In the nonparametric DEA models, a zero output poses no problem in a radial output-oriented model. Similarly, a zero input is easily accommodated in an input-oriented model. But with a zero input, in an output-oriented model the benchmark can be constructed only from units that employ zero quantity of the relevant input.

<u>Negative Inputs or Outputs</u>

In many papers (especially in the nonparametric DEA literature) authors have developed various mathematical models for evaluating production efficiency in the presence of negative inputs or outputs. One such model simply reverses the algebraic sign and treating negative inputs as outputs or outputs as inputs (when all observations of the relevant input or output are negative).

When some observations have negative and others have positive values of an input or output, one adds a sufficiently large positive number to all observations so that the modified values are all positive. One appeals to the translation invariable property of the DEA models to justify this data transformation.[3] But the fundamental point one needs to remember is that the neoclassical production possibility set is a subset of R_{+}^{m+n} and for any feasible input-output bundle (x, y), $x \in R_{+}^{n}$ and $y \in R_{+}^{m}$. Thus, negative inputs or outputs are precluded. One often regards the profit of a firm as an output and points out that it can be negative. But a firm does not produce profit. It uses *inputs* to produce *output*. Profit is neither an input nor an output but the difference between the revenue from the output and the cost of the inputs.

3. Input Aggregation:

Selecting the right number of distinct inputs and outputs for the specification of the underlying technology is often quite difficult in applied production analysis. In many cases, for a variety of reasons, several inputs (or outputs) may be aggregated into a single composite input (or output). In agricultural production, for example, the number of even the important crops may be too large to treat them as separate outputs. This requires aggregating them into a single index of crop output. Similarly, especially in developing countries, one constructs a measure of total cultivated area by adding up irrigated and unirrigated acreage, assigning a higher weight to irrigated land. Because there are farms that cultivate only unirrigated land, including irrigated land as a distinct input would create the usual problems associated with zero input levels. In manufacturing, quantities of different kinds of fossil fuels are sometimes aggregated in terms of their British Thermal Units (BTU) equivalent. In many cases, information about input use is available only in value terms. One may, for example, have to use the total wage payments instead of the numbers of different categories of employees to measure the labor input. This is especially true for applications where the input-output data are indirectly constructed from financial statements of firms.

In applications of Data Envelopment Analysis (DEA), maintaining a large number of inputs (or outputs) in the model may result in all or most of the decision-making units (DMUs) to

[3] As explained in Ray (2004), the output oriented radial DEA model is invariant to input translation but not to output translation. Similarly, in the input-oriented model is invariant to output translation but not to input translation.

be rated as efficient. As explained by Leibenstein and Maital (1992) this is a result of the dimensionality of the input /output space relative to the number of observations. Meaningful aggregation of multiple inputs or outputs, therefore, reduces the number of constraints in the linear programming model, thereby increasing the "degrees of freedom". However, while input aggregation has some advantages when it is valid, it does pose problems in cases when it is not valid.

The theoretical implications of such aggregation have been addressed in some previous studies. For instance, based on Monte Carlo simulations, Thomas and Tauer (1994) argue that with linear aggregation of inputs the technical efficiency measure becomes an economic measure of efficiency comprising of both technical and allocative components. As such, this introduces bias in the measurement of technical efficiency. In a subsequent study by Tauer (2001), the results of simulations indicate that technical efficiency estimates computed by DEA are biased even when the exact aggregator function is used to aggregate inputs. Färe and Zelenyuk (2002) and Färe et al (2004) demonstrate that if some inputs are introduced in the DEA model in price aggregated form, the technical efficiency measure will be *biased downward* as compared to if the inputs were included in disaggregated terms. Further, the bias is equal to the allocative inefficiency. They further illustrate that the DEA technique will yield unbiased technical efficiency scores even when inputs are price aggregated, if and only if there is no allocative inefficiency in the sub-vector of inputs that is aggregated.

The above theoretical developments reveal that while including too many distinct inputs (or outputs) in the DEA model may lead to problems associated with overspecification, aggregating inputs into a fewer number of inputs to be included in the model may lead to biased measures of technical efficiency. The validity of input aggregation in any specific application is therefore an empirical question. However, the statistical side of this issue has been far less explored. Of the few studies that address this aspect of the problem, some are nonparametric while others assume some parametric form of the statistical distribution. Simar and Wilson (2001) discuss bootstrap procedures for testing several restrictions in nonparametric models, including tests for input or output aggregation. Sirvent et al (2005) use simulations to evaluate the performance of a number of statistical tests that can be used for the selection of variables in DEA efficiency models. In a recent paper Banker et al. (2007) propose a method of estimating allocative inefficiency when only aggregate cost or revenue data and quantity data (but not price data for individual inputs or outputs) are available.[4] They provide a test of *allocative efficiency* through an application of Banker's (1993) F-test based on the DEA *technical efficiency* residuals from the aggregate and the disaggregated models.

In an empirical application, even when price and quantity data are all available, one might prefer to aggregate subgroups of inputs (or outputs) to overcome the 'curse of dimensionality'. There are, of course, various ways in which inputs may be aggregated. In some contexts, the aggregation weights may be guided by *technical norms* - like aggregating various kinds of fuels into a total energy input based on their individual Btu contents. In some other cases the aggregation may be *data driven* – like using Principal Component Analysis where several inputs (or outputs) are replaced by a small number of their principal components, to reduce the problem of dimensionality.[5] An *economically meaningful* aggregation procedure is one where a sub-group of inputs (or outputs) is replaced by its total cost (or revenue). Using prices would be the valid way to aggregate when choice of the relevant inputs (or outputs) is allocatively efficient.

3.1 Input Aggregation in Nonparametric Models[6]

Consider an industry producing a scalar output, y, from a bundle of n inputs, $x=(x_1,x_2,\ldots,x_n)$. Let (x^j, y_j) be the observed input-output bundle of firm j $(j= 1,2,\ldots, N)$. The technology is defined by the production possibility set

$$T=\{(x,y): y \le f(x); y \in R_+, x \in R_+^n\} \qquad (5)$$

where $f(x)=\max y:(x,y)\in T$ is the production function.[7]

The production function evaluated at the input bundle x^0 is

[4] The idea is that the prices are not known to the analyst. But if all firms face identical prices and are allocatively efficient then the aggregated and disaggregated technical efficiencies are identical. It is a test for this equality.

[5] For a discussion of the use of Principal Component Analysis (PCA) within DEA see Adler and Golany (2001, 2002). An inherent problem with PCA is that typically there is no intuitive or economic interpretation of what these principal components represent.

[6] This section draws heavily from Ray and Mukherjee (2005)

[7] Firms move from one point to another on the same isoquant when (relative) input prices change – either over time or across regions. But under competitive input market assumption (justifying price-taking behavior) firms in the same market will face same prices.

$$f(x^0) = \max \varphi y^0 : (x^0, \varphi y_0) \in T. \qquad (6)$$

Obviously, every observed input-output bundle (x^j, y_j) is feasible. Under the standard assumptions of convexity of the production possibility set and free disposability of inputs and output, an estimate of the set T is:

$$S = \left\{ (x, y) : x \geq \sum_{j=1}^{N} \lambda_j x^j ; y \leq \sum_{j=1}^{N} \lambda_j y_j ; \sum_{j=1}^{N} \lambda_j = 1; \lambda_j \geq 0; j = 1,2,...,, N \right\} \qquad (7).$$

A nonparametric estimate of the production frontier at the input bundle x^0 is obtained by solving the following linear programming problem:

$$\begin{aligned} &\hat{f}(x^0) = \max \phi y_0 \\ &s.t. \sum_{j=1}^{N} \lambda_j x^j < x^0; \\ &\quad \sum_{j=1}^{N} \lambda_j y_j < \varphi y_0; \\ &\quad \sum_{j=1}^{N} \lambda_j = 1; \lambda_j \geq 0 \, (j = 1, 2, ..., N) \end{aligned} \qquad (8)$$

The support of $\hat{f}(.)$ is the free disposal convex hull of the observed input bundles

$$X^* = \left\{ x : x \geq = \sum_{j=1}^{N} \lambda_j x^j ; \sum_{j=1}^{N} \lambda_j = 1; \lambda_j \geq 0; j = 1,2,..., N \right\}. \qquad (9)$$

The dual of the LP problem (8) is

$$\begin{aligned} &\min \alpha_0 + \beta^0{}' x^0 \\ &s.t. \alpha_0 + \beta^0{}' x^j \geq p y_j ; (j = 1, 2, ..., N) \\ &\quad p y^0 = y^0 ; \\ &p \geq 0; \beta^0 \geq 0; \alpha_0 \text{ unrestricted.} \end{aligned} \qquad (10)$$

For a simple example, consider the 3-input 1-output case and the observed input-output bundle of firm k in the sample. For this example, the explicit form of problem (8) above is

$$
\begin{aligned}
&\hat{f}(x_{1k}, x_{2k}, x_{3k}) = \max \varphi y_k \\
&s.t. \sum_{j=1}^{N} \lambda_j y_j \geq \varphi y_k; \\
&\sum_{j=1}^{N} \lambda_j x_{1j} \leq x_{1k}; \\
&\sum_{j=1}^{N} \lambda_j x_{2j} \leq x_{2k}; \\
&\sum_{j=1}^{N} \lambda_j x_{3j} \leq x_{3k}; \\
&\sum_{j=1}^{N} \lambda_j = 1; \lambda_j \geq 0 \, (j = 1, 2, ..., N)
\end{aligned}
\tag{11}
$$

The corresponding dual problem is

$$
\begin{aligned}
&\hat{f}(x_{1k}, x_{2k}, x_{3k}) = \min \alpha_k + \beta_1^k x_{1k} + \beta_2^k x_{2k} + \beta_3^k x_{3k} \\
&s.t. \alpha_k + \beta_1^k x_{1j} + \beta_2^k x_{2j} + \beta_3^k x_{3j} \geq p y_j; (j = 1, 2, .., N) \\
&p y_k = y_k; \\
&\beta_1^k, \beta_2^k, \beta_3^k, p \geq 0; \ \alpha_k \text{ unrestricted}
\end{aligned}
\tag{12}
$$

Clearly, in (12) above, $p = 1$.

Define the function[8]

$$R^k(x_1, x_2, x_3) = \alpha_k + \beta_1^k x_1 + \beta_2^k x_2 + \beta_3^k x_3 \tag{13}$$

Then $R^k(x^j) \geq y_j$ for each j and $R^k(x^k) = \hat{f}(x^k)$.

We may express the deviation between y_j and $R^k(x^j)$ as

$$\varepsilon_j^k = R^k(x^j) - y_j \, (j = 1, 2, ..., k, .., N). \tag{14}$$

Given that $R^k(x^k) = \hat{f}(x^k)$, as noted above, the problem (12) can be reformulated as

[8] The function $R^k(.)$ is a supporting hyperplane constituting a local outer approximation to the frontier of the production possibility set and is one of the facets of the *overproduction function* (Varian, 1984). For a more detailed discussion see Ray (2004) chapter 10.

$$\begin{aligned} &\min\ \varepsilon_k^k = R^k(x^k) - y^k \\ &s.t.\ \varepsilon_j^k = R^k(x^j) - y_j \geq 0 (j = 1, 2, ..., k, .., N). \end{aligned} \qquad (15)$$

Several points need to be noted here:

(i) While the non-negative residual ε_k^k measures the deviation of y_k from the frontier $\hat{f}(.)$ evaluated at the input bundle x^k, the other residuals $\varepsilon_j^k (j \neq k)$ have no such interpretation.

(ii) For any other input bundle $x^j (j \neq k)$, the corresponding tangent hyperplane $R^j(x_1, x_2, x_3) = \alpha_j + \beta_1^j x_1 + \beta_2^j x_2 + \beta_3^j x_3$ will have a different gradient vector.

(iii) The efficiency residuals $\varepsilon_j^j (j = 1,2,...,k,..,N)$ are obtained independently rather than jointly.

In (12) or (14) above, we impose no additional constraints beyond the non-negativity of the β s. Now impose a further constraint

$$a\beta_1^k - b\beta_2^k = 0. \qquad (16)$$

That is, $\beta_2^k = \frac{a}{b}\beta_1^k$. The restricted version of (12) would then be

$$\begin{aligned} &\tilde{R}(x^k) = \min \alpha_k + \beta_1^k (x_{1k} + \tfrac{a}{b} x_{2k}) + \beta_3^k x_{3k} \\ &s.t. \alpha_k + \beta_1^k (x_{1j} + \tfrac{a}{b} x_{2j}) + \beta_3^k x_{3j} \geq p y_j; \\ &p y_k = y_k; \\ &\beta_1^k, \beta_2^k, \beta_3^k, p \geq 0;\ \alpha_k \text{ unrestricted.} \end{aligned} \qquad (17)$$

Define, now, the aggregated input

$$X_{1j} = x_{1j} + \tfrac{a}{b} x_{2j}. \qquad (18)$$

Problem (17) would then become

$$\begin{aligned}
&\tilde{R}(x^k) = \min\ \alpha_k + \beta_1^k X_{1k} + \beta_3^k x_{3k} \\
&s.t.\ \ \alpha_k + \beta_1^k X_{1j} + \beta_3^k x_{3j} \geq py^j; \\
&py_k = y_k; \\
&\beta_1^k, \beta_3^k, p \geq 0; \alpha_k \text{ unrestricted}
\end{aligned} \tag{19}$$

The dual of this problem is

$$\begin{aligned}
&\tilde{f}(x_{1k}, x_{2k}, x_{3k}) = \max \varphi y_k \\
&s.t. \sum_{j=1}^{N} \lambda_j y_j \geq \varphi y_k; \\
&\sum_{j=1}^{N} \lambda_j x_{1j} \leq X_{1k}; \\
&\sum_{j=1}^{N} \lambda_j x_{3j} \leq x_{3k}; \\
&\sum_{j=1}^{N} \lambda_j = 1; \lambda_j \geq 0\,(j = 1, 2, ..., N)
\end{aligned} \tag{20}$$

Obviously, when $a = b$, X_1 is simply the sum of the quantities of the inputs x_1 and x_2. In that case, the two inputs are treated as perfect substitutes. Further, because (20) is a restricted version of (12), the minimum value of the objective function at the optimal solution of (20) will be no lower than what is obtained at the optimal solution of (12). Therefore,

$$\tilde{\varepsilon}_k^k = \tilde{f}^k(x^k) - y_k \geq \varepsilon_k^k = \hat{f}(x^k) - y_k. \tag{21}$$

The former is the restricted and the latter the unrestricted residual.

3.2 Input Aggregation in Parametric Models

For input aggregation in a parametric specification of the production function, one assumes homothetic separability. Consider, again, the production function

$$y = f(x_1, x_2, x_3) = F(a(x_1, x_2), x_3). \tag{22}$$

Assume further $X_1 = a(x_1, x_2)$ is homogeneous of degree 1 and is functionally separable from x_3. Linear homogeneity implies $X_1 = a_1 x_1 + a_2 x_2$, where $a_1 = \frac{\partial a}{\partial x_1}$ and $a_2 = \frac{\partial a}{\partial x_2}$. Separability implies that the marginal rate of substitution between x_1 and x_2 is independent of x_3. That is $\frac{\partial(\frac{a_2}{a_1})}{\partial x_3} = 0$.

Now, if the firm selects the inputs (x_1, x_2) to minimize $C(X_1) = w_1x_1 + w_2x_2$ for the sub-aggregate input X_1, $\frac{a_2}{a_1} = \frac{w_2}{w_1}$. Hence, $X_1 = x_1 + \frac{w_2}{w_1}x_2 = \frac{w_1x_1 + w_2x_2}{w_1}$. Using w_1 as the numeraire, we get

$$X_1 = w_1x_1 + w_2x_2. \qquad (23)$$

3.3 Statistical Tests for Input Aggregation

In econometric models, test of linear restrictions on parameters are usually conducted by comparing the residual sums of squares from the restricted and the unrestricted model in an appropriate F test. A comparable F test developed by Banker (1993) can be employed to test whether the efficiency residuals from the restricted DEA model are significantly bigger as a whole than the unrestricted DEA residuals.[9] Note that the restriction (10) results in the input aggregation defined in (17). Hence, this F test is in effect a test of the validity of the way the inputs have been aggregated.

The multipliers β_1^k and β_2^k at the optimal solution of (12) denote the shadow prices of the corresponding inputs of firm *k*: x_1^k and x_2^k. A problem often encountered in DEA (as in other linear programming models) is that at the optimal solution, the shadow price becomes zero because a constraint is non-binding. To avoid this problem one may impose upper and lower bounds of the form

$$L \le \frac{\beta_2^k}{\beta_1^k} \le U \qquad (24)$$

on the ratio of a pair of shadow prices. This is known as assurance region (AR) analysis.

Note that one gets the restriction (16) above by setting

$$L = \tfrac{a}{b} = U. \qquad (25)$$

In that sense, aggregation of the two inputs in the manner described above is a degenerate case of AR analysis.

The lower and upper bounds in AR analysis are often arbitrarily chosen. However, the cost minimizing behavior of a competitive firm can provide guidance in selecting the ratio $\frac{a}{b}$ from available data. When firm *k* is allocatively efficient in its choice of the input quantities x_1^k and

---

[9] Pastor, Ruiz, and Sirvent (2002) performed a nonparametric statistical test of nested radial DEA models to determine the optimal choice of inputs and outputs.

x_2^k, the marginal rate of substitution will be equal to the ratio of the two input prices faced by the firm. In that case,

$$\frac{\beta_2^k}{\beta_1^k} = \frac{w_2^k}{w_1^k}. \tag{26}$$

A valid measure of the aggregate input would then be[10]

$$X_{1j} = x_{1j} + \tfrac{w_2^k}{w_1^k} x_{2j}. \tag{27}$$

As argued by Banker et al. (2007), Banker's F test would then become a test of allocative efficiency.

In the existing literature it is standard practice to assume that all firms face the same prices of the inputs. In that case

$$\frac{w_2^k}{w_1^k} = \frac{\overline{w}_2}{\overline{w}_1} \text{ for each firm } k. \tag{28}$$

When this is not the case, one must construct a different measure of the aggregate input by using a different price ratio in order to evaluate the efficiency of each firm *k*. That is when evaluating firm *r,* the aggregate input is constructed as

$$X_{1j}^r = x_{1j} + \tfrac{w_2^r}{w_1^r} x_{2j}. \tag{29}$$

However, when evaluating a different firm, *s*, the appropriate way to measure the aggregate input would be to define

$$X_{1j}^s = x_{1j} + \tfrac{w_2^s}{w_1^s} x_{2j} (j = 1,2,..,N). \tag{30}$$

The relevant LP problem (17) for any firm *k* (*k=1,2,...,,N*) would then become

[10] In aggregating over more than two inputs, any one input price may be regarded as a *numeraire* and the other inputs may be weighted by their relative prices.

$$
\begin{aligned}
&\tilde{f}(x_{1s}, x_{2s}, x_{3s}) = \max \varphi y_s \\
&s.t. \sum_{j=1}^{N} \lambda_j y_j \geq \varphi y_s; \\
&\sum_{j=1}^{N} \lambda_j x_{1j}^s \leq X_{1k}^s; \\
&\sum_{j=1}^{N} \lambda_j x_{3j} \leq x_{3k}; \\
&\sum_{j=1}^{N} \lambda_j = 1; \lambda_j \geq 0 \, (j = 1, 2, ..., N)
\end{aligned}
\tag{31}
$$

Note that the aggregate input X_1 needs to be measured differently while solving the problem (31) for each individual firm unless all firms face the same prices of the two inputs.

It is possible, of course, that in a particular application input (or output) prices are not available and one must use the actual cost (or revenue) data in the model as in Banker, Chang, and Natarajan (2007). This, it may be noted, implicitly assumes that input prices are identical across firms.

3.3.1 Banker's F Test in DEA

It may be noted that the DEA efficiency residuals ε_j are obtained independently of each other. This is in contrast with the frontier production function model proposed by Aigner and Chu (1968). In their case, a single parametric function is fitted to the entire data set and the efficiency residuals are jointly derived and, therefore, are not independent of one another. Now suppose that we choose a probability density function $f(\ .)$ such that $f(\varepsilon_j)$ is monotone decreasing in the efficiency residuals. In that case, because the DEA estimate of the production function minimizes each ε_j, it thereby maximizes each $f(\varepsilon_j)$. Hence, the DEA frontier $g^*(x)$ maximizes the likelihood function subject to the constraints specified above.

Banker specifies the deterministic frontier where the random inefficiency component of y appears in an additive manner. One may directly link the one-sided econometric frontier with the DEA frontier by specifying (18) differently as

$$
y = g(x)e^{-\varepsilon}; \varepsilon \geq 0 \tag{32}
$$

leading to

$$g(x) = \varphi y; \varphi \geq 1. \qquad (33)$$

Thus,

$$\varepsilon = \ln(\varphi) \geq 0. \qquad (34)$$

Note that as argued in Banker (1993), interpretation of the DEA frontier $g^*(x)$ as a maximum likelihood estimator of the unknown frontier $g(x)$ remains valid.

An alternative conceptualization of the frontier would be to directly specify (21) with the non-negative efficiency residual defined as

$$\varepsilon_k^k = \frac{\tilde{f}^k(x^k) - y_k}{y_k} = (\varphi_k - 1) \geq 0. \qquad (35)$$

Banker has proposed a number of statistical tests for comparing two groups of firms to assess whether one group is more efficient than the other. Assume that there are N firms in the sample of which m_1 are in group 1 and m_2 are in group 2. Firms in group1 have the exponential distribution of (in)efficiency ε_j with parameter σ_1 and those in group 2 also have the exponential distribution but with parameter σ_2. Designate the first group of firms as M_1 and the second group as M_2. Consider the residuals $\varepsilon_j^*(j = 1, \ldots, N)$ obtained from DEA. Under the maintained hypothesis, the sample statistic

$\sum_{j \in M_i} \frac{\varepsilon_j^*}{\sigma_i}$ has the χ^2 distribution with $2m_i$ $(i= 1, 2)$ degrees of freedom.

Under the null hypothesis $\sigma_1 = \sigma_2$, the test statistic

$$F = \frac{\sum_{j \in M_1} \varepsilon_j^* \Big/ m_1}{\sum_{j \in M_2} \varepsilon_j^* \Big/ m_2} \qquad (36)$$

has the F distribution with $(2m_1, 2m_2)$ degrees of freedom.

On the other hand, if the ε_j s have the half Normal distribution, (i.e., the Normal distribution with mean 0 and variance σ^2 truncated from below at 0), then $\sum_{j\in M_1}\left(\frac{\varepsilon_j^*}{\sigma_1}\right)^2$ has the χ^2 distribution with m_1 degrees of freedom. Similarly, $\sum_{j\in M_2}\left(\frac{\varepsilon_j^*}{\sigma_2}\right)^2$ has the χ^2 distribution with m_2 degrees of freedom.

Hence, in this case, under the null hypothesis $\sigma_1 = \sigma_2$, the statistic

$$F = \frac{\sum_{j\in M_1}(\varepsilon_j^*)^2 \Big/ m_1}{\sum_{j\in M_2}(\varepsilon_j^*)^2 \Big/ m_2} \tag{37}$$

has the F distribution with (m_1, m_2) degrees of freedom.

One would use $\varepsilon_k^* \equiv \ln(\hat{\varphi}_k)$ for the aggregated (i.e., the restricted) model in the numerator and $\varepsilon_k^* \equiv \ln(\varphi_k)$ for the disaggregated (i.e., the unrestricted) model in the denominator in testing for the validity of input aggregation. The aggregated model would be rejected only when the test statistic exceeds the critical value for the relevant degrees of freedom, for any specific distributional assumption.

3.4 A Statistical Test for Nested Radial DEA Models

In parametric regression analysis, one typically applies the *t* test at the desired level of significance to decide whether or not any individual regressor should be included in the estimated model. Pastor, Ruiz, and Sirvent (2002) proposed an analogous test for model selection in terms of what they describe as *marginal efficiency contribution* (MEC) of any individual input or output.

The underlying logic behind their test is quite simple. Suppose that an output-oriented model includes m outputs and n inputs and the resulting radial output expansion factors for the N individual units are $\varphi_j^* \ (j=1,2,...,N)$. Next consider the model without one of the inputs. Because the new model has one fewer restriction (corresponding to the excluded input) we get a new set of optimal values $\varphi_j^{**} \ (j=1,2,...,N)$ for the same units. Define $\rho_j = \frac{\varphi_j^{**}}{\varphi_j^*}$. Clearly, because the model excluding one input is less restrictive, $\varphi_j^{**} \geq \varphi_j^{*}$ so that $\rho_j \geq 1$. At this point, the analyst has to select a critical value $\overline{\rho} > 1$ (e.g., $\overline{\rho} = 1.10$) and define the binary variable T_j which equals 1 if $\rho_j > \overline{\rho}$ and 0 otherwise. Stated differently, $T_j = 1$ indicates that exclusion of the input under consideration significantly alters the optimal value of the DEA problem for unit *j*. Viewed this way, the random variables T_j are binary outcomes of a series of independent Bernoulli trials with some constant probability *p* of 'success' (i.e. $T_j = 1$). One may select some hypothesized probability p_0 and test the null hypothesis

$H_0 : p < p_0$ against the alternative $H_1 : p > p_0$.

If H_0 is rejected, one may conclude that exclusion of the relevant input significantly affects the distribution of efficiency and, hence, it is a relevant input. The p-value[11] for the test statistics is $1 - F^B(N-1, p_0)$ where F^B is the cumulative density function for the Binomial distribution $B(N-1, p_0)$.

3.5 DEA and Bootstrap

Simar (1992, 1996), Simar and Wilson (1997a, 1997b) set the foundation for the consistent use of bootstrap techniques to generate empirical distributions of efficiency scores and have developed tests of hypotheses relating to returns to scale through bootstrapping. Following Simar and Wilson (1997a) we can describe the existing bootstrap techniques for the output-oriented technical efficiency measure from an output oriented radial DEA model with the following algorithm:

i) Solve the DEA LP problem to obtain $\hat{\varphi}_j$ for each DMU j=1,2,…,N.

[11] The p-value is the probability of getting T_0 or more 'success' in a series of N-1 independent trials where the probability of success in any trial is p_0.

ii) Select the b-th (b=1,2,…,B) independent naive bootstrap sample $\{\varphi^*_{1,b}, \varphi^*_{2,b}, ..., \varphi^*_{N,b}\}$, which consists of N data values drawn with replacement from the estimated values $\hat{\varphi}_j$ s.

iii) Construct the smoothed bootstrap sample $\{\varphi^{**}_{1,b}, \varphi^{**}_{2,b}, ..., \varphi^{**}_{N,b}\}$, from the naïve bootstrap sample. Notice that all the φ_j s are greater than or equal to 1. Therefore, the smoothed bootstrap sample should be appropriately bounded. It will be computed according to:

$$\varphi^{**}_{j,b} = \begin{cases} \varphi^*_j + h\varepsilon_j & \text{if } \varphi^*_j + h\varepsilon_j \geq 1 \\ 2 - (\varphi^*_j + h\varepsilon_j) & \text{otherwise} \end{cases} ; \text{ for j } = 1,2,...,N. \tag{38}$$

As proposed in Silverman (1986), $h = 0.9\, A\, n^{-1/5}$, where $A = \min\left\{s.d.; \frac{IQR}{1.34}\right\}$ is the optimal bandwidth that minimizes the approximate mean integrated square error of the distribution of the DEA values of φ.

iv) Create the b-th pseudo-data set as $\{(x^{j*}, y_j^* = y_j(\frac{\hat{\varphi}}{\varphi^{**}_j})); j=1,2,...,N\}$.

v) Use the pseudo-data set to compute new $\hat{\varphi}^*_j$ s from the output oriented DEA LP problem.

vi) Repeat steps (ii)-(iv) B-times to obtain $\{\hat{\varphi}^*_{j,b};\ b=1,2,...,B\}$ for each DMU j, $j=1,2,...,N$.

vii) Calculate the average of the bootstrap estimates of φ s for each unit j $\overline{\varphi}^*_j = \frac{1}{B}\sum_{b=1}^{B}\varphi^*_{j,b}$. The bias corrected estimate of φ_j is $\varphi^{bc}_j = 2\hat{\varphi}_j - \overline{\varphi}^*_j$ .Use the empirical distribution function of $\hat{\varphi}^*_{j,b}$ to construct the relevant bootstrap confidence intervals for φ_j.[12]

It should be noted that an interpretation of the results obtained from the bootstrap procedure is not always clear. For example, in the bth replication using the pseudo-data consisting of the

[12] See Simar and Wilson (2000) and Sickles and Zelenyuk (2019) section 9.3.1 for a detailed discussion.

actual input bundles coupled with the fictitious output levels of firms, the optimal solution φ^* shows the scalar expansion factor for the fictitious output quantity and its inverse is *not a measure of the efficiency of the actual input output bundle.* It is possible that the actual input-output bundle may lie above the production frontier constructed from the pseudo-data obtained in any one bootstrap sample. One may, of course, use the optimal solutions from the (bootstrap) DEA problems to construct measures of the *frontier output level* producible from the fixed input bundle of a firm. Thus, it is more meaningful to construct a 95% confidence interval of the maximum output with lower and upper bounds $[y_L^*, y_U^*]$. In principle, the upper bound (y_U^*) may be used to derive a probabilistic measure of the technical efficiency of an observed input-output bundle. It is still possible that the actually observed output from a given input bundle may exceed its corresponding upper bound.

4. Non-discretionary Inputs[13]

Variable inputs are varied while fixed inputs are held constant (even as the level of production changes) *at the discretion of the firm*. There are, however, other factors, that affect the output produced from the inputs chosen by the firm but are beyond its control. For a simple example, consider irrigation in farming. The farmer can increase the level of irrigation by choosing a bigger pump but may want to wait until its existing pump is sufficiently depreciated. Rainfall is an alternative source of water for crops. However, the farmer cannot alter the amount of rainfall at his discretion. These are treated as *non-discretionary* inputs. One can think of the discretionary inputs as directly productive and the non-discretionary inputs as facilitating production.[14]

As noted earlier, the production possibility set consists of all feasible input-output bundles and is thus defined in the input-output space. However, production takes place in a specific physical, social, and cultural environment. Differences in environmental conditions can play a decisive role in defining the feasibility of a particular input-output bundle. In measuring the efficiency of

[13] See also the section on 'contextual variables' in Ray (2022) Ch 12 of volume 1of this handbook.
[14] This description was suggested to me by Subal Kumbhakar.

a decision-making unit, we assume that it can choose the input bundle it uses, or the output bundle it produces. Unlike inputs or outputs, the environmental factors cannot be chosen by the firm and must be treated as 'non-discretionary'.

An obvious example of an environmental factor is rainfall in the context of agricultural production. The maximum output producible from a given bundle of inputs (say labor, fertilizer, and land) depends on the amount of rainfall. In that sense, rainfall contributes to the output much the same way as irrigation. However, while the farmer can choose the level of irrigation, the amount of rainfall is not within his control. Here, rainfall acts as a non-discretionary input. Note that in defining the feasible set for a DEA LP problem, one has to include a constraint for the amount of rainfall. However, when the radial input-oriented technical efficiency is to be measured, the proportional scaling factor applies only to the discretionary inputs (like labor, fertilizer, and land) but not to rainfall.

For another example, consider a secondary school where the average performance of its pupils in a standardized test in mathematics is one of the outputs and hours of classroom instruction in math is one of the inputs. An increase in this input is expected to improve the average test score in math. Now consider another variable – the median family income of the town where the school is located. There is ample evidence to conclude that students from more affluent families where the parents are professionals are better motivated and spend more time on homework and perform better in tests. In that sense, the economic status of the pupil acts like class time spent on math. However, the former is an input while the latter is a contextual variable. They are also referred to as *non-discretionary inputs*. Another example of a contextual variable is the marital status of parents of a pupil. A child from a single parent family (irrespective of income) is unlikely to get the same level of parental attention as when both parents are present. Thus, an increase the proportion of pupils from single parent families will lower the average math score even when all other variables are unchanged. Two things emerge out of this illustrative example. First, unlike an input, a contextual variable may be either favorable (like family income) or detrimental (like single parent families). Second, a decision maker at an appropriate level of authority (the school superintendent or the Board of Education) can select the input bundle used. This is not true for the contextual variables.

For yet another example, consider the efficiency of a water utility. The outputs are the number of customers served and the gallons of water distributed. The inputs are pumps, length of pipe lines and hours of labor. Note that in an urban area the higher density of population implies that the same number of customers can be served and the same volume of water dispensed with a smaller network of pipelines than what is required in a rural area. Moreover, when many customers are located in the same building (as is the case in an urban community) the labor hours needed for meter reading will be lower than in a rural area where customers are located at distant points. In this case, density of population is an environmental variable.

Where contextual variables are considered to be significant determinants of performance, an appropriate way to conceptualize the production technology is to define the production possibility set conditional on a specific vector of contextual variables z^0:

$$T(z^0) = \{(x, y) : y \text{ can be produced from } x \text{ given the contextual variables } z^0\}. \quad (39)$$

One then has a family of conditional production possibility sets. Efficiency is still evaluated at the inputs used and outputs produced. But the appropriate benchmark bundle depends on the applicable vector of contextual variables. The disposability and convexity assumptions about the technology apply to the input-output set but is not necessarily extended to the contextual variables z.

Consider the revised transformation function

$$F(x, y; z) = k. \quad (40)$$

The conditional production possibility set would then be:

$$T(z^0) = \left\{(x, y)" F(x, y; z^0) \leq 0\right\}. \quad (41)$$

There are two different ways to formulate the DEA problem depending on how the revised transformation function is conceptualized. One possibility is to consider the function multiplicatively separable between (x, y) and z as

$$F(x, y, z) = a(x, y).b(z) = k\ . \quad (42)$$

In this case, any change in the contextual variables causes a neutral shift in the production frontier would altering the marginal rates of substitution between inputs or marginal rates of transformation between outputs. Such multiplicative separability permits a 2-stage analysis where a conventional DEA model is solved using only the inputs and outputs in the first stage followed by a regression of the DEA efficiency scores on the contextual variables in the second stage. It is important to recognize that this 2-stage analysis is permissible only if the environmental variable (the regressors) in the second stage are uncorrelated with the inputs and outputs in the first stage DEA. This 2nd stage regression analysis was proposed by Ray (1988, 1991) and is quite widely used in empirical applications to *explain* variations in measured levels of DEA efficiency.

The other option is to include the contextual variables along with the inputs and outputs in a comprehensive DEA problem. In this case, favorable environmental factors are treated *like inputs* because then contribute positively towards outputs while unfavorable environmental factors are treated *like outputs* because they demand additional inputs. However, even though appropriate inequality constraints are included for these factors, the proportional scaling factors do not apply to them. This one step analysis would be recommended when the contextual factors are correlated with some inputs. In the water utility example, the population density in the area served affects the specific inputs like length of network or labor but not other inputs like the number of pumps. In the school example on the other hand, there is no reason to believe that socioeconomic factors (either good or bad) would be related specifically to some inputs and considering a neutral shift is more appropriate.

There are two important points to note about a one-step DEA. First, the analyst must make a prior judgment about whether any relevant contextual variable is favorable or unfavorable because that would determine the nature of the inequality in the relevant constraint. Second, and more importantly, the convexity assumption about the input-output bundles may not be applicable for some contextual variables. This is particularly true for categorical variables. Suppose that in our water utility example, the service areas are classified as rural, urban, and metropolitan but the exact measure of population density is not available for each observation. In this case, all we know is that water delivery is most difficult in the rural areas and the least difficult in the metropolitan areas. Creating convex combinations of a categorical variable representing population density is not meaningful in this context. Following Banker and Morey (1986) one may handle this by treating the conditional production possibility sets as nested in the sense that all input-output bundles that are feasible in a less favorable condition are also feasible in in a more favorable condition but not the other way around. In this case, we include only the rural observations to construct the frontier for evaluating utilities serving rural areas but all observations to construct the frontier that is to be used for evaluating the utilities serving the most densely populated areas. It should be noted though that for multiple contextual variables that are categorical such cross-classification may severely restrict the number of observations available for constructing the frontier for the less favored groups.

4.1 2nd Stage Regression[15]

Ray (1988, 2004) conceptualized a 1-output production function of the form

$$y = g(x).h(z)e^{-\eta}; \eta \geq 0; 1 \geq h(z) \geq 0 \qquad (43)$$

and argued that the DEA radial output-oriented efficiency score $\tau_y = \frac{1}{\varphi^*}$ is a measure of $\frac{y}{g(x)} = h(z)e^{-\eta}$ when $g(x)$ is additive.[16] Assume further that $\ln h(z) = z'\beta$. This leads to the 2nd stage regression

$$\ln \tau_y = z'\beta + u; u \leq 0. \qquad (44)$$

One can estimate this regression by OLS to get consistent estimators of the slope parameters. However, by construction, some of the fitted residuals $(\hat{u})$ will be positive. To overcome this problem, one can adjust the intercept upwards by the largest positive residual so that the so

[15] This section is not designed to be comprehensive review of different approaches to measuring the impact of contextual variables on the measured efficiency through a second stage regression. The interested reader may refer to chapters 9 and 10 of Sickles and Zelenyuk (2019)

[16] Unless x and z are correlated and if the g(x) is additive, there should not be any bias. However, the DEA efficiency measure includes effects of difference in z (like soil type, public infrastructures, etc). The 2nd stage regression is designed to appropriately shift the frontier downwards relative to the best possible environment. In the absence of CRS, the additivity property of g(x) does not hold and there will be a bias due to convexity.

called "Greene corrected" residuals $(u_j^* = \hat{u}_j - \hat{u}_{\max})$ are all non-positive. Ray suggests using $e^{u_j^*} \leq 1$ to measure efficiency corrected for the contextual variables.

Two specific problems should be highlighted. First, when $g(x)$ is strictly concave, the DEA efficiency measures will be biased upwards.[17] Also, the 'fitted value' $\hat{h}(z) = e^{z'\hat{\beta}}$ may exceed 1 for some observations. Moreover, as noted above, the 2nd stage OLS regression will not have the standard statistical properties of the classical linear regression.

Banker and Natarajan (2008) and Banker, Natarajan, and Zhang (2019) frame the 2nd stage regression model differently. First, in the spirit of the parametric stochastic frontier analysis, they include a two-sided random noise along with the one-sided efficiency. Thus, their model is

$$y = g(x).h(z)e^{v-u}; 1 \geq h(z) \geq 0 \qquad (45)$$

Further, while $u \geq 0, v_M \geq v \geq -v_M$. Thus, the random noise, although two-sided, is bounded both from above and below. Also, they specify $h(z) = e^{-z'\beta}$ and only positive values of both the coefficient vector β and the contextual variables z. These non-negativity restrictions automatically ensure $1 \geq h(z) \geq 0$. Thus, their model becomes

$$\ln(\tfrac{y}{g(x)}) = v - u - z'\beta \qquad (46)$$

For the simplest case of a single contextual variable the regression with an intercept becomes

$$\ln(\tfrac{y}{g(x)}) = v - u - \beta_1 z \qquad (47)$$

Define $\varepsilon = v - u$ and $\tilde{\varepsilon} = \varepsilon - E(\varepsilon)$. Then the regression can be written as

$$\ln(\tfrac{y}{g(x)}) = \beta_0 - \beta_1 z + \tilde{\varepsilon}; where\, \beta_0 = E(\varepsilon) - v^M. \qquad (48)$$

[17] Because in that case, $\sum_j \lambda_j g(x^j) < g(\sum_j \lambda_j x^j) \mid \sum_j \lambda_j = 1.$

One can use the log of the DEA score $\tau_y = \frac{1}{\varphi^*}$ as the dependent variable and estimate the 2nd stage regression by OLS.

While the OLS regression leads to consistent estimators of the coefficients of the contextual variables, the pure inefficiency measures (unrelated to these variables) cannot be derived directly. Banker and Natarajan (2008) also consider maximum likelihood estimation of the 2nd stage regression and derive conditional mean and mode of the distribution of the one-sided term (*u*) (comparable to the formulas in Jondrow, Lovell, Materov, and Schmidt (1982) but advise using the OLS residuals to rank order the units in terms of efficiency.

4.2 Truncated Regression in the 2nd Stage

Simar and Wilson (2007) extend their earlier models of smoothed bootstrap of DEA efficiency scores to incorporate contextual variables in a second stage regression. However, they object to using either OLS or a Tobit model (i.e., censored regression) and instead propose a truncated regression. Using the optimal value of the DEA output-oriented LP problem, they consider a 2nd stage model

$$\varphi_i^* = z^{i\prime}\beta + u_i \geq 1 \Rightarrow u_i \geq 1 - z^{i\prime}\beta. \qquad (49)$$

One can specify a normal distribution $N(0,\sigma_u^2)$ for u and use MLE for truncated regression to estimate the model parameters ($\tilde{\beta},\tilde{\sigma}_u^2$) using only those observations for which the φ^* is strictly greater than 1. They advise bootstrap by drawing pseudo values $(\tilde{u}_{ib})$ of the residuals from a Normal distribution with variance $\tilde{\sigma}_u^2$ truncated at $1 - z^{i\prime}\tilde{\beta}$ and constructing a new bootstrap sample $\varphi_{ib}^* = z^{i\prime}\tilde{\beta} + \tilde{u}_{ib}$ for truncated regression.

In the computational procedure described above bootstrapping is applied only in the 2nd stage but the DEA efficiency scores are computed only once with the actual data for the initial estimation of the truncated regression. Simar and Wilson (2007) also proposed an alternative and more elaborate computational algorithm that can be described as ‘double bootstrap’ where the DEA results are also bootstrapped.[18]

4.3 Contextual Variables in Parametric Models

In stochastic frontier analysis, a comprehensive maximum likelihood estimation is applied to estimate the production function along with the ‘efficiency function’ nested into it. Typically, the specified model is

$$\ln y = \beta' \ln x + v - u \qquad (50)$$

where the random noise v is distributed as $N(0,\sigma_v^2)$ while the inefficiency component u is some one-sided distribution either naturally non-negative (e.g., the exponential or he gamma distribution) or a two-sided distribution truncated from below at 0. A popular choice for u is the truncated Normal distribution $|N(\mu_u,\sigma_u^2)|$. To incorporate the contextual variables, one expresses the mean of the pre-truncated distribution as $\mu_u = \alpha' z$. All parameters of the model $(\beta,\alpha,\sigma_v^2,\sigma_u^2)$ are estimated by maximizing the likelihood function. Estimates of efficiency at individual data points are obtained by the mean (or mode) of the one-sided component conditional upon the composite disturbance.[19] One can also make σ_u^2 a function of the z variables. Wang (2002) considers a framework where z affects both the mean and variance.

4.4 Choice between Inputs and Contextual Variables

Except in the obvious cases like rainfall, classification of factors affecting output as inputs or contextual variables is not quite straightforward. In principle, an input is something that the producer can choose. For example, the proportion of students from non-English language speaking families has an impact on the average test score of a school district. But the school superintendent cannot change this socio-demographic characteristic of its student population. In that sense it is an exogenously determined (non-discretionary) variable. However, the town’s

[18]For details of the steps involved, see Simar and Wilson (2007).
[19] See the chapters on SFA by Kumbhakar, Parmeter, and Zelenyuk in volume 1 of this Handbook for details.

Board of Education may adopt a policy of encouraging enrollment from such families. In that case it becomes an input (like the number of teachers per student).

In the end, classification of variables as inputs or contextual is a judgment call for the analyst. In general, however, the greater the scope of decision making, the fewer are the variables deemed contextual.

5. Input-Output Choice in Some Areas of Application

5.1 Manufacturing[20]

A very popular subject of efficiency and productivity analysis is manufacturing.

Output:

In this sector, output can be defined in alternative ways. The three main choices are:

(i) value added;

For value added the only two inputs included should be Labor and Capital

(ii) sales/shipment,

Sales or shipment does not reflect total production. Should be adjusted for change in inventories.

(iii) gross output

Gross output should be used only when intermediate inputs (materials and energy) are included.

Inputs: The inputs included are labor, capital, energy, and materials.[21]

(i) Labor:

Typically, one distinguishes between production workers and non-production workers.

[20] See also the chapter by Bhoumik in this volume.
[21] Recently, purchased services are included as yet another input is the so called KLEMS (Capital-Labor-Energy-Materials-Services) data sets constructed at the 2-digit industry level for different countries. But there is hardly any KLEMS analysis at the firm level.

Because production workers often scheduled for overtime during peak periods and fewer hours during slack times, a more accurate measure of this labor input is the number of hours worked.

Non-production workers are usually full-time employees. Non-production labor is measured by the number of employees.

(ii) Capital:

Capital is, by far, the most difficult input to measure. Apart from the fact that the age and vintage differences of machinery makes it hard to get an aggregate measure of the capital input, there is also the question of stock vs. flow.

Usually, one can get the data only on the value of gross fixed assets. This can be appropriately deflated by a price index of machinery and transport equipment to get a real value of the capital stock. This can be used as a measure of the capital input.

One major shortcoming of this approach is that it fails to take account of differences in capacity utilization across firms and/or over years. The true measure of the capital input is the actual flow of capital services that contributes to production of the output during the production period.

A different approach is to construct a measure of the capital flow input indirectly using the 'user cost approach'. The total user cost associated with the capital flow input can be measured by the sum of interest and amortization, depreciation, rent, and repair and maintenance expenses. Next one needs to construct a price of capital services. One can start with the wholesale price index of transport and machinery and compute the interest cost using an appropriate interest rate. For example, suppose that the machinery price index in a particular year is 250 and the interest is 5%. Next assume that the relevant depreciation rate is 7% and repairs and maintenance add another 3%. Hence, the user cost is 15% of 37.50 per year of capital service of one machine. If the total user cost of capital for a firm is 7500. We can measure the flow of capital service as $\frac{7500}{37.50} = 200$ machine years.

A note of caution here is that one should never use new capital formation to measure capital input. New capital formation is investment and neither the stock of capital nor the flow of capital services.

(iii) Energy:

In general, information on energy input is available mainly in terms of the expenditure on energy. The principal sources of energy are (a) coal, (b) oil, (c) electricity, and (d) natural gas. One needs to construct a price of energy as a weighted average of these four different energy inputs. The problem is that these fuels are measured in different units. For example, coal is measured in tons while electricity in kilowatt hours. One must, first, reduce these units of measurement to British Thermal Units (BTUs). Thus, price per ton of coal is to be transformed into the price per BTU of energy from coal. Similarly, for other fuels. Next a price of energy per BTU is constructed as the weighted average of these energy source specific prices using the expenditure share of industrial consumption of energy in the geographical region (e.g., the states of the US) where the firm is located. Finally, a measure of the quantity of energy consumed (in BTU) by the firm is obtainable by the ratio of the expenditure on fuels and this price per BTU of energy.

(iv) Materials:

The materials input in most cases includes too many different items and even though the total expenditure on materials may be available, it is difficult to construct a price and deflate the expenditure to get a measure of the quantity of materials consumed. The only reasonable way to measure materials is by the value (effectively setting the price per unit equal to 1). Of course, for inter temporal analysis one can use the price index of industrial raw materials to deflate the nominal amount of expenditure on materials.

5.2 Banking[22]

Although an overwhelming share of empirical applications relate to measurement of efficiency in banking, there is no consensus on what constitute the inputs and outputs of a bank.

[22] For a detailed analysis of the banking industry see the chapter by Miller in this volume.

There are two competing views about the technological character of a bank as a production decision making unit. Berger and Humphrey (1991) describe the two alternative views as *production approach* and *intermediation approach.*

In the production approach a bank is visualized as a provider of different kinds of services to deposit and credit account holders by performing transactions and processing documents like loan application, credit reports, demand drafts, and so on. In this view of a bank, labor, and physical capital (like building and equipment) are the inputs and the numbers of deposit and credit transactions and documents processed are the outputs. Because information on the number of deposit and credit transactions performed or the number of documents processed is not available, researchers use either the numbers of deposit and credit accounts or the monetary values of the balances in these accounts.

In the intermediation approach (also known as the Asset Approach) a bank is seen to be a financial intermediary between savers with funds to lend and investors looking for funds to borrow. The principal act of value addition by a bank is to transform the deposits and other borrowed funds into loans and other investments that yield returns. As a financial intermediary, a bank uses labor and capital along with funds accepted from depositors (used as raw materials) to create income generating assets like loans and investments.

In the intermediation approach based primarily on the model of the short run optimization problem of a financial firm (like a bank) developed by Sealy and Lindley (1977), loans and investments are the outputs while labor, deposits, and (physical) capital are the inputs of the bank.

A choice between the two different approaches is not merely a matter of interpretation of how a bank works but has important implications for selection of inputs and outputs. As Berger and Humphrey (1991) argue, the production approach is more appropriate for evaluating performance at the branch level because branches primarily process documents for the bank as a whole and branch managers have little control over bank funding and investment decisions. For the bank as the unit of evaluation, the intermediation approach is more appropriate because it explicitly includes interest expenses, which account for a major part of the total expenses of a bank.

There are numerous studies, however, at the bank level which treat deposits as output. It is often argued that the bank as a financial firm does offer valuable service to the depositors by keeping their savings secure and deposits, therefore, should be treated as output. But these funds in deposit accounts generate no revenue to the banks unless transformed into credit given out to borrowers.

Sealy and Lindley (1977) argue that the deposited funds are themselves inputs that are turned into 'loanable funds" as intermediate output which is ultimately transformed into revenue generating assets.

Another argument often voiced against considering deposits as input is that it would imply that a bank becomes more efficient if it reduces deposits because economizing on the use of inputs enhances efficiency. But this line of argument is somewhat misleading in that it does not tell the whole story. One must recognize that acquiring more deposits cannot be an end in itself for the bank and if the funds in the deposit accounts are held idle (entailing interest expenses) without being processed into loans and investments generating revenue, an increase in deposits would indeed make the bank less efficient!

A variable that has received little attention in the production efficiency literature on banking is the amount of equity capital of a bank. Very few studies have included it in the list of inputs and outputs. Berger and Mester (1997) argue that within the intermediation or asset approach, the equity capital of the bank along with other liability items like core deposits and purchased funds are sources of funds and should be treated as inputs that enable a bank to produce the revenue generating assets. However, unlike deposits and other funds, the equity capital cannot be changed in the short run. Hence, it ought to be treated as a fixed input. Ray and Das (2010) in their study of short run cost and profit efficiency of Indian commercial banks included a bank's capital and reserve as fixed input.

Finally, there is no consensus about how to treat the number of branches of a bank in the input-output classification. On one hand, increase in the number of branches amount to easier access for retail customers of the bank and can be interpreted as better service quality. In that sense, a branch is an output of the bank. On the other hand, a branch generates more business and can, therefore, be treated as an input. In fact, like deposits, number of branches have been treated as an output in some studies and as input in some others.

An ingenious way to treat branches in the production technology of a bank is to include it as a conditioning variable in a cost function with unequivocal output quantities and input prices as regressors. If the coefficient of the number of branches is negative, then it can be treated as an input because a negative marginal effect implies that increase in this ‘input’ permits reduction in the use of other inputs resulting in a lower cost without reducing the outputs. If, however, the marginal effect is positive, branches should be treated as output because in that case, with more branches, the cost of producing the unchanged quantities of the other outputs becomes higher.

5.3 Health Care

In view of the fast-increasing cost of health care all over the world, it is not surprising that efficiency in delivery of health care has attracted considerable attention from both academic researchers and policy makers and DEA has been extensively used to evaluate the efficiency of hospitals, nursing homes, and other providers of health care.

Conceptually the output of health care is the extent of improvement in health status or the physical quality of life of a patient. Just as in the case of education, the cognitive skills acquired by a pupil can only be indirectly (and imperfectly) measured by standardized test scores, the output of a provider of health care must be measured indirectly by a number of observable and quantifiable indicators.

A hospital as the unit of analysis is regarded as a production decision making unit combining different kinds of labor and capital inputs to provide treatment to different kinds of patients.

<u>Inputs</u>

Three broad kinds of labor considered as separate inputs are numbers of physicians, nurses and paramedical staff, and other workers. Where detailed data are available, a further distinction is made between surgeons and general practitioners.

Measuring the capital input is far more problematic. While the hi-tech equipment account for the bulk of the physical capital input of a hospital lack of data precludes makes it difficult to include them in a measure of the capital input used.

Similarly, the size of the hospital building and the number of intensive care units available are measures of capital input that should be included if the data are available.

In most cases, the number of beds is used as a 'catch all' measure of the capital input. Basically, the number of beds is regarded as a measure of the size of a hospital and the implicit assumption is that other capital inputs are proportional to the number of beds.

Outputs

Outputs are measured by the numbers of inpatient days and outpatient cases treated. In some cases, inpatient care is further broken up into emergency, maternity, and other kinds of treatment. In the case of health care, the pre-existing condition of the patient at the time of admission is an important determinant of the resources need for treatment.

Therefore, adjusting for complexity and intra-diagnostic severity of cases is important while measuring output by the number of patients. In practice, however, even if detailed information on severity and complexity of cases may be obtained for one or two hospitals, it is unrealistic to expect that the analyst will have such information for all hospitals in the sample. A minimal adjustment for complexity may be made by grouping patients as seniors (needing geriatric care), children (receiving pediatric care) and other adults (requiring general care).

Another important aspect of health care that should be taken into consideration in measuring efficiency is the quality of care provided. A provider may be able to treat a larger number of patients from the same bundle of labor and capital inputs if the quality of care is lower.

Few studies of health care efficiency have taken explicit account of quality. Sometimes an index of quality is constructed from patient satisfaction survey. But constructing an objective measure of quality of care is difficult.

Intuitively, a better quality of care restores the patient to normal health sooner and keeps the person healthy for a longer time period in the absence of any new and unrelated ailment.

In the extreme case, negligence in critical situations may result in death of the patient. In some studies, the incidence of patient mortality is treated as a measure of the *lack of quality.* In other words, lower mortality rate *ceteris paribus* implies better quality of care.

Another variable used to measure quality is the proportion of unscheduled readmissions. Here the presumption is that if a patient needs to be readmitted without a scheduled follow up the patient was given less than proper care in the hospital and was discharged even before treatment was complete.

6. Conclusion

Inputs are related to the outputs by the technology. The role of any empirically constructed the production or transformation function is to quantify this relationship. The technology can also be formulated through cost, revenue, and profit functions. But these require additional assumption about optimizing behavior by the producer. In principle, the production technology captures a physical relationship between inputs and outputs. In that sense, identical input bundles should produce identical outputs unless there are differences in the contextual variables – systematic or random. A second stage regression in DEA or incorporating such variables in the statistical distribution of the efficiency component of the error term in stochastic frontier analysis allows one to extract the impact of these contextual variables from the measured efficiency.

Finally, appropriate choice of inputs and outputs is a precondition of any meaningful analysis of the production technology. Whether one wishes to evaluate productive performance through technical efficiency or simply wants to examine returns to scale properties, homotheticity, sub/super additivity or any other characteristic of the technology, improper selection of inputs and outputs will render any empirical analysis of production meaningless. This chapter ends with a note of caution. In some cases (especially among OR analysts) there is a perception that anything that the decision makers want to reduce is an input while anything that they want to increase is an output. An extreme example is one where in an international comparison of efficiency in emission control, fossil fuel is treated as input (because the author believes that nations should use less of it) while non-fossil fuel is treated as an output (because the author considers it desirable to use more of it). One can avoid such ill-conceived model formulation by paying attention to the underlying production economic theory.[23]

References

Adler, N. and B. Golany (2001) "Evaluation of Deregulated Airline Networks using Data Envelopment Analysis with Principal Component Analysis with an Application to Western Europe" *European Journal of Operational Research*, 132, 260-273.

Adler, N. and B. Golany (2002) "Including Principal Component Weights to Improve Discrimination in Data Envelopment Analysis", *Journal of Operational Research Society*, 53, 985-991.

Aigner, D.J. and S. F. Chu (1968), "On Estimating the Industry Production Function," *American Economic Review* 58:4 (September) 826-39.

Banker, R. D. (1993), "Maximum Likelihood, Consistency, and Data Envelopment Analysis: A Statistical Foundation", *Management Science,* 39, 1265-1273.

Banker, R. D., H. Chang, and R. Natarajan (2007), "Estimating DEA Technical and Allocative Inefficiency Using Aggregate Cost or Revenue Data" *Journal of Productivity Analysis* (27) 115-121.

Banker, R.D., A. Charnes, and W.W. Cooper (1984), "Some Models for Estimating Technical and Scale Inefficiencies in Data Envelopment Analysis," *Management Science,* 30:9 (September), 1078-92.

Banker, R.D. and R.C. Morey (1986) "The Use of Categorical Variables in Data Envelopment Analysis", *Management Science,* 32:12 (December), 1613-1627.

Banker, R.D., and R. Natarajan. 2008. Evaluating Contextual Variables Affecting Productivity Using Data Envelopment Analysis. Operations Research, Vol 56, No. 1, pp 48-58

Banker, R.D., R. Natarajan, and D. Zhang. 2019. Two-Stage estimation of contextual variables in stochastic production function models using Data Envelopment Analysis: Second stage OLS versus bootstrap approaches. *European Journal of Operational Research*. Vol 278, pp 368-384.

[23] See the DEA example in section 6.2.5 in Ramanathan (2003).

Berger, A. N., and Humphrey, D. B., 1991. Efficiency of financial institutions: International survey and directions for future research. *European Journal of Operational Research* 98, 175-212.

Berger, A. N., and Mester, L. J., 1997. Inside the black box: What explains differences in the efficiencies of financial institutions? *Journal of Banking and Finance* 21, 895-967.

Bhaumik, S. Technical Efficiency and Its Determinants in the Manufacturing Sector: What We Know and What We Should Know; **;** in Ray, S.C., R. Chamber, and S.C. Kumbhakar (eds) *Handbook of Production Economics* Volume 2 (Springer Nature) (2022 forthcoming)

Charnes, A., W.W. Cooper, and E. Rhodes (1978) Measuring the Efficiency of Decision Making Units, *European Journal of Operational Research* 2:6 (November), 429-44.

Färe, R., S. Grosskopf, and V. Zelenyuk (2004) "Aggregation Bias and its Bounds in Measuring Technical Efficiency" *Applied Economics Letters*, 11, 657-660.

Färe, R. and V. Zelenyuk (2002) "Input Aggregation and Technical Efficiency" *Applied Economics Letters*, 9, 635-636.

Jondrow, J., I.S. Materov, C.A.K. Lovell, and P. Schmidt (1982) "On the estimation of technical inefficiency in the stochastic frontier production function model"; *Journal of Econometrics*, vol. 19, issue 2-3, 233-238.

Kumbhakar, S.C., C. Parmeter, and V. Zelenyuk. Stochastic Frontier Analysis: Foundations and Advances I, in Ray, S.C., R. Chamber, and S.C. Kumbhakar (eds) *Handbook of Production Economics* Volume 1 (Springer Nature) (2022 forthcoming)

Leibenstein, H. and S. Maital (1992) "Empirical Estimation and Partitioning of X-Inefficiency: A Data-Envelopment Approach" *The American Economic Review*, 82, 2, 428-433.

Marx, K. and F. Engels. (translated by Looch, W. and C.P. McGill) *The German Ideology*(Lawrence and Wishart, London) 1938.

Miller, S. Empirical Analysis of Production Economics: Applications to Banking; in Ray, S.C., R. Chamber, and S.C. Kumbhakar (eds) *Handbook of Production Economics* Volume 2 (Springer Nature) (2022 forthcoming)

Murty, S. and R. R. Russell. Bad Outputs in Production Economics**;** in Ray, S.C., R. Chamber, and S.C. Kumbhakar (eds) *Handbook of Production Economics* Volume 1 (Springer Nature) (2022 forthcoming)

Pastor, J.T., J. L. Ruiz, I. Sirvent. 2002. A Statistical Test for Nested Radial DEA Models. *Operations Research*, Vol 50, No. 4, pp 728-735.

Ramanathan, R. *An Introduction to Data Envelopment Analysis*. (Sage Publications India 2003).

Ray, S. (2004) *Data Envelopment Analysis: Theory and Techniques for Economics and Operations Research*, Cambridge University Press, Cambridge, UK.

Ray, S.C. Data Envelopment Analysis: A Nonparametric Approach to Production Analysis; in Ray, S.C., R. Chamber, and S.C. Kumbhakar (eds) *Handbook of Production Economics* Volume 1 (Springer Nature) (2022 forthcoming)

Ray, S.C. Conceptualization and Measurement of Productivity Growth and Technical Change; in Ray, S.C., R. Chamber, and S.C. Kumbhakar (eds) *Handbook of Production Economics* Volume 1 (Springer Nature) (2022 forthcoming)

Ray, S. C., and Das, A., 2010. Distribution of cost and profit efficiency: Evidence from Indian banking. *European Journal of Operational Research* 201, 297-307.

Ray, S.C. and K. Mukherjee. 2005. The Validity of Input Aggregation in DEA Models: Some Statistical Tests; University of Connecticut *Economics Working Paper* 2005-54.

Sealey, C., and Lindley, J., 1977. Inputs, outputs, and a theory of production and cost at depository financial institutions. *Journal of Finance* 32, 1251-1266.

Sickles, R. and V. Zelenyuk. 2019. *Measurement of Productivity and Efficiency* (Cambridge).

Simar, L. 1992. Estimating efficiencies from frontier models with panel data: A comparison of parametric, nonparametric, and semiparametric methods with bootstrapping; *Journal of Productivity Analysis*, 3(1): 171-203.

Simar, L. 1996. Aspects of statistical analysis in DEA-type frontier models; *Journal of Productivity Analysis*, 7(27): 177-2185.

Simar, L. and P.W. Wilson. 1998. Sensitivity Analysis of efficiency scores: how to bootstrap in nonparametric frontier models, *Management Science*. Vol 44, pp 49-61.

Simar, L. and P.W. Wilson. 2000. Statistical Inference in Nonparametric Frontier Models: The State of the Art, *Journal of Productivity Analysis*. Vol 13, pp 49-78.

Simar, L. and P. W. Wilson (2001) "Testing Restrictions in Nonparametric Efficiency Models" *Communications in Statistics: Simulation and Computation*, 30, 1, 159-184.

Simar, L. and P.W. Wilson. 2007. Estimation and inference in two-stage, semi-parametric models of production processes. *Journal of Econometrics*. Vol 136, pp 31-64.

Sirvent, I., J. L. Ruiz, F. Borrás, and J. T. Pastor (2005) "A Monte Carlo Evaluation of Several Tests for the Selection of Variables in DEA Models" *International Journal of Information Technology and Decision Making*, 4, No. 3, 325-344.

Tauer, L.W. (2001) "Input Aggregation and Computed Technical Efficiency" *Applied Economics Letters*, 8, 295-297.

Thomas, A.C. and L.W. Tauer (1994) "Linear Input Aggregation Bias in Nonparametric Technical Efficiency Measurement" *Canadian Journal of Agricultural Economics*, 42, 77-86.

Varian, H. R. (1984), "The Nonparametric Approach to Production Analysis," *Econometrica* 52:3 (May) 579-97.

Wang, H.J.2002. Heteroscedasticity and Non-Monotonic Efficiency Effects of a Stochastic Frontier Model; *Journal of Productivity Analysis*, 18, 241-253.