Semi-Parametric Estimation of Allocative Inefficiency Using Smooth Non-Parametric Frontier Analysis

Taylor McKenzie

Abstract

Data envelopment analysis (DEA) has been widely applied to study the efficiency of firms, efficiency of various stages in production, and how productivity, efficiency, and technology change through time, among many other applications. DEA is a non-parametric analysis, giving it a major advantage over parametric methods that require specification of a specific production function or rely on (usually) low-order Taylor-series approximations, which may not fully capture interactions between inputs. However, DEA produces production frontiers that are not everywhere-differentiable, making it impossible to calculate marginal input productivities needed to directly analyze allocative efficiency of firms. Alternative DEA-based methods that estimate cost or revenue frontiers to study allocative inefficiency assume firms are price-takers and are therefore inappropriate for non-competitive markets. This research investigates smooth non-parametric frontier analysis, a smooth analogue of DEA that permits calculation of marginal input productivities, and applies them to models of allocative inefficiency while making no assumption on the nature of competition. An empirical methodology is developed and applied to U.S. freight rail firms to determine if overcapitalization is prevalent in the rate-regulated industry, as predicted by the Averch-Johnson hypothesis.

1 Literature Review

1.1 Technical and Allocative Efficiency

Technical inefficiency is defined as the deviation between observed and maximum-possible output, measuring the technical transformation of inputs into outputs. Allocative inefficiency is defined as the deviation between observed and profit-maximizing allocations of inputs, capturing the effectiveness with which firms choose input quantities (Aigner et al., 1977). The two concepts are distinct: A firm that is technically efficient may not exhibit allocative efficiency, and vice versa. As an illustration, consider Figure 1. In this plot, the black curve represents the production possibilities frontier, and the slope of the red line is the ratio of output to input price, p_Y/p_X. Since point A is below the frontier, it is not technically efficient. Point B is technically efficient because it lies on the frontier, but not allocatively efficient since the marginal productivity of X does not equal the price ratio at that point. Point C is both technically and allocatively efficient.

Technical efficiency has been estimated and analyzed using a variety of methods. Data envelopment analysis (DEA), described in greater detail in the following subsection, is a non-parametric method that uses linear programming techniques to find a production frontier that envelops the data, which can then be used to infer inefficiency of individual firms. Stochastic frontier analysis (SFA), a parametric method that requires specification of the form of the production function, estimates inefficiency by assuming deviations from the unobserved frontier follow a one-sided distribution (typically half-normal or log-normal) and using data to fit parameters of that distribution. Aigner et al. (1977) describe methods to separately identify technical inefficiency from measurement error using stochastic frontier models.

Allocative inefficiency has been studied using both SFA and DEA. Prior to the widespread use of DEA, Farrell (1957) described methods of evaluating the price efficiency (i.e., allocative efficiency) of firms using production frontiers, under the assumption that firms are price-takers. Färe et al. (1985) later applied DEA to this approach in a seminal work, using revenue as the output in the DEA model to quantify allocative inefficiency. Again, this approach relies on the assumption that firms are price-takers and is therefore inappropriate for non-competitive markets. Should this methodology be applied to firms that are not price-takers, estimation of the cost frontier necessarily omits a key (unobservable) input, *market power*, and can therefore produce biased estimates of allocative inefficiency.

In a seminal work, Schmidt and Lovell (1979) use a stochastic frontier model to study allocative inefficiency. The authors begin with a first-order approximation of the log-production function (i.e., log-log form). Using this form allows the first-order conditions for profit maximization to be expressed in the closed-form

$$\frac{x^1}{x^j} = \frac{w_j \alpha_1}{w_1 \alpha_j}, \tag{1}$$

where x^1 and x^j are the quantities of capital and input j used, respectively, w_1 and w_j are the factor prices of capital and input j, respectively, and α_1 and α_j are elasticities of output with respect to capital and input j, respectively. The authors assume that Equation (1) may not hold empirically. Instead, the authors assume

$$\ln(x^1) - \ln(x^j) = \ln\left(\frac{w_j \alpha_1}{w_1 \alpha_j}\right) + \varepsilon_j, \tag{2}$$

where $\varepsilon = (\varepsilon_2, ..., \varepsilon_M)$ follows a multivariate normal distribution with mean μ and covariance Σ. Under the assumption of a monotonically increasing, concave production function, a positive value of μ_j indicates overcapitalization relative to input j; that is, the firm would be able to realize greater profits by decreasing its capital utilization relative to its utilization of input j. Unlike DEA-based analyses of allocative inefficiency, this SFA method makes no assumptions on the nature of competition. Instead, it relies on direct estimation of the real production function to obtain marginal rates of transformation, then compares those to observed prices, making no assumption about how those prices were realized. However, SFA does require the researcher to choose a functional form for the production function, making the process prone to misspecfication error. If the selected functional form does not fully capture interactions between inputs, estimates of technical and allocative inefficiency could be biased.

Studies of allocative inefficiency complement studies of technical inefficiency, each offering a valuable and varied perspective on production. Importantly, many hypotheses have been made relating inefficiency to competition. For example, Averch and Johnson (1962) predicted that firms in rate-regulated industries would tend to overcapitalize to relax profit constraints imposed by regulators. The empirical analysis presented in this paper will test this hypothesis in the rate-regulated U.S. freight rail industry.

1.2 Data Envelopment Analysis

Data envelopment analysis (DEA) is a nonparametric technique that has been extensively used to analyze productivity and efficiency. The analysis involves solving a linear programming problem that generates a production possibilities frontier that envelops the data and using those results to estimate inefficiency. Specifically, output-oriented efficiency for firm i is defined as

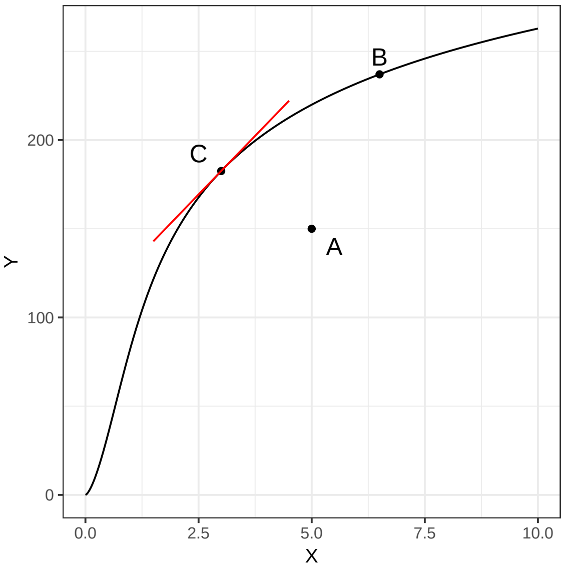

Figure 1: Technical and Allocative Efficiency

$$\phi(x_i, y_i) = \inf\{\theta | (x_i, y_i/\theta) \in S\}, \tag{3}$$

where S is the set of possible production plans (Shepherd, 2015). Under the assumption of constant returns-to-scale, output-oriented efficiency can be found for firm $i = 0$ using DEA by solving the linear programming problem

$$\begin{aligned} &\max \phi \\ &\text{subject to} \\ &\sum_{i=1}^{n} \lambda_i x_i^j \leq x_0^j \qquad j = 1, ..., M \\ &\sum_{i=1}^{n} \lambda_i y_i^k \geq \phi y_0^k \qquad k = 1, ..., K \\ &\lambda_i \geq 0 \qquad i = 1, ..., N. \end{aligned}$$

In this system, m indexes inputs, k indexes outputs, and i indexes firms. The efficiency of firm $i = 0$ (i.e., the ratio of observed to maximum output, given values for inputs) is given by ϕ^{-1}. Alternative specifications for returns-to-scale can be used by imposing additional constraints on λ; for example, a variable

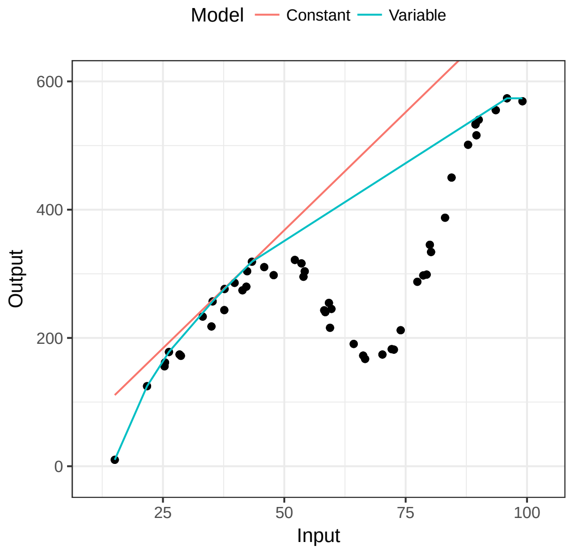

Figure 2: DEA Model Estimates

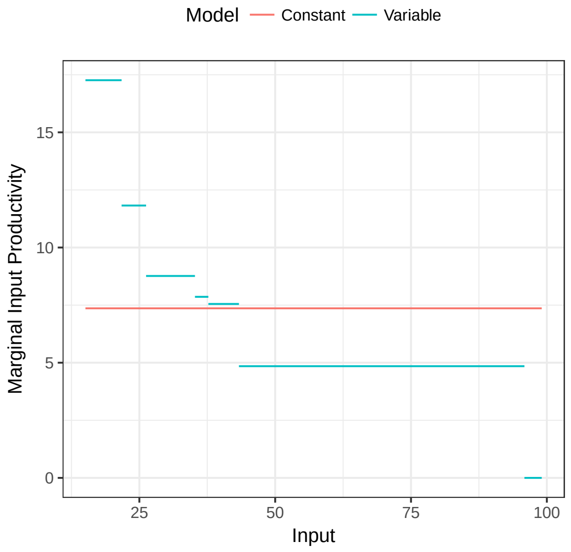

Figure 3: DEA Slope Estimates

returns-to-scale model can be used by constraining $\sum_i \lambda_i = 1$. Example input and output data and estimated DEA frontiers (with constant and variable returns-to-scale) using capital stock as the sole input and GDP as the output is shown in Figure 2.[1] Further, a plot of marginal input productivity (i.e., the slope of the production frontier) is shown in Figure 3.

DEA models have been used to identify a number of production characteristics. Notably, Färe et al. (1997) used DEA methods to separately identify productivity growth, technical progress and efficiency changes. Specifically, the authors found that efficiency change can be expressed as

$$\frac{\phi_{t+1}(x_{i,t+1}, y_{i,t+1})}{\phi_t(x_{i,t}, y_{i,t})} \tag{4}$$

and technical change can be written as

$$\left[\left(\frac{\phi_t(x_{i,t+1}, y_{i,t+1})}{\phi_{t+1}(x_{i,t+1}, y_{i,t+1}}\right)\left(\frac{\phi_t(x_{i,t}, y_{i,t})}{\phi_{t+1}(x_{i,t}, y_{i,t})}\right)\right]^{1/2}. \tag{5}$$

In these expressions, $\phi_t(x_{i,s}, y_{i,s})$ is the efficiency of firm i's production plan in year s relative to the production possibilities frontier in year t. Further, the total change in productivity can be expressed as the product of efficiency and technical changes.

As seen in Figure 2, DEA produces a piecewise-linear production possibilities frontier. While this characteristic is not problematic for many analyses, it does prevent direct estimation of allocative inefficiency from marginal rates of transformation because marginal input productivities are not defined at kinks in the frontier, as shown in Figure 3. For this reason, researchers have turned toward using costs or revenue in DEA analyses to estimate allocative inefficiency. However, as previously discussed, this approach assumes that firms are price-takers, and applying this method to non-competitive markets can produce biased estimates of allocative inefficiency. Instead, a smooth analogue of DEA can be used to obtain estimates of allocative inefficiency in a non-competitive environment in a approach similar to that of SFA, where marginal rates of transformation of real inputs to real outputs are estimated and compared to observed prices. Smooth non-parametric frontier analysis (SNFA) is one such method and is discussed further in the following subsection.

1.3 Smooth Non-Parametric Frontier Analysis

Smooth non-parametric frontier analysis uses kernel smoothing with additional bounding constraints to produce a smooth production possibilities frontier. Racine et al. (2009) developed a framework to estimate SNFAs and permitted the imposition of addi-

[1] This simple example is only meant to illustrate techniques and contrast different types of models. For a full empirical application, see the following section.

tional monotonicity and concavity constraints. First, the authors use a Nadaraya-Watson estimator for kernel smoothing, defined as

$$A(x, x_i) = \frac{K(x, x_i)}{\sum_{h=1}^{N} K(x, x_h)}, \tag{6}$$

where K is a generalized product kernel and observations of the independent variable x are indexed with subscripts. This paper uses a multivariate normal kernel for K; that is,

$$K(x, x_h) = \exp\left(-\frac{1}{2}(x - x_h)' H^{-1}(x - x_h)\right), \tag{7}$$

where H is a positive-definite bandwidth matrix chosen by the researcher. There are a number of methods that can be used to choose or select H, and bandwidth selection is outside of the scope of this research. Instead, this paper uses the rule-of-thumb

$$\begin{aligned} H_{jj}^{1/2} = &\left(\frac{4}{M+2}\right)^{1/(M+4)} \\ &\times N^{-1/(M+4)} \times \mathrm{sd}(x^j) \quad \text{for } j = 1, ..., M \\ &H_{ab} = 0 \quad \text{for } a \neq b, \end{aligned}$$

where $\mathrm{sd}(x^j)$ is the sample standard deviation of input j and N is the total number of observations used in estimation (Silverman, 1986). Then, the kernel smoothing estimate of y can be written as

$$\hat{y} = m(x) = \sum_{i=1}^{N} p_i A(x, x_i) y_i, \tag{8}$$

where y_i is the observed output of firm i and p_i are weights selected by the estimation procedure.[2]

It is well-known that kernel smoothing methods can suffer from bias near boundaries of the data due to asymmetric weighting (Silverman, 1986). This effect may be amplified in SNFA models because of the additional constraints placed on the fitted curve. There are a number of boundary correction methods available to reduce this bias, and this research utilizes the reflection method.

Optimal weights for Equation (8) can be found by solving the quadratic programming problem

$$\min_{p} \quad -\mathbf{1}'p + \frac{1}{2}p'p. \tag{9}$$

Additional constraints can then be imposed to ensure the kernel smoothing estimates envelop the data. Specifically, the following bounding constraints are imposed:

$$m(x_i) - y_i = \sum_{h=1}^{N} p_h A(x_i, x_h) y_h - y_i \geq 0 \quad \forall i. \tag{10}$$

Constrained quadratic programming problems such as the one above can be solved with many software packages. This research utilizes the `quadprog` package in the R programming language to solve these problems (Weingessel, 2013).

Racine et al. (2009) also describe constraints that can be imposed if the researcher assumes the frontier is monotonically increasing and/or concave. The authors suggest enforcing monotonicity by imposing

$$\nabla_x m(x) \cdot \mathbf{1} = \sum_{h=1}^{N} p_h \left(\nabla_x A(x, x_h) \cdot \mathbf{1}\right) y_h \geq 0 \quad \forall x. \tag{11}$$

In this expression, $\nabla_x m(x_i)$ is the gradient of m evaluated at x_i; that is,

$$\nabla_x m(x) = \left(\frac{\partial m(x)}{\partial x^1}, \frac{\partial m(x)}{\partial x^2}, ..., \frac{\partial m(x)}{\partial x^M}\right)'. \tag{12}$$

However, the constraint in Equation (11) only states that the sum of partial derivatives need be non-negative, which could result in estimates of individual marginal productivities that are negative. This research instead assumes all marginal productivities are non-negative; that is,

$$\frac{\partial m(x)}{\partial x^j} = \sum_{h=1}^{N} p_h \frac{\partial A(x, x_h)}{\partial x^j} y_h \geq 0 \quad \forall x, j. \tag{13}$$

Further, since concavity constraints expressed by in terms of second derivatives can produce quadratic programming problems that are computationally infeasible for large numbers of inputs, the authors use conditions for concavity given by Afriat (1967), expressed as

$$m(x) - m(z) \leq \nabla_x m(z) \cdot (x - z) \quad \forall x, z. \tag{14}$$

Note that Equation (13) and Equation (14) must hold for each point x and z in the domain for monotonicity and concavity to hold globally. In practice, however, it is not possible to impose these constraints over the entire domain. Fortunately, it is often sufficient to impose the constraints at a fixed number of points (possibly more than those used to fit the frontier). For this reason, the analysis in this paper

[2] In standard kernel smoothing, $p = \mathbf{1}$.

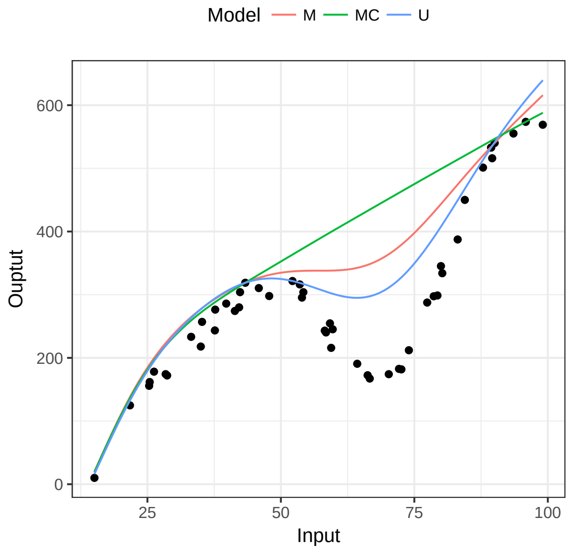

Figure 4: SNFA Model Estimates

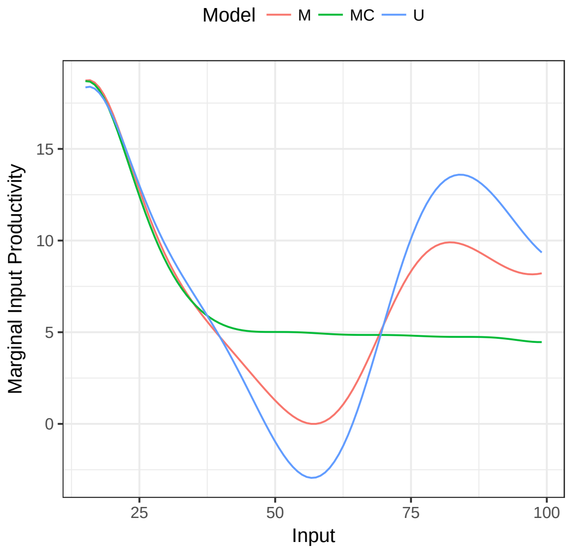

Figure 5: SNFA Slope Estimates

imposes constraints at additional points, described in detail in the following section.

Finally, marginal productivities of inputs at a point x are given by

$$\nabla_x m(x) = \sum_{i=1}^{N} \hat{p}_i \nabla_x A(x, x_i) y_i, \tag{15}$$

where $\hat{p}_i$ are estimated weights.

SNFA was run on the same data used in Figure 2, and results are plotted in Figure 4. Three separate specifications were used: An unconstrained model ("U"), a model that assumes the frontier is monotonically increasing ("M"), and a model that assumes the frontier is both monotonically increasing and concave ("MC"). The marginal input productivity is also plotted for each of the three specifications in Figure 5.

As can be seen in Figure 5, the slope of the frontier is well-defined at each point, unlike with the DEA model. Further, the slope of the frontier with monotonicity constraints is non-negative, and the slope of the frontier with monotonicity and concavity constraints is non-negative and (weakly) decreasing, as desired. As in DEA analysis, the efficiency of each data point can be calculated, so analysis similar to that conducted by Färe et al. (1997) can also be performed using SNFA analysis.

Though the analysis in this section was performed on bivariate data, it can be extended to include multiple inputs.[3] Since marginal productivities are defined and easily calculable for SNFA models, they can be used to assess allocative inefficiency (as well as technical inefficiency) in a non-parametric data envelopment framework. The following section describes an empirical SNFA analysis for estimating allocative inefficiency.

2 Empirical Model

The allocative inefficiency model presented in this section closely resembles that in Schmidt and Lovell (1979), with additional considerations given to address the non-parametric estimation of the production frontier. Begin by considering the set of production possibilities in year t, denoted

$$S_t = \{(x, y) | x \text{ can produce } y \text{ in year } t\}. \tag{16}$$

Note that the production possibilities set can change over time due to technical progress. If S_t is a closed set, the production possibilities frontier in year t can be expressed as

$$P_t = \{(x, y) \in S_t | \nexists\, \theta \in (0, 1) \text{ s.t. } (x, y/\theta) \in S_t\}. \tag{17}$$

[3] It is also possible to extend SNFA analysis to multiple outputs.

Then, there exists a production possibilities function f_t that satisfies

$$p_t(x) = y \text{ if and only if } (x, y) \in P_t. \qquad (18)$$

I assume that p_t is a continuous function that can be estimated with kernel smoothing methods. Note that p_t may exhibit other properties such as monotonicity or concavity.

Next, consider the observed input and output data for firms i in year t, denoted $\{(x_{it}, y_{it})\}_{i=1}^{N}$. Assume that there is a production function for firm i in year t such that

$$f_{it}(x_{it}) = y_{it}. \qquad (19)$$

I assume that firms' production functions are differentiated by a multiplicative efficiency factor ϕ_{it}, so that

$$p_t(x_{it}) = f_{it}(x_{it})/\phi_{it}. \qquad (20)$$

From this relation we can also define the efficiency function

$$\phi_{it}(x_{is}, y_{is}) = \frac{f_{is}(x_{is})}{p_t(x_{is})} = \frac{y_{is}}{p_t(x_{is})}. \qquad (21)$$

From the definition of the production possibilities function, we know that $\phi_{it}(x_{it}, y_{it}) \in (0, 1]$.[4]

Now denote the price of output for firm i in year t as p_{it} and the per unit cost of input j for firm i in year t as w_{it}^{j}. Then, a profit maximizing firm will allocate inputs so that

$$\frac{\partial f_{it}(x_{it})}{\partial x^j} = \frac{w_{it}^{j}}{p_{it}}. \qquad (22)$$

Using Equation (20), the left-hand side of this equation can be written as

$$\frac{\partial f_{it}(x_{it})}{\partial x^j} = \phi_{it}\frac{\partial p_t(x_{it})}{\partial x^j}, \qquad (23)$$

which is easily calculable since the slope of the production possibilities function is a result of SNFA analysis. Further, since input and output prices are observable, the validity of Equation (22) can be evaluated.

In practice, Equation (22) will never be exactly true and may not even be true in expectation, as in the case of systematic misallocation of inputs. To test for misallocation, consider the model

$$\left(\phi_{it}\frac{\partial p_t(x_{it})}{\partial x^j}\right) \Big/ \left(\frac{w_{it}^{j}}{p_{it}}\right) = \exp(\alpha_i^j + \varepsilon_{it}^{j}), \qquad (24)$$

where ε_{it}^{j} is a mean-zero normally distributed error. Using a log transformation, this model is equivalent to

$$\log\left(\phi_{it}\frac{\partial p_t(x_{it})}{\partial x^j}\right) - \log\left(\frac{w_{it}^{j}}{p_{it}}\right) = \alpha_i^j + \varepsilon_{it}^{j}. \qquad (25)$$

Under the assumption of a concave production function, $\alpha_i^j > 0$ unambiguously indicates systematic underallocation of input j by firm i and a sub-optimal realization of profit. Without the assumption of concavity, $\alpha_i^j > 0$ indicates profit can be increased using a greater quantity of x^j, but not necessarily that doing so would bring the firm closer to the globally optimal allocation.[5] This research will also investigate the following alternative specification

$$\log\left(\phi_{it}\frac{\partial p_t(x_{it})}{\partial x^j}\right) - \log\left(\frac{w_{it}^{j}}{p_{it}}\right) = \alpha_t^j + \varepsilon_{it}^{j}, \qquad (26)$$

where α_t^j measures systematic underallocation of input j by all firms in year t.

Both Equation (24) and Equation (26) represent a system of M equations, one for each input. Since each equation contains the same set of independent variables, the systems can be estimated via equation-by-equation OLS. Further, cross-equation correlations can be computed from error terms, which are of interest since they can reveal whether misallocation of one input is correlated with misallocations of other inputs.

3 Data

The previously described empirical analysis is applied to Class I freight railroads in the United States. This analysis considers the period of time after effects of the industry's partial deregulation had largely been realized. In 1980, there were 40 Class I railroads in operation; by 1999, mostly as a result of mergers and consolidation, there were just seven, as there are today. The output of freight railroads is measured by revenue-ton-miles, defined as one ton of product shipped one mile that generates revenue for the railroad.

[4] Note that $\phi_{it}(x_{is}, y_{is})$ may exceed 1 when $s \neq t$ due to shifts in the frontier over time.

[5] This characteristic is a result of the possibility that multiple production plans can satisfy first-order conditions for profit maximization when production functions are not necessarily concave.

References

Afriat, S. N. (1967). The construction of utility functions from expenditure data. *International economic review 8*(1), 67–77.

Aigner, D., C. K. Lovell, and P. Schmidt (1977). Formulation and estimation of stochastic frontier production function models. *Journal of econometrics 6*(1), 21–37.

Averch, H. and L. L. Johnson (1962). Behavior of the firm under regulatory constraint. *The American Economic Review 52*(5), 1052–1069.

Färe, R., S. Grosskopf, J. Logan, and C. K. Lovell (1985). Measuring efficiency in production: with an application to electric utilities. In *Managerial issues in productivity analysis*, pp. 185–214. Springer.

Färe, R., S. Grosskopf, and M. Norris (1997). Productivity growth, technical progress, and efficiency change in industrialized countries: reply. *The American Economic Review 87*(5), 1040–1044.

Farrell, M. J. (1957). The measurement of productive efficiency. *Journal of the Royal Statistical Society. Series A (General) 120*(3), 253–290.

Racine, J. S., C. F. Parmeter, and P. Du (2009). Constrained nonparametric kernel regression: Estimation and inference. In *Working paper*.

Schmidt, P. and C. K. Lovell (1979). Estimating technical and allocative inefficiency relative to stochastic production and cost frontiers. *Journal of econometrics 9*(3), 343–366.

Shepherd, R. W. (2015). *Theory of cost and production functions*. Princeton University Press.

Silverman, B. W. (1986). *Density estimation for statistics and data analysis*, Volume 26. CRC press.

Weingessel, A. (2013). *quadprog: Functions to solve Quadratic Programming Problems.*