ESSAYS ON

EFFICIENCY AND PRODUCTIVITY

Contributions on Bootstrap, DEA and Stochastic Frontier Models

by

Mickael Löthgren

A Dissertation for the
Degree of Doctor of Philosophy, Ph.D.

Stockholm School of Economics
Department of Economic Statistics
1997

STOCKHOLM SCHOOL
OF ECONOMICS
THE ECONOMIC RESEARCH INSTITUTE

ISBN 91-7258-441-6

Stockholm 1997

Acknowledgments

I am indebted to many people who, in one way or the other, have contributed towards the completion of this thesis. First of all I want to thank my advisor Anders Westlund who gave me the opportunity to start working in the area of efficiency and productivity analysis. He spent many hours reading several versions of often unstructured drafts and provided comments that helped me to focus on the important issues. I also want to thank Magnus Tambour, co-author of three papers in the thesis. The cooperation with Magnus offered many hours of hard and enjoyable work and our many stimulating discussions gave new insights and ideas that still remain to be explored. Sune Karlsson co-authored one paper and provided insightful comments and answers on my seemingly endless questions on econometrics and computers. Tor Jacobson read some of the later manuscripts and provided constructive and valuable comments. Mikael Gredenhoff and Michael Andersson have provided comments and remarks on some of the problems I have run into. Even more important is the fact that they have made the time at SSE both fun and enjoyable. I am also grateful to all other not mentioned members of the Department of Economic Statistic at Stockholm School of Economics that have provided comments in any way on my work. Monica Peijne and Carina Morton-Fincham have provided personal and friendly administrative assistance.

I also want to thank Pontus Roos for introducing me to the basics of DEA and distance functions and Rolf Färe and Shawna Grosskopf who kindly hosted me during my visit in the winter of 1995 to the Department of Economics, Southern Illinois University, Carbondale where they introduced me to parts of the axiomatic production theory.

I am grateful to Lars Sundin for the time and the large amount of red ink he spent when helping me improve the thesis summary language.

Finally, I want to thank Anna who has encouraged me to keep the work going and stimulated my mind with the necessary energy to complete this thesis. She has been more than patient with my many hours of late night work and my sometimes too absent mind. I hope that I can support you in the same way whenever you need it!

Financial support from the Swedish Research Council for the Humanities and Social Sciences (HSFR) is gratefully acknowledged.

Stockholm, February 1997

Mickael Löthgren

Thesis Summary

This thesis deals with estimation of production frontiers, technical efficiency and productivity for general multiple-input and multiple-output technologies. The production frontier is defined as either the maximum output obtainable from an input bundle (in an output based setting) or as the minimum input required to produce a given output level (in an input based setting). Technical efficiency is defined by the distance from the observed input-output vector to the frontier, using some distance measure. A commonly applied distance measure is the distance function introduced by Shephard (1953, 1970). The distance function is a scalar valued representation of the multiple-input, multiple-output technology. The function represents the radial distance from the observed output/input vector to the production frontier. This distance measure corresponds to the definition of radial technical efficiency introduced in Farrell (1957). The distance function furthermore provides the basis in the Malmquist productivity index introduced in the production context by Caves, Christensen, and Diewert (1982). The Malmquist productivity index can, following Färe, Grosskopf, and Roos (1995), be decomposed in two components: a technical change and a technical efficiency change component. The technical change component represents the movement of the production frontier over time, whereas the technical efficiency change component represents the change in the distance from the frontier.

Estimation of the production frontiers and the efficiency and productivity measures can be performed using basically two approaches: the nonparametric Data Envelopment Analysis (DEA) and the parametric, econometric, stochastic frontier production function approach.

DEA, introduced by Charnes, Cooper, and Rhodes (1978), is a non-parametric linear programming (LP) estimation method based on a piecewise linear envelopment of the observed input-output data. It can readily handle multiple-inputs and multiple-outputs, and point estimates of efficiency and productivity can easily be obtained using only input and output data without behavioral assumptions of cost minimization or profit maximization. The econometric approach, introduced by Aigner, Lovell, and Schmidt (1977) and Meeusen and van den Broeck (1977), is based on composed error regression models of parametric stochastic production frontiers.

Both DEA and the stochastic frontier models have undergone theoretical developments and have found extensive usage in various applications. See, e.g., Seiford (1996) and Greene (1993) for surveys of DEA and stochastic frontier models,

respectively. The two methods have, however, some drawbacks. The primary weakness with the DEA method is that it does not allow for statistical noise, and inference on the obtained estimates cannot be performed using standard methods. The stochastic frontier models incorporate random noise in the composed error term entering the regression models, but have on the other hand a major drawback in that the models only apply to single output (input) and multiple-input (output) production technologies. A recent issue of the *Journal of Productivity Analysis*, see Lovell, Olesen, and Petersen (1996), discusses these drawbacks and presents an agenda for future research within efficiency and productivity analysis. For DEA, the main future research issue is identified as the development of a stochastic DEA that allows for random noise and statistical inference on efficiency and productivity. For the stochastic frontier models, the primary future research issue is identified as the development of multiple-input and multiple-output models.

Included papers

This thesis contains six papers, listed below, proposing various extensions of DEA and stochastic frontier models well in line with the research agenda outlined in Lovell, Olesen, and Petersen (1996).

Paper 1 "Bootstrapping DEA-based Malmquist Index" and *Paper 2* "Scale Efficiency and Scale Elasticity in DEA-models - A Bootstrapping Approach" deal with the issue of statistical inference using DEA. These papers make use of the flexible computational intensive bootstrap resampling method, introduced by Efron (1982), to perform the desired inference. A bootstrap algorithm to obtain confidence intervals based on DEA estimates of different production characteristics is developed. The proposed algorithm is applied to data from Swedish eye-care service provision. In the first paper, the Malmquist productivity index and its components are estimated and tested. The main question is whether an obtained firm-specific productivity estimate indicates a significant progressed or regressed productivity. Similar tests are performed for the two productivity components to investigate whether a significant productivity change is due to a shift in the production technology or to changed technical efficiency.

In the second paper, scale properties in the Swedish health care production of eye care services in 1993 are studied. A new scale elasticity approximation is proposed and the bootstrap algorithm is used to determine the significance of the estimates of scale efficiency and scale elasticity.

Paper 3 "Computationally Efficient Double Bootstrap Variance Estimation"

focuses on the bootstrap method and develops an approach to obtain computationally efficient double bootstrap confidence intervals for an arbitrary parameter. The parameter could for instance be the productivity or the scale measure studied in *paper 1* and *2*. The double bootstrap confidence intervals have better coverage accuracy than the simple percentile intervals usually employed when constructing bootstrap confidence intervals. The proposed method is based on a first order expansion of the statistic and estimation of the first order terms in this expansion. These estimates are obtained from the bootstrap outer loop data and produce the variance estimator when weighted with the resampling frequencies in the double bootstrap sample. Small sample simulation studies indicate that the proposed variance estimator produces bootstrap-t confidence intervals with approximately the same coverage accuracy as bootstrap-t intervals obtained from an actual double bootstrap.

Paper 4 "Generalized Stochastic Frontier Production Models" and *Paper 5* "A Multiple Output Stochastic Ray Frontier Production Model" deal with the issue of developing multiple-input and multiple-output stochastic frontier models. A generalization of the standard single output stochastic frontier production model is proposed. The approach is based on the use of a polar coordinate representation of the output. A stochastic ray frontier production function is defined, where the norm of the output vector is specified as a function of the inputs and the output mix, represented by the multidimensional polar coordinate angles. The technical efficiency measure is represented by the radial distance from the observed output norm to the frontier norm. The approach allows econometric estimation of distance functions and technical efficiency for general multiple-input and multiple-output technologies. The approach offers an alternative to the standard dual frontier cost or profit function estimation usually applied to estimate efficiency in general multiple-input and multiple-output technologies.

Paper 5 extends the ideas in *Paper 4* and offers an empirical illustration of the stochastic ray frontier model. The model is extended further by incorporating the technical efficiency effects model by Battese and Coelli (1995). The technical efficiency is specified as a separate function of variables assumed to determine the level of efficiency. The model is applied to public Swedish health care data. A balanced panel data set of 26 county councils providing health care during the period $1989 - 1994$ is studied. The main question is whether an organizational, "internal market"-reform introduced by some councils during the observed time periods has had an effect on the production possibilities and the level of technical efficiency. The technical efficiency effects model allows identification and estima-

tion of the organizational effect on the production frontier and the efficiencies. The time varying pattern of technical efficiency and the technical change in the production frontier can also be identified and estimated.

Paper 6 "Productivity and Customer Satisfaction - A DEA Network Model" focuses on the problem of defining valid measures of efficiency and productivity. The main purpose of the paper is to define measures of efficiency and productivity that account for the customer-perceived quality of services and products offered by the firms under study. The approach is based on a separation of the production and consumption activity. A network technology is defined, in which the production and consumption are represented by two nodes in the network. The model allows separate allocation of the inputs to pure production and to customer oriented activities. An empirical application is included, where the network model is estimated using data from Swedish pharmacies. The derived network Malmquist productivity index, accounting for customer satisfaction, is compared with a more traditional Malmquist index.

References

AIGNER, D., C. A. K. LOVELL, AND P. SCHMIDT (1977): "Formulation and Estimation of Stochastic Frontier Production Function Models," *Journal of Econometrics*, 6, 21–37.

BATTESE, G. E., AND T. COELLI (1995): "A Model for Technical Inefficiency Effects in a Stochastic Frontier Production Function," *Empirical Economics*, 20, 325–332.

CAVES, D. W., L. R. CHRISTENSEN, AND W. E. DIEWERT (1982): "The Economic Theory of Index Numbers and the Measurement of Input and Output and Productivity," *Econometrica*, 50.

CHARNES, A., W. W. COOPER, AND E. RHODES (1978): "Measuring the Efficiency of Decision Making Units," *European Journal of Operational Research*, 2, 429 – 444.

EFRON, B. (1982): *The Jackknife, the Bootstrap and other Resampling Plans*, CBMS-NSF Regional Conference Series in Applied Mathematics. Society for Industrial and Applied Mathematics.

FÄRE, R., S. GROSSKOPF, AND P. ROOS (1995): "Productivity and Quality Changes in Swedish Pharmacies," *The International Journal of Production Economics*, 29, 137–144.

FARRELL, M. S. (1957): "The Measurement of Productive Efficiency," *Journal of the Royal Statistical Society, Series A*, 120, 253–281.

GREENE, W. H. (1993): "The Econometric Approach to Efficiency Analysis," in H. O. Fried, C. A. K. Lovell and S. S. Schmidt (eds.) The Measurement of Productive Efficiency, Oxford University Press, New York.

LOVELL, C. A. K., O. B. OLESEN, AND N. C. PETERSEN (1996): "Editor's Introduction," *Journal of Productivty Analysis*, 7, 87–98.

MEEUSEN, W., AND J. VAN DEN BROECK (1977): "Efficiency Estimation from Cobb-Douglas Production Functions with Composed Error," *International Economic Review*, 18, 435–444.

SEIFORD, L. M. (1996): "Data Envelopment Analysis: The Evolution of the State of the Art (1978-1995)," *Journal of Productivity Analysis*, 7, 99–137.

SHEPHARD, R. W. (1953): *Cost and Production Functions*, vol. 194 of *Lecture Notes in Economics and Mathematical Systems*. Springer-Verlag. Originally published by Princeton University Press 1953.

——— (1970): *The Theory of Cost and Production Functions*. Princeton University Press.

List of Papers

I. Löthgren, M. and M. Tambour (1995), "Bootstrapping DEA-based Malmquist Index", Working Paper Series in Economics and Finance, No. 78, Stockholm School of Economics. Revised version December 1996.

Resubmitted to *Empirical Economics.*

II. Löthgren, M. and M. Tambour (1996), "Scale efficiency and Scale Elasticity in DEA-models - A Bootstrapping Approach", Working Paper Series in Economics and Finance, No. 91, Stockholm School of Economics. Revised version December 1996.

Forthcoming in *Journal of Productivity Analysis.*

III. Karlsson, S. and M. Löthgren (1997), "Computationally Efficient Double Bootstrap Variance Estimation", Working Paper Series in Economics and Finance, No. 151, Stockholm School of Economics.

Submitted to *Computational Statistics & Data Analysis.*

IV. Löthgren, M. (1996), "Generalized Stochastic Frontier Production Models", Working Paper Series in Economics and Finance, No. 149, Stockholm School of Economics.

Submitted to *Economics Letters.*

V. Löthgren, M. (1997), "A Multiple Output Stochastic Ray Frontier Production Model", Working Paper Series in Economics and Finance, No. 158, Stockholm School of Economics.

Submitted to *Journal of Productivity Analysis.*

VI. Löthgren, M. and M. Tambour (1996), "Productivity and Customer Satisfaction - A DEA Network Model", Working Paper Series in Economics and Finance, No. 140, Stockholm School of Economics.

Submitted to *European Journal of Operational Research.*

I

Bootstrapping the DEA-based Malmquist Index

Mickael Löthgren* and Magnus Tambour†
Stockholm School of Economics

Working Paper Series in Economics and Finance No. 78
October 1995
Revised and Retitled December 1996

Abstract

This paper presents a bootstrap approach to calculate confidence intervals for firm specific Malmquist productivity indices obtained from data envelopment analysis (DEA) models. The bootstrap is easily implemented and allows identification of production units that have significant productivity changes. An application using data from Swedish eye-care departments is included. We find that 40% of the departments have significant progress in productivity whereas only 10% of the departments have a significant regress in productivity.

Keywords: Bootstrap, Data Envelopment Analysis, Distance Function, Malmquist Productivity Index.

JEL-Classification: C14, C15, D24, H42.

Acknowledgements: We thank Sune Karlsson and Anders Westlund for helpful comments and the Swedish Ophthalmology Society and SPRI for providing data. Financial support to M. Tambour from the National Corporation of Swedish Pharmacies is greatfully acknowledged.

The paper was presented at the third European Conference on Health Economics, Stockholm, August, 20 - 22, 1995.

*Corresponding author. Department of Economic Statistics, Stockholm School of Economics, P.O. Box 6501, S-113 83 Stockholm, Sweden. *E-mail*: stml@hhs.se. *Fax*: + 46-8-34 81 61.
†Centre for Health Economics.

1. Introduction

One methodological approach in efficiency and productivity measurement that has attracted much attention since the late 1970s is the non-parametric linear programming (LP) approach, dubbed data envelopment analysis (DEA) by Charnes, Cooper and Rhodes (1978). This approach is based on estimation of technology frontiers against which firms (or other decision making units) are evaluated. An appealing property of the DEA-approach is that multiple-input, multiple-output technologies can be modeled without behavioral assumptions and cost data, contrary to the standard cost-function based approach. This is important if price or cost data are unavailable as is often the case in the health care sector or other public production.

An extension of the single period DEA-models is the estimation of productivity growth by Malmquist indices (Atkinson and Wilson (1995) list a number of recent studies). A weakness with the standard DEA model is that it does not incorporate any random noise. The method only gives point estimates of efficiency and productivity that do not offer any information of the uncertainty in the firm specific estimates. One approach to remedy this shortcoming is to apply bootstrap techniques in order to obtain measures of statistical precision in the estimates. The bootstrap can be implemented in various ways and earlier studies have considered somewhat different questions. Firm-specific efficiency scores (or distance functions) are considered in Ferrier and Hirschberg (1994), Gstach (1995), Wilson and Simar (1995) and Simar and Wilson (1995), while Atkinson and Wilson (1995) bootstrap sample-averages of efficiency and Malmquist indices.

The purpose of this paper is to extend the literature by presenting a bootstrap method for calculation of confidence intervals for firm-specific DEA-based Malmquist productivity indices. Our model of the data generating process underlying the bootstrap states that the observed data are generated as random radial (inefficiency) deviations off the production frontier. Earlier studies have concentrated on the stochastic nature of the production frontier in the bootstrap resampling. Pseudo data have been generated randomly to resample the frontier estimation. The bootstrap estimates of the firm specific distance functions have been conditioned on the original data of the inputs and outputs. We extend this idea and specify the DEA-estimates of the distance functions completely in terms of the pseudo data. Hence, the bootstrap production frontier, the distance function and the Malmquist index are completely based on the pseudo data.

The bootstrap method allows identification of firms with significant productivity changes. Using this information, managers can focus their efforts to increase productivity where improvements are most needed. Furthermore, firms with significant positive productivity change can be identified and constitute benchmarks

against which other firms can be compared and evaluated.

The proposed bootstrap method is applied on data from a sample of Swedish eye-care departments in 1992 and 1993. One conclusion from the original results is that about half the sample of the departments had a positive change in productivity and the other half had a negative change. The bootstrap results show that about 40% of the departments had a significant positive productivity growth, while only 10% of the departments had a significant regress in productivity. Thus, for almost half of the observations the sign of the productivity change is insignificant.

The remainder of the paper is organized as follows: Section 2 contains a presentation of the DEA-model and the Malmquist productivity index. Section 3 follows with a description of the bootstrap procedure. The construction of confidence intervals is outlined in section 4. Section 5 presents the empirical application and Section 6 concludes the paper with a summary.

2. Production technology and DEA

Consider a sample of K firms using $x^t \in R_+^N$ inputs in the production of $y^t \in R_+^M$ outputs in time period $t = 1, ..., T$. A multiple-input, multiple-output production technology can be represented by the output set, defined as

$$P^t\left(x^t\right) = \left\{y^t : y^t \text{ can be produced by } x^t \text{ at time } t\right\}, t = 1, ..., T. \tag{2.1}$$

In an output-based approach, the production technology is completely characterized by the output distance function (Shephard (1970)), defined as

$$D_o^t(y, x) = \min\left\{\nu \in (0, 1] : \frac{y}{\nu} \in P^t(x)\right\}, t = 1, ..., T. \tag{2.2}$$

The distance function is less than, or equal to one if and only if the output y belongs to the output set $P(x)$. In an input-based approach the input distance function can represent the technology . It is defined as $D_i(x, y) = max\left\{\lambda \geq 1 : \frac{x}{\lambda} \in L(y)\right\}$, where $L(y) = \{x : x \text{ can produce } y\}$ is the input requirement set. In both approaches a firm is considered as technically efficient if the distance function equals one. An output-based approach is used in this paper because it is most reasonable for the empirical application where the hospital departments receive a (fixed) budget. The output distance function is illustrated in Figure 2.1.

Figure 2.1 displays the output set for a technology with two outputs. The firm with output y is inefficient since it is possible to scale the output vector

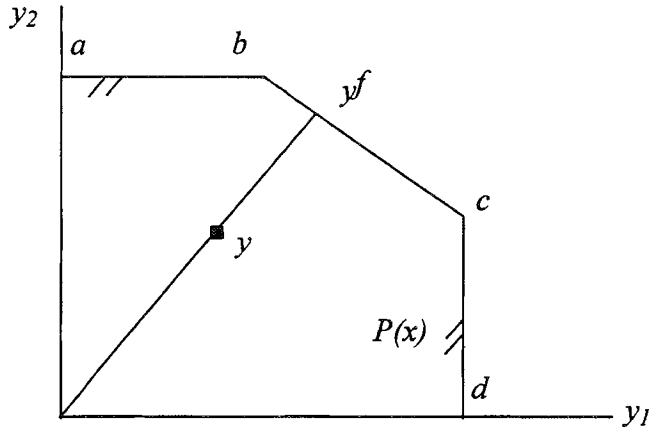

Figure 2.1: The output set with two outputs.

by a number larger than one, and the transformed output vector is still feasible. The value of the distance function is given by $D_o^t(y,x) = \|y\| / \left\|y^f\right\| < 1$, where $y^f \in IsoqP^t(x) = \{y : y \in P^t(x), \mu y \notin P^t(x), \mu > 1\}$ is the frontier output. The isoquant, $IsoqP^t(x)$, is given by the line segment $a-b-c-d$.

Following Färe *et al.* (1989) the distance function can be estimated using a piecewise linear reference technology for firm k as the solution to the LP problem

$$\left[\widehat{D}_o^t\left(y_k^t, x_k^t\right)\right]^{-1} = \max_z \left\{\theta \in R_{++} : \theta y_k^t \leq Y^t z, x_k^t \geq X^t z, z \in R_+^K\right\}, k = 1, ..., K, t = 1, ..., T, \tag{2.3}$$

where y_k^t is an M-vector of outputs, x_k^t is an N-vector of inputs, Y^t is a $(M \times K)$ matrix of outputs, X^t is a $(N \times K)$ matrix of inputs and z is a K-vector of non-negative intensity variables. Note that the solution of (2.3) always satisfies $\widehat{D}_o^t(y_k^t, x_k^t) \leq 1$, since firm k is a member of the reference technology.

2.1. The Malmquist productivity index

Caves, Christensen and Diewert (1982) showed how distance functions can be used to define Malmquist indices of productivity change. Färe *et al.* (1989) further extended this work by showing how this index can be estimated by non-parametric LP-models. We follow Färe *et al.* (1989) and define the output-based Malmquist productivity index between time period t and $t+1$ as a geometric mean of the indices proposed by Caves, Christensen and Diewert (1982)

$$M_o^{t,t+1}\left(y^t,y^{t+1},x^t,x^{t+1}\right)=\left[\frac{D_o^t\left(y^{t+1},x^{t+1}\right)}{D_o^t\left(y^t,x^t\right)}\frac{D_o^{t+1}\left(y^{t+1},x^{t+1}\right)}{D_o^{t+1}\left(y^t,x^t\right)}\right]^{\frac{1}{2}},t=1,...,T-1. \tag{2.4}$$

Färe *et al.* (1989) showed that the index can be decomposed into two components as

$$M_o^{t,t+1}(\cdot)=\underbrace{\frac{D_o^{t+1}\left(y^{t+1},x^{t+1}\right)}{D_o^t\left(y^t,x^t\right)}}_{E_o^{t,t+1}}\underbrace{\left[\frac{D_o^t\left(y^t,x^t\right)}{D_o^{t+1}\left(y^t,x^t\right)}\frac{D_o^t\left(y^{t+1},x^{t+1}\right)}{D_o^{t+1}\left(y^{t+1},x^{t+1}\right)}\right]^{\frac{1}{2}}}_{TC_o^{t,t+1}},t=1,...,T-1. \tag{2.5}$$

The term outside the brackets ($E_o^{t,t+1}$) is a ratio of two distance functions, which measures the change in efficiency between period t and $t+1$ as a "catching up to the frontier" effect. The next term ($TC_o^{t,t+1}$) measures the change in the production technology as a shift in the production frontier. Improvement in productivity, as well as improvement in efficiency and technology, is indicated by values greater than one, whereas values less than one indicates regress. The Malmquist productivity index can be interpreted as a measure of total factor productivity growth. This is most easily seen within a single-input single-output technology where the Malmquist index (given constant returns to scale technology) simplifies to the average product ratio $M_o^{t,t+1}=\dfrac{y^{t+1}/x^{t+1}}{y^t/x^t}$.

To estimate the Malmquist index (2.4) and the decomposition in (2.5), four separate LP-problems have to be solved. The first two single period distance functions are solved using the formulation in (2.3). The other two distance functions, the cross-period distance functions, are based on observations from two periods. In those two cases, observations in period t are evaluated against the reference technology in period $t+1$, and vice versa. The distance function $D_o^t\left(y^{t+1},x^{t+1}\right)$, where input and output observations from period $t+1$ are evaluated relative to the technology in period t, is estimated by solving the following LP-problem

$$\left[\widehat{D}_o^t\left(y_k^{t+1},x_k^{t+1}\right)\right]^{-1}=\max_z\left\{\theta:\theta y_k^{t+1}\leq Y^tz,x_k^{t+1}\geq X^tz,z\in R_+^K\right\},k=1,...,K. \tag{2.6}$$

Note that in the case of progressive technical change, y^{t+1} does not necessarily belong to $P^t\left(x^t\right)$. Hence, the cross-period distance function can take on values both greater or less than one. The other cross-period distance function estimate, $\widehat{D}_o^{t+1}\left(y^t,x^t\right)$, is obtained by solving a similar LP-problem as in (2.6), with switched time superscripts t and $t+1$.

3. The Bootstrap Method

The bootstrap method is a well established computationally intensive statistical resampling method used to perform inference in complex problems, see Efron and Tibshirani (1993) for a presentation of the method. Due to the complexity of the DEA-estimator, no analytical solutions of the bootstrap are generally available. The basic idea of the bootstrap method is to approximate the sampling distributions of the DEA estimators by Monte Carlo simulation where repeated resamples of the observed data produce repeated estimates. The empirical distributions of the simulated estimates approximates the sampling distributions of the estimators. The resamples are generated using an estimate of the data generating process (DGP).

The DGP underlying the bootstrap performed in this paper is output-oriented. The outputs are given by random radial deviations off the frontier, i.e., the isoquant of the output set. Formally, the input-output observations are given by

$$\left(x_k^t, y_k^t\right) = \left(x_k^t, D_{o,k}^t y_k^{f,t}\right), k = 1, ..., K, \tag{3.1}$$

where $y_k^{f,t} \in IsoqP^t\left(x_k^t\right)$ is the unobservable frontier output for the k:th firm in time period t. The distance functions are assumed to be drawn from the same distribution, i.e., $D_{o,k}^t \sim F_{D_o}^t$, $k = 1, ..., K$, $t = 1, ..., T$, where $F_{D_o}^t$ denotes the time-specific common distribution.

This DGP-model represents the idea that, conditioned on the inputs and the output proportions, the stochastic elements in the production process are completely represented by the random output efficiency measures.

The main idea in the simulation is to mimic the DGP. The procedure in each resample is as follows: Conditioned on observed inputs and output proportions, the resample data are constructed in two steps. For each time period, the frontier outputs are estimated and bootstrap pseudo-outputs are generated by replicating the DGP in (3.1) using the estimated frontier outputs and pseudo-distance functions drawn from some estimate of the distribution $F_{D_o}^t$. In this paper we use the non-smoothed and non-parametric bootstrap where the empirical distribution of the estimated distance functions is used to generate the pseudo-efficiencies.

3.1. The bootstrap Monte Carlo algorithm

The algorithm is given by the following steps:

1. Let the transformed input-output vectors be given by

$$\left(x_k^t, \widehat{y}_k^{f,t}\right) = \left(x_k^t, \left(\widehat{D}_{o,k}^t\right)^{-1} y_k^t\right), k = 1, ..., K, \tag{3.2}$$

where the transformed output vectors are efficient in the sense that $\widehat{y}_k^{f,t} = \left(\widehat{D}_{o,k}^t\right)^{-1} y_k^t \in Isoq\widehat{P}^t(x_k) \quad \Leftrightarrow \quad \widehat{D}_o^t\left(\widehat{y}_k^{f,t}, x_k^t\right) = 1.$

2. Resample, with replacement, K distance functions from $\widehat{D}_o^t(\cdot) = \left(\widehat{D}_{o,1}^t(\cdot), \ldots, \widehat{D}_{o,K}^t(\cdot)\right)$. Let the vector $\delta^{*t} = (\delta_1^{*t}, \ldots, \delta_K^{*t})$ denote the K resampled distance functions and let $\mathcal{I}^{*t}$ denote a K-vector of resampled firm-indices.

3. Let the bootstrap pseudo-cases, for each time period, be given by
$$\left(x_k^{*t}, y_k^{*t}\right) = \left(x_k^t, \delta_k^{*t}\widehat{y}_k^{f,t}\right), k = 1, ..., K. \tag{3.3}$$

4. Estimate the bootstrap output distance functions as the solutions to LP-problems similar to (2.3) as
$$\left[\widehat{D}_o^{*t}\left(y_k^{*t}, x_k^t\right)\right]^{-1} = \max\left\{\theta : \theta y_k^{*t} \leq Y^{*t}z, x_k^t \geq X^t z, z \in R_+^K\right\}, k = 1, ..., K, t = 1, ..., T. \tag{3.4}$$
The bootstrap cross-period distance functions $\widehat{D}_o^{*t}\left(y_k^{*t+1}, x_k^{t+1}\right)$ are obtained from (2.6), where the distance functions for the period $t+1$-pseudo-cases $\left(y_k^{*t+1}, x_k^{t+1}\right)$ are estimated relative to the output set constructed from period t bootstrap cases (y_k^{*t}, x_k^t) . The other cross-period distance functions $\widehat{D}_o^{*t+1}\left(y_k^{*t}, x_k^t\right)$ are obtained analogously.

5. Repeat steps (2) - (4) B times to create a set of B K-vectors $\widehat{D}_o^{*b,t} = \left(\widehat{D}_{o,1}^{*b,t}, \ldots, \widehat{D}_{o,K}^{*b,t}\right)$, $b = 1, ..., B$. The number of bootstrap replications is set to $B = 1000$. Efron and Tibshirani (1993), p. 275, recommend this size of B in order to make the variability of the boundaries of the confidence intervals constructed from the bootstrap "acceptably" low.

The bootstrapped firm-specific Malmquist output-based productivity index, and the decomposition of the index, is obtained as (2.5)

$$\widehat{M}_o^{*t,t+1}(\cdot) = \underbrace{\frac{\widehat{D}_o^{*t+1}\left(y^{*t+1}, x^{t+1}\right)}{\widehat{D}_o^{*t}\left(y^{*t}, x^t\right)}}_{\widehat{E}_o^{*t,t+1}} \underbrace{\left[\frac{\widehat{D}_o^{*t}\left(y^{*t}, x^t\right)}{\widehat{D}_o^{*t+1}\left(y^{*t}, x^t\right)} \frac{\widehat{D}_o^{*t}\left(y^{*t+1}, x^{t+1}\right)}{\widehat{D}_o^{*t+1}\left(y^{*t+1}, x^{t+1}\right)}\right]^{\frac{1}{2}}}_{\widehat{TC}_o^{*t,t+1}}, t = 1, ..., T-1. \tag{3.5}$$

In order to keep the firm-specific dynamic structure of the productivity over time, the same firms are used for the two time periods, i.e., $\mathcal{I}^{*b,t} = \mathcal{I}^{*b,t+1} = \mathcal{I}^{*b}$, for each bootstrap iteration $b = 1, ..., B$.

3.2. Some remarks on the bootstrap procedure

The DGP in (3.1) underlying the bootstrap is similar to earlier work on bootstrapping of DEA-models. Simar and Wilson (1995), Wilson and Simar (1995) as well as some other work cited in Grosskopf (1996) specify this DGP.

Given the DGP, a fundamental issue is whether to condition the bootstrap distance functions on the original data or not. Our approach is based on the resampled data for the bootstrap estimation of both the frontiers and the distance functions. Most earlier work use the resampled data to estimate the bootstrap frontiers but condition the bootstrap distance function estimates on the original observations. The key difference is in step 4 in the algorithm, where a conditioned bootstrap distance function estimate is obtained by substituting the bootstrap output data y_k^{*b} by the original output y_k.

We motivate our resampling approach by the fact that the sampling distribution of the DEA estimator is a function of realizations of two stochastic events. First, the observations are generated by a stochastic mechanism as specified in the DGP. Second, the distance function estimators are functions of the frontier estimates. A valid resampling procedure should take this double stochastic feature into consideration. In our resampling algorithm the bootstrap frontier estimates and the bootstrap distance function estimates are based on the resampled data in the same manner as the original estimates are based on the original data. In this way both the stochastic DGP and the sampling distribution of the frontier estimates will be incorporated in the bootstrap distance function estimates in (3.4). The empirical distributions of the bootstrap distance functions approximate the common distribution of the efficiencies, conditional on the inputs and the output proportions.

The DGP in (3.1) models the production frontier as deterministic in the sense that the frontier is fixed and not affected by random events. The stochastic element is the random distance functions generating inefficient observations off this frontier. Since the distance functions are truncated above at one the DGP specifies that the observed outputs always lie within the output set. Our approach guarantees that this condition is fulfilled in the sense that $\widehat{D}_o^{*b} \leq 1, \forall b$. This property is not necessarily satisfied if the bootstrap distance function estimates are conditioned on the original observations as in Simar and Wilson (1995) since in their approach it is possible that $\widehat{D}_o^{*b} > 1$ for some b.

We choose, by arguments of simplicity, to use the simple non-parametric non-smoothed resampling scheme. If the bootstrap efficiency estimates are conditioned on the original observations, Wilson and Simar (1995) note that the simple bootstrap gives inconsistent estimates of the sampling distribution of the distance function estimates. This inconsistency argument cannot, however, be raised against the unconditional bootstrap procedure in this paper since our approach is not

conditioned on the original observations. Simar and Wilson propose the use of a smoothed bootstrap to obtain a consistent procedure. The proposed smoothing procedure can of course be incorporated into our bootstrap algorithm as well, although at the cost of increased complexity.

4. Confidence Intervals

A number of alternatives to calculate confidence intervals are available. The presentation that follows is explicitly concerned with construction of confidence intervals for the Malmquist index. Intervals for the efficiency change term ($E_o^{t,t+1}$) and the technical change term ($TC_o^{t,t+1}$) are calculated analogously.

The percentile method, see e.g., Efron and Tibshirani (1993), is the most straightforward method to obtain bootstrap confidence intervals. This type of intervals is used in Ferrier and Hirschberg (1994), where firm-specific efficiencies are bootstrapped. The percentile method is based on the empirical distribution function of $\widehat{M}_{o,k}^{*b}$, $b = 1, ..., B$, defined as $\widehat{G}_k(s) = \frac{1}{B}\sum_{b=1}^{B} I\left(\widehat{M}_{o,k}^{*b} \leq s\right)$, for any real value s, where $I(\cdot)$ denotes the indicator function. A $(1-2\alpha)$ equal-tail confidence interval for the true Malmquist index $M_{o,k}$ is given by the interval

$$\left(\widehat{M}_{o,k}^{*(\alpha)}, \widehat{M}_{o,k}^{*(1-\alpha)}\right), \tag{4.1}$$

where $\widehat{M}_{o,k}^{*(\alpha)}$ is the α:th quantile of $\widehat{G}_k$, i.e., $\widehat{M}_{o,k}^{*(\alpha)} = \widehat{G}_k^{-1}(\alpha)$. The quantiles of $\widehat{G}_k$are given by the $[(B+1)\alpha]$:th and the $[(B+1)(1-\alpha)]$:th ordered values of $\widehat{M}_{o,k}^{*b}$, $b = 1, ..., B$, respectively, where $[r]$ denotes the integer part of any real value r.

Simar and Wilson (1995) present a simple and direct approach to bias correct the percentile intervals in (4.1) using a simple additive bias correction. The bias of the DEA estimator is defined as $\widehat{bias}_k = E\left(\widehat{M}_{o,k}\right) - M_{o,k}$, where the expectation $E(\cdot)$ is taken under the specified DGP. The bootstrap estimate of this bias is given by

$$\widehat{bias}_k^* = \frac{1}{B}\sum_{b=1}^{B} \widehat{M}_{o,k}^{*b} - \widehat{M}_{o,k}. \tag{4.2}$$

The bias corrected $(1-2\alpha)$ equal-tail intervals are simply obtained by shifting the bounds in the intervals in (4.1) by the factor $2 \cdot \widehat{bias}_k^*$ as

$$\left(\widehat{M}_{o,k}^{*(\alpha)} - 2 \cdot \widehat{bias}_k^*, \widehat{M}_{o,k}^{*(\alpha)} - 2 \cdot \widehat{bias}_k^*\right). \tag{4.3}$$

Simar and Wilson (1995) motivate the correction of $2 \cdot \widehat{bias}_k^*$ by the fact that this correction centers the empirical bootstrap distribution on the bias corrected estimate $\widetilde{M}_{o,k} = \widehat{M}_{o,k} - \widehat{bias}_k^*$. Shifting the intervals in (4.3) by a factor of

$1 \cdot \widehat{bias}_k^*$ will center the intervals on the biased $\widehat{M}_{o,k}$. Simar and Wilson express their intervals in terms of percentile intervals for the corrected bootstrap estimate $\widehat{M}_{o,k}^{*b} = \widehat{M}_{o,k}^{*b} - 2 \cdot \widehat{bias}_k^*$. Since the bias correction term is the same and constant for each bootstrap replication b, this obviously results in intervals identical to (4.3).

Remark: The original estimates are ignored in the percentile intervals since this type of intervals are based only on the resampled estimates. The firm-specific original estimates are however taken into account in the bias-corrected intervals in the bias estimate (4.2).

It can be noted that it is possible that the original estimates may lie outside the bounds of the confidence intervals. This is most likely to occur for estimates located in the tail of skewed empirical distributions of the original estimates. This is not inherent to our resampling algorithm. It may also occur using the conditional resampling in Simar and Wilson (1995).

5. An Empirical Application

To illustrate the method we apply the bootstrap procedure on data from 29 Swedish (public) eye-care departments in 1992 and 1993. All data are obtained from a database compiled by SPRI (the Swedish Institute for Health Services Development) in collaboration with the Swedish Ophthalmology Society. Three input and four output variables are chosen to represent the eye-care production technology. While these variables are perhaps not the most ideal representation of the eye-care production technology, we believe that they are sufficient to illustrate the method. Descriptive statistics of the inputs and outputs are shown in Table 1

TABLE 1 IN HERE

The input variables contain two measures on labor input, full time equivalent (FTE) months for two physician categories, specialized and other physicians. Unfortunately, FTE input for other labor categories are not included in the surveys. Good measures on capital use in health care provision are (at least in Sweden) notoriously difficult to obtain. Although one could argue that capital costs amount to a small fraction of the total cost in health care provision we choose not to ignore it. The number of available beds is included as a crude proxy for capital input. To some extent this input variable also reflects resource use in that beds have been used as an instrument to determine the size of the budgets that are allocated within the county councils that provide almost all health care in Sweden.

As output variables we use three common eye-care procedures (cataract, glaucoma and squint surgery) and number of visits. Cataract surgery is the far most

common procedure in ophthalmology. It has increased at a substantial rate the last decade, and is now the most common surgical procedure in Swedish health care. Since data on number of operation procedures are available, we have chosen not to include the more commonly used number of bed-days or discharges as output variables.

5.1. Results

Bootstrap confidence intervals for the Malmquist index together with the decomposition and the original estimates for the 29 departments are displayed in Table 2. The results are obtained using PC versions of Gauss and Lingo.

TABLE 2 IN HERE

One conclusion from the original results is that about half of the departments had a positive change and the other half had a negative change in productivity. The bias-corrected bootstrap results show that the productivity growth was significant for ten departments, although only three of the fourteen cases with an estimated negative productivity change in the original model are significant. There are thus several cases where the positive and negative changes are not significant. The percentile method identifies six departments with a significant positive change in productivity and one department with a significant negative productivity change. Hence, using the percentile method, there are 22 cases where the sign of the productivity change estimates is not significant and hence cannot be determined. The bias-corrected method, on the other hand, gives somewhat more "detailed" information. In a number of cases the significant results coincide between the percentile and the bias-corrected methods.

Differences in conclusions between the original and the bootstrap results can be seen by inspecting the results for department no. 13 and 29, for example. From the original results we would conclude that department 13 had a positive change in productivity, whereas department 29 had a decline in productivity. However, the conclusion from the bootstrap results is that neither of these two departments had a significant change in productivity. This shows that the productivity estimates from the original model are sensitive to sampling variation and comparisons of production units, based on the original point estimates should therefore be made with caution.

Turning to the efficiency change component, 13 of the 29 departments experienced a regress in efficiency, nine departments improved their efficiency and seven departments had no change in efficiency. Although no significant efficiency

changes are found using the percentile method, the bias-corrected intervals, however, show significant negative change in efficiency for six departments and significant improved efficiency for three departments. The simple percentile method, which is used by, e.g., Ferrier and Hirschberg (1994), thus offers little information for the decision maker in this case. The conclusion based on the percentile results is that all departments can have had either a positive or a negative change in efficiency. As an example it can be noted that the percentile method does not identify department 19 with a significant improvement in efficiency although the original estimate was 1.23. The bias-corrected method on the other hand, indicates that the change in efficiency was significant and range from 1.27 to 1.74.

Finally, for the technical change component a positive shift in technology is obtained for 19 departments in the original results of which the percentile method recognize nine significant cases. No significant negative changes are found using the percentile method. The bias-corrected intervals show significant positive changes in technology for nine departments and significant negative changes for three departments.

6. Summary and conclusions

In this paper we apply the non-parametric bootstrap method on a standard DEA-model. The approach is easy to implement and makes it possible to compute confidence intervals for firm-specific Malmquist productivity indices.

The bootstrap procedure is applied on data from Swedish eye-care departments. We note that the introduction of a stochastic element into the DEA-model, for this specific dataset, leads to some changes in conclusions relative to the original results. The original results show that about half the sample of the departments had a positive change in productivity and the other half had a negative change. The bootstrap results show that about 40% of the departments had a significant positive productivity growth, while only 10% of the departments had a significant regress in productivity. Thus for almost half of the observations the sign of the productivity change is insignificant.

The productivity change can be decomposed in two components. For the first component, change in efficiency, almost half of the departments experienced a regress according to the original results. Several of these cases are also significant in the bias-corrected bootstrap results, although no significant negative results were found using the percentile method. The second component, technical change, contributed positively to the change in productivity for two thirds of the departments in the original results and almost half of those cases are also significant positive according to the bias-corrected bootstrap confidence intervals.

The main point to note is that the bootstrap method allows identification of

firms with significant changes in productivity. This information is important from a management perspective since efforts can be focused on increasing productivity where improvements are most needed. Furthermore, firms with significant positive productivity change can be identified as a benchmark group against which other firms can be compared and evaluated.

Bootstrapping of DEA-models is still under development and some important properties of the bootstrap remains to be explored. It is not clear, for example, which resampling approach has the best coverage accuracy of the obtained bootstrap confidence intervals. Should a conditional or unconditional, smoothed or nonsmoothed, approach be used? This question must also consider the possible trade-off between accuracy and complexity of the application of the bootstrap. Furthermore, the importance of the number of replicates in the Monte Carlo algorithm remains to be established. The usually applied number of 1000 replicates are based on general recommendations for bootstrap confidence intervals. It is not clear if this is a sufficient number of replications for bootstrapping the complex DEA estimators. However, we leave these issues to future research.

7. References

Atkinson, S. E. and Wilson, P. W., (1995), "Comparing Mean Efficiency and Productivity Scores from Small Samples: A Bootstrap Methodology", *Journal of Productivity Analysis*, Vol. 6, 137-152 .

Caves, D., Christensen, L. and Diewert, E. W., (1982), "The Economic Theory of Index Numbers and the Measurement of Input, Output and Productivity", *Econometrica*, Vol. 50, 1393-1414.

Charnes, A., Cooper, W. W. and Rhodes, E., (1978), "Measuring the Efficiency of Decision Making Units", European Journal of Operational research, Vol. 2, 429-444.

Efron, B. and Tibshirani, R. J., (1993), "*An Introduction to the Bootstrap*", Monographs on Statistics and Applied Probability, No. 57, Chapman and Hall, New York, London.

Ferrier, G. D. and Hirschberg, J. G., (1994), "Bootstrapping DEA Efficiency Scores: With an Application to Italian Banks", unpublished working paper.

Färe, R., Grosskopf, S., Lindgren, B. and Roos, P., (1989), "Productivity Developments in Swedish Hospitals: A Malmquist Output Index Approach", Department of Economics Discussion Paper 89-3, Southern Illinois University, Carbondale. Published in Charnes, A., Cooper, W. W., Lewin, A. Y. and Seiford, L. S. (eds.), (1995), "*Data Envelopment Analysis: Theory, Methodology and Applications*", Kluwer Academic Publishers, Amsterdam.

Grosskopf, S., (1996), "Statistical Inference and Nonparametric Efficiency: A Selective Survey", *Journal of Productivity Analysis,* Vol. 7, No. 2/3, 161-176.

Gstach, D., (1995), "Comparing Structural Efficiency of Unbalanced Subsamples: A Resampling Adaptation of Data Envelopment Analysis", *Empirical Economics*, Vol. 20, 531-542.

Shephard, R., (1970), "*Theory of Cost and Production Functions*", Princeton University Press, Princeton, New Jersey.

Simar, L. and Wilson, P. W., (1995), "Sensitivity Analysis of Efficiency Scores: How to Bootstrap in Nonparametric Frontier Models", Discussion Paper 9503, Institut de Statistique, Universite Catholique de Louvain.

Wilson, P. W. and Simar L., (1995), "Bootstrap Estimation for Nonparametric Efficiency Estimates", Unpublished Working Paper, July 1995.

Tables

TABLE 1: Descriptive statistics, inputs and outputs 1992 and 1993

	1992			1993		
	Max	Mean	Min	Max	Mean	Min
Specialist physicians	164	87	22	180	89	31
Other physicians	71	24	0	78	23	0
Number of beds	23	11	2	22	9	1
Cataract	1,495	778	271	1,509	783	257
Glaucoma	312	62	17	212	55	11
Squint	113	45	2	172	44	8
Visits	41,700	22,747	9,098	43,830	23,855	8,940

Table 2: The Malmquist productivity index and decomposition. Original estimates and 95% bootstrap confidence intervals, percentile and bias-corrected. Significant results in bold.

Dep.	Malmquist index					Efficiency change					Technical change					Dep.
	Orig	Percentile		Bias-corrected		Orig	Percentile		Bias-corrected		Orig	Percentile		Bias-corrected		
		lower	upper	lower	upper		lower	upper	lower	upper		lower	upper	lower	upper	
1	0.84	0.67	1.05	0.65	1.03	0.95	0.73	1.18	0.72	1.17	0.89	0.81	1.01	**0.79**	**0.99**	1
2	1.41	**1.41**	**2.14**	0.74	1.47	0.80	0.71	1.32	**0.32**	**0.94**	1.75	**1.42**	**2.16**	**1.39**	**2.13**	2
3	1.43	**1.18**	**1.72**	**1.22**	**1.76**	1.00	0.90	1.24	0.87	1.21	1.43	**1.17**	**1.67**	**1.25**	**1.74**	3
4	1.30	**1.05**	**1.60**	**1.08**	**1.63**	1.00	1.00	1.00	1.00	1.00	1.30	**1.04**	**1.59**	**1.08**	**1.63**	4
5	1.08	0.79	1.24	0.98	1.43	1.10	0.80	1.30	0.99	1.49	0.98	0.87	1.09	0.87	1.09	5
6	1.28	**1.07**	**1.57**	**1.01**	**1.51**	1.00	1.00	1.67	0.77	1.44	1.28	0.82	1.52	0.98	1.69	6
7	0.92	0.86	1.37	0.55	1.06	0.88	0.81	1.34	0.53	1.05	1.04	0.95	1.15	0.94	1.13	7
8	0.97	0.83	1.26	0.73	1.17	0.95	0.80	1.29	0.68	1.17	1.03	0.91	1.09	0.94	1.12	8
9	0.91	0.84	1.32	0.54	1.02	0.81	0.76	1.23	**0.44**	**0.91**	1.11	1.00	1.17	**1.05**	**1.22**	9
10	0.98	0.87	1.40	0.61	1.14	0.89	0.74	1.24	0.56	1.05	1.11	**1.04**	**1.26**	0.98	1.20	10
11	1.39	**1.10**	**1.65**	**1.22**	**1.77**	1.07	0.99	1.47	0.92	1.39	1.31	**1.04**	**1.39**	**1.23**	**1.57**	11
12	2.09	**1.76**	**2.51**	**1.73**	**2.48**	1.00	1.00	1.57	0.83	1.41	2.09	**1.49**	**2.37**	**1.74**	**2.62**	12
13	1.12	0.86	1.31	1.00	1.45	1.06	0.79	1.25	0.93	1.38	1.07	1.00	1.14	0.99	1.13	13
14	0.90	0.85	1.32	0.53	1.00	0.84	0.78	1.23	**0.49**	**0.93**	1.08	**1.01**	**1.14**	**1.01**	**1.14**	14
15	0.86	0.85	1.33	**0.46**	**0.94**	0.81	0.81	1.30	**0.41**	**0.90**	1.06	0.95	1.10	0.98	1.14	15
16	0.91	0.85	1.31	0.56	1.01	0.84	0.79	1.23	**0.49**	**0.93**	1.09	**1.01**	**1.13**	**1.04**	**1.17**	16
17	0.99	0.74	1.13	0.87	1.26	0.99	0.68	1.01	0.81	1.14	1.01	0.85	1.20	0.84	1.20	17
18	1.34	0.98	1.57	**1.19**	**1.77**	1.12	0.83	1.35	0.99	1.51	1.20	**1.08**	**1.34**	**1.08**	**1.35**	18
19	1.22	0.78	1.24	**1.25**	**1.72**	1.23	0.75	1.22	**1.27**	**1.74**	0.99	0.96	1.10	0.91	1.05	19
20	0.90	0.73	1.07	0.78	1.11	1.00	1.00	1.00	1.00	1.00	0.90	0.74	1.07	0.78	1.11	20
21	1.06	0.79	1.20	0.96	1.38	1.07	0.77	1.19	0.97	1.39	0.99	0.94	1.07	0.92	1.05	21
22	1.19	0.79	1.22	**1.21**	**1.64**	1.20	0.78	1.23	**1.21**	**1.66**	0.99	0.89	1.08	0.88	1.07	22
23	0.82	0.76	1.21	**0.49**	**0.95**	0.88	0.83	1.39	0.52	1.08	0.94	0.73	1.09	0.75	1.11	23
24	1.16	0.92	1.38	**1.05**	**1.51**	1.00	0.66	1.00	0.80	1.14	1.16	0.96	1.63	0.89	1.56	24
25	1.03	0.80	1.23	0.88	1.31	1.01	0.79	1.26	0.83	1.31	1.02	0.89	1.14	0.92	1.17	25
26	0.74	**0.62**	**0.95**	**0.50**	**0.82**	0.88	0.62	1.00	**0.61**	**0.99**	0.85	0.80	1.05	**0.66**	**0.91**	26
27	0.88	0.69	1.05	0.72	1.08	0.97	0.71	1.10	0.79	1.18	0.91	0.81	1.03	**0.77**	**0.99**	27
28	1.24	0.87	1.37	**1.20**	**1.70**	1.18	0.82	1.32	**1.14**	**1.64**	1.05	0.96	1.14	0.95	1.12	28
29	0.89	0.64	1.01	0.79	1.16	1.00	0.65	1.00	0.83	1.17	0.89	0.75	1.08	0.74	1.07	29

II

Scale Efficiency and Scale Elasticity in DEA-Models: A Bootstrapping Approach*

Mickael Löthgren† and Magnus Tambour‡
Stockholm School of Economics
P.O. Box 6501
S-113 83 Stockholm
Sweden

Working Paper series in Economics and Finance No. 91
January 1996
Revised November 1996

Abstract

The purpose of this paper is to present methods for estimation and tests of firm specific returns to scale measures using data envelopment analysis. Both qualitative (scale efficiency) and quantitative (scale elasticity) measures are considered. Scale elasticity is estimated using a simple primal-based approximation. We propose an easily implementable bootstrap algorithm to perform statistical tests on firm-specific estimates. These types of tests are not possible to perform directly in a standard DEA-model. An empirical application using Swedish hospital data is provided. The results are similar for both returns to scale measures in the original results, but some differences occur in the hypotheses tests.

Keywords: Bootstrap, DEA, Scale Efficiency, Scale Elasticity.
JEL-Classification: C14, C15, D24.

*We thank Sune Karlsson, Anders Westlund, Henry Tulkens and two anonymous referees for valuable comments and SPRI and the Swedish Ophthalmology Society for providing data.

†Department of Economic Statistics. *E-mail:* stml@hhs.se. *Fax:* +46 - 8 - 34 81 91.

‡Centre for Health Economics. *E-mail:* hemt@hhs.se. *Fax:* +46 - 8 - 30 21 15. Research support from the National Corporation of Swedish Pharmacies (Apoteksbolaget) is gratefully acknowledged.

1. Introduction

Two approaches to characterize scale economies are given in the economics literature. The first is the neoclassical - production function based - approach (see, e.g., Frisch (1965)). The second is the axiomatic approach (see, e.g., Shephard (1970)). While the first approach is mainly concerned with *quantitative* scale measures (elasticity of scale) - the second approach has within data envelopment analysis (DEA) until recently been devoted to obtain *qualitative* information of scale economies. One example is the "scale efficiency method," outlined in, e.g., Färe, Grosskopf and Lovell (1994), which identifies returns to scale qualitatively by comparing efficiency measures under different returns to scale restrictions. Equivalent methods have been presented by Banker (1984) and Banker, Charnes and Cooper (1984). In recent work, efforts have been made to obtain quantitative DEA measures of returns to scale (Banker and Thrall (1992), Forsund (1996)). The neoclassical notion of scale elasticity, defined in terms of efficiency measures in Färe, Grosskopf and Lovell (1988), can be used in DEA-models to measure the magnitude of returns to scale. As Forsund (1996) notes, knowledge of scale efficiency does not, however, permit an estimation of scale elasticity.

Returns to scale is a frontier concept and it only makes sense to define returns to scale for frontier inputs-outputs. The choice of orientation - input decreasing or output increasing - plays an important role for estimating scale elasticity (and scale efficiency). It is possible that an input-based approach gives different returns to scale results compared to an output-based approach. This holds both for scale efficiency and scale elasticity estimation and the more inefficient a firm is, the more diverse returns to scale results can be obtained. The choice of orientation is ultimately up to the researcher and depends on the application under study.

The DEA-method has several nice features. It readily models multiple-input, multiple-output technologies, even when price data are not available. No behavioral assumptions need to be imposed, which is an appealing feature when studying efficiency in public owned production. One drawback, though, of the DEA-method is that statistical hypotheses testing of firm-specific scale properties cannot be performed directly on the DEA-estimates.

There has been comparatively few studies on statistical testing of returns to scale in DEA. Banker (1996) summarizes his earlier work and discuss hypothesis testing in DEA-models. His approach concerns global tests for returns to scale properties for the technology. No tests are presented for firm-specific scale measures. Simar (1996) points out the bootstrap as a general solution to perform inference in DEA-models for firm-specific measures. The DEA application of the bootstrap method is still under development. A number of studies listed in Grosskopf (1996) have applied the bootstrap to Farrell (1957) efficiency measures and Löthgren and Tambour (1995) bootstrap the Malmquist index. However, bootstrap studies of scale measures in DEA-models are unknown to the authors.

The purpose of the paper is to present a bootstrap algorithm that allows testing

firm-specific scale efficiency and scale elasticity. Furthermore, a simple approximation of firm-specific scale elasticity in DEA-models is presented. The approximation provides lower and upper bounds of the scale elasticity. This approximation is used in the bootstrap. In the paper we use a new approach to calculate bias corrected and accelerated BC_a bootstrap confidence intervals based on a linear approximation of the statistic. The proposed methodology is illustrated in an empirical application using data from Swedish eye-care departments. We use an output oriented approach since it is most appropriate for the empirical application where data on health care departments are used. These departments receive a (fixed) budget and strive to provide as much health care services as possible.

The paper unfolds as follows: section 2 contains a theoretical presentation of the production technology and the DEA-models. Section 3 presents two approaches to estimate returns to scale in DEA and describe the proposed approximation of elasticity of scale. Section 4 - 6 contain formal descriptions of the hypotheses tests performed, the bootstrap algorithm and the construction of the bootstrap confidence intervals used in the tests. Section 7 continues with the empirical example. Section 8 concludes the paper with a summary.

2. Production theory and DEA

Let $x \in R_+^N$ denote a vector of inputs used by one firm in the production of $y \in R_+^M$ outputs. In an output-based setting, where the input quantities are taken as given and outputs are the choice variables, the production technology can be described by the output set, defined as

$$P(x) = \{y : x \, can \, produce \, y\}. \tag{2.1}$$

The technology is assumed to satisfy a set of maintained axioms discussed in, e.g., Färe (1988).

A scalar valued representation of the technology, in terms of the output set, can be given by the efficiency measure (c.f. Farrell (1957))[1],

$$F_o(y,x) = \max\{\lambda : \lambda y \in P(x)\}. \tag{2.2}$$

The efficiency measure takes on values larger than, or equal to unity, for a feasible observation, i.e., $F_o(y,x) \geq 1$ if and only if $y \in P(x)$. A value larger than one indicates technical inefficiency.

[1]This measure is often called the output-based Farrell efficiency measure, but as pointed out in Forsund (1996) the efficiency measure originally proposed by Farrell (1957) is in fact the output distance function.

2.1. Data envelopment analysis

Consider K firms employing N inputs in the production of M outputs. The efficiency measures can be estimated by solving K linear programs for each technology satisfying either constant returns to scale (CRS), non-increasing returns to scale (NIRS) or variable returns to scale (VRS). The estimated (technical) efficiency for firm k, when CRS is imposed, is obtained from the solution to the LP problem (Charnes, Cooper and Rhodes (1978))

$$\widehat{F}_o\left(y_k, x_k | CRS\right) = \max_z \left\{\theta : \theta y_k \leq Yz, x_k \geq Xz, z \in R_+^K\right\}, \tag{2.3}$$

where Y is a $(M \times K)$ matrix of outputs, X is a $(N \times K)$ matrix of inputs and z is a K-vector of intensity variables. The efficiency estimate, under an assumption of NIRS, is obtained from the LP problem (see, e.g., Färe, Grosskopf and Lovell (1994))

$$\widehat{F}_o\left(y_k, x_k | NIRS\right) = \max_z \left\{\theta : \theta y_k \leq Yz, x_k \geq Xz, 1_K' \, z \leq 1, z \in R_+^K\right\}, \tag{2.4}$$

where 1_K is a K-vector of one's. Yet another efficiency measure, under variable returns to scale (VRS), is estimated by (see, e.g., Färe, Grosskopf and Lovell (1994))

$$\widehat{F}_o\left(y_k, x_k | VRS\right) = \max_z \left\{\theta : \theta y_k \leq Yz, x_k \geq Xz, 1_K' \, z = 1, z \in R_+^K\right\}. \tag{2.5}$$

It is clear from the LP formulations that the three efficiency measures are nested as follows

$$\widehat{F}_o\left(y_k, x_k | CRS\right) \geq \widehat{F}_o\left(y_k, x_k | NIRS\right) \geq \widehat{F}_o\left(y_k, x_k | VRS\right) \geq 1\,. \tag{2.6}$$

3. Returns to scale in DEA-models

3.1. Scale efficiency

One way to identify the nature of returns to scale is to use the scale efficiency measure as outlined in, e.g., Färe, Grosskopf and Lovell (1994). An estimate of the output-based scale efficiency measure is defined as

$$\widehat{S}_{o1}\left(y, x\right) = \frac{\widehat{F}_o\left(y, x \,|\, CRS\right)}{\widehat{F}_o\left(y, x \,|\, VRS\right)}. \tag{3.1}$$

Since $\widehat{F}_o\left(y, x | CRS\right) \geq \widehat{F}_o\left(y, x | VRS\right)$, the ratio satisfies $S_{o1} \geq 1$. A value equal to one indicates scale efficiency. A firm with $S_{o1} = 1$ is scale efficient in the sense that the chosen input-output mix is optimal and maximizes the average (multiple-output) productivity. Furthermore, the input-output mix is equally efficient to the CRS as to the VRS technology.

Given $\widehat{S}_{o1} \geq 1$, the input-output mix is not scale efficient, and the firm in question is operating either in a region of increasing returns to scale (inefficient small scale), or in a region of decreasing returns to scale (inefficient large scale). This can be determined by comparing the CRS output-based efficiency measure to the equivalent NIRS-measure. Thus, the ratio

$$\widehat{S}_{o2}(y,x) = \frac{\widehat{F}_o(y,x \mid CRS)}{\widehat{F}_o(y,x \mid NIRS)}, \tag{3.2}$$

indicates whether the scale-inefficiency is due to a too small scale or a too large scale. Following, e.g., Färe, Grosskopf and Lovell (1994) increasing returns to scale is inferred when $\widehat{S}_{o2}\left(y,x \left| \widehat{S}_{o1} > 1\right.\right) = 1$, and decreasing returns to scale when $\widehat{S}_{o2}\left(y,x \left| \widehat{S}_{o1} > 1\right.\right) > 1$.

The output-based scale efficiency measure is illustrated in Figure 1 where a single input is used in the production of a single output. The letters A, B, C, D, E and F denote six input-output observations. Firm B is scale efficient since $\widehat{S}_{o1} = 1$, whereas A, C, D, E and F are not scale efficient. Both B and D are, however, technical efficient relative to the VRS technology. A is operating at a too small scale (increasing returns to scale, i.e., $\widehat{S}_{o2} = 1$) and C, D, E and F are operating at too large scales (decreasing returns to scale, i.e., $\widehat{S}_{o2} > 1$). Note that the results are altered in an input-based scale efficiency analysis for observation C and F. For those observations the input scale efficiency would lead to a conclusion of increasing returns to scale.

FIGURE 1 IN ABOUT HERE

3.2. Scale elasticity

The scale efficiency measure provides a qualitative measure of returns to scale and classify observations as belonging to increasing, decreasing or constant returns to scale regions of the technology. In some instances the quantitative magnitude of scale economies is also of interest. The elasticity of scale provide such a measure It is well known that for a single output technology the scale elasticity is defined as

$$\varepsilon(x) = \frac{\partial \ln f(\lambda x)}{\partial \ln \lambda}\mid_{\lambda=1} = \sum_{n=1}^{N} \frac{\partial f(x)}{\partial x_n}\frac{x_n}{f(x)}, \tag{3.3}$$

where $f(x)$ denotes a production function. In the multiple-output situation an output oriented scale elasticity can be defined in terms of the output-based efficiency measure as (Färe, Grosskopf and Lovell (1988), Forsund (1996))

$$\varepsilon_o(y,x) = \sum_{n=1}^{N} \frac{\partial F_o\left(y^f,x\right)}{\partial x_n} \cdot x_n = \nabla_x F_o\left(y^f,x\right) \cdot x, \tag{3.4}$$

where $y^f = F_o(y,x) \cdot y$ is a frontier output belonging to the isoquant of the output set, defined as $IsoqP(x) = \{y : y \in P(x), \lambda y \notin P(x), \lambda > 1\}$.

One approach to estimate scale elasticity in DEA-models, based on the dual to (2.5), is given in Banker and Thrall (1992). As an alternative to their approach we estimate the scale elasticity using a direct primal based approximation.

3.3. Approximating scale elasticity using DEA

We propose the following approximation of the elasticity of scale within a DEA framework. For fixed outputs, consider equal proportional changes in each input dimension, i.e., $dx = \delta x$ for a positive scalar. The total differential of the output efficiency measure can be expressed as

$$dF_o(y,x) = \delta \sum_{n=1}^{N} \frac{\partial F_o}{\partial x_n} x_n = \delta \nabla_x F_o(y,x) \cdot x. \tag{3.5}$$

Hence, the inner product $\nabla_x F_o(y,x) \cdot x$ can be expressed in terms of the differential as $\nabla_x F_o(y,x) \cdot x = \frac{dF_o(y,x)}{\delta}$. Using this, and condition on proportional changes in inputs, the scale elasticity is given by

$$\varepsilon_o\left(y^f, x\right) = \nabla_x F_o\left(y^f, x\right) \cdot x = dF_o\left(y^f, x\right)/\delta. \tag{3.6}$$

A straightforward approximation of the scale elasticity is obtained by approximating the differential $dF_o\left(y^f, x\right)$ by the DEA-based difference of the efficiency measure. This leads to the following scale elasticity approximations:

$$\widehat{\varepsilon}_o^{-}\left(\widehat{y}^f, x, \delta\right) = \Delta^{+}\widehat{F}_o\left(\widehat{y}^f, x, \delta\right)/\delta \tag{3.7}$$

and

$$\widehat{\varepsilon}_o^{+}\left(\widehat{y}^f, x, \delta\right) = \Delta^{-}\widehat{F}_o\left(\widehat{y}^f, x, \delta\right)/\delta, \tag{3.8}$$

where $\widehat{y}^f = \widehat{F}_o(y,x) \cdot y \in Isoq\widehat{P}(x)$ and $Isoq\widehat{P}(x)$ is the DEA-estimate of the output isoquant. $\Delta^{+}\widehat{F}_o\left(\widehat{y}^f, x, \delta\right) = \widehat{F}_o\left(\widehat{y}^f, (1+\delta)\, x \,|VRS\right) - \widehat{F}_o\left(\widehat{y}^f, x \,|VRS\right)$ and $\Delta^{-}\widehat{F}_o\left(\widehat{y}^f, x, \delta\right) = \widehat{F}_o\left(\widehat{y}^f, x \,|VRS\right) - \widehat{F}_o\left(\widehat{y}^f, (1-\delta)\, x \,|VRS\right)$ denote the "right" and "left" DEA-difference approximations based on proportional increases and decreases in inputs, respectively.

The proposed approximations give lower and upper bounds on the scale elasticity, as stated in the following proposition.

Proposition 3.1. *The scale elasticity approximations satisfy*

$$\widehat{\varepsilon}_o^{-}\left(\widehat{y}^f, x, \delta\right) \leq \widehat{\varepsilon}_o^{+}\left(\widehat{y}^f, x, \delta\right), \ \forall \delta \in [0,\ 1]. \tag{3.9}$$

Proof. Löthgren and Tambour (1996) establish a negative monotonicity of the scale elasticity approximation.[2] I.e.,

$$\widehat{\varepsilon_o^-}\left(\widehat{y}^f, x, \delta''\right) = \frac{\Delta^+ \widehat{F}_o\left(\widehat{y}^f, x, \delta''\right)}{\delta''} \leq \frac{\Delta^+ \widehat{F}_o\left(\widehat{y}^f, x, \delta'\right)}{\delta'} = \widehat{\varepsilon_o^-}\left(\widehat{y}^f, x, \delta'\right) \quad \forall\, \delta' \geq \delta''. \tag{3.10}$$

Use $\delta'' = -\delta' = \delta$, $0 \leq \delta \leq 1$, in (3.10) and note that $\widehat{\varepsilon_o^-}\left(\widehat{y}^f, x, -\delta\right) = -\Delta^+ \widehat{F}_o\left(\widehat{y}^f, x, -\delta\right)/\delta = \Delta^- \widehat{F}_o\left(\widehat{y}^f, x, \delta\right)/\delta = \widehat{\varepsilon_o^+}\left(\widehat{y}^f, x, \delta\right)$. Using this we have

$$\widehat{\varepsilon_o^-}\left(\widehat{y}^f, x, \delta\right) \leq \widehat{\varepsilon_o^-}\left(\widehat{y}^f, x, -\delta\right) = \widehat{\varepsilon_o^+}\left(\widehat{y}^f, x, \delta\right), \tag{3.11}$$

and the proposition is proved. ■

The proposed approximations equal the theoretical scale elasticity measure for a CRS-technology. This follows since under CRS the DEA differences satisfies $\Delta^- \widehat{F}_o(\cdot) = \Delta^+ \widehat{F}_o(\cdot) = \delta \widehat{F}_o(y, x)$, which in turn implies that $\widehat{\varepsilon_o^-}\left(\widehat{y}^f, x, \delta\right) = \widehat{\varepsilon_o^+}\left(\widehat{y}^f, x, \delta\right) = 1$.

4. Hypothesis testing

Given the DEA-estimates of the scale efficiency and elasticity, it is reasonable to ask whether the estimates are statistically significant or not. For example, if the scale efficiency estimate $\widehat{S}_{o1} > 1$ one may ask whether this really indicates scale inefficiency or not in a statistical sense. Similarly, if the estimate of the scale elasticity differs from unity, one may ask if this is a significant deviation from CRS or not.

We propose firm-specific test procedures based on bootstrap confidence intervals for tests of local returns to scale properties. For ease of notation, we ignore the firm k index in the presentation.

4.1. Scale efficiency testing

The scale efficiency test is performed using a nested test procedure given by:

H_{01} : $S_{o2} = 1$ - Scale efficient or Increasing returns to scale

H_{11} : $S_{o2} > 1$ - Decreasing returns to scale

and

H_{02} : $S_{o1} = 1 \,|S_{o2} = 1$ - Scale efficient

H_{12} : $S_{o1} > 1 \,|S_{o2} = 1$ - Increasing returns to scale

[2]The monotonicity proof in Löthgren and Tambour (1996) is given for the input based scale elasticity. The monotonicity of the output based scale elasticity follows analogously.

The ordering of the tests follows from the fact that $S_{o1} \geq S_{o2} \geq 1$.This procedure allows controlling for the overall significance level for rejecting scale efficiency. Let α denote the common significance level in the first and second test. The overall significance level of rejecting scale efficiency, given that scale efficiency is prevalent, is given by $\alpha_0 = \alpha + (1 - \alpha)\,\alpha$.

Let $\widehat{S}^*_{oi}(\alpha)$, $i = 1, 2$, denote the lower bound of a one-sided $(1 - \alpha)$ bootstrap confidence interval for S_{oi}. The test procedure is as follows: if $\widehat{S}^*_{o2}(\alpha) > 1$, then H_{01} is rejected and we conclude that the firm operates under decreasing returns to scale. If, on the other hand, $\widehat{S}^*_{o2}(\alpha) = 1$, H_{01} cannot be rejected and we continue to the second test. If $\widehat{S}^*_{o1}(\alpha) > 1$, then H_{02} is rejected and it is concluded that the firm operates under increasing returns to scale. Finally, if $\widehat{S}^*_{o1}(\alpha) = 1$, H_{02} cannot be rejected and it is concluded that the firm is scale efficient.

4.2. Scale elasticity testing

Let $\widehat{\varepsilon}^{*-}_o(\alpha)$ denote the lower bound on a $(1 - \alpha)$ one-sided bootstrap confidence interval for the scale elasticity lower bound approximation. Let $\widehat{\varepsilon}^{*+}_o(1 - \alpha)$ denote the upper bound on $(1 - \alpha)$ one-sided confidence interval for the scale elasticity upper bound approximation.

The test procedures are as follows:

1. The hypothesis $H_0 : \varepsilon_o \leq 1$ is rejected if $\widehat{\varepsilon}^{*-}_o(\alpha) > 1$. I.e., the hypothesis of DRS or CRS is rejected if the lower limit of the confidence interval for the lower bound of the elasticity approximation is greater than one.

2. The hypothesis $H_0 : \varepsilon_o \geq 1$ is rejected if $\widehat{\varepsilon}^{*+}_o(1 - \alpha) < 1$. I.e., the hypothesis of IRS or CRS is rejected if the upper limit of the confidence interval for the upper bound of the elasticity approximation is less than one.

If $\widehat{\varepsilon}^{*-}_o(\alpha) < 1 < \widehat{\varepsilon}^{*+}_o(1 - \alpha)$, neither $H_0 : \varepsilon_{o1} \leq 1$ nor $H_0 : \varepsilon_{o1} \geq 1$ can be rejected. In other words, depending on the null hypothesis, CRS (or DRS/IRS) cannot be rejected in this case. The main point is that confidence intervals for both the lower and the upper bound of the approximation are needed to completely discriminate between the different forms of returns to scale.

5. The Bootstrap Method

The bootstrap method is a well established computationally intensive statistical resampling method used to perform inference in complex problems, see, e.g., Efron and Tibshirani (1993) for a presentation of the method. Due to the complexity of the DEA-estimator, no analytical solutions of the bootstrap are generally available. The basic idea of the bootstrap method is to approximate the sampling distributions of the DEA scale estimators by Monte Carlo simulation where repeated resamples of the

observed data produce repeated scale estimates. The empirical distributions of the simulated scale estimates approximates the sampling distributions of the estimators. The resamples are generated using an estimate of the data generating process (DGP).

The DGP underlying the bootstrap is output-oriented. The outputs are given by random radial deviations off the frontier, i.e., the isoquant of the output set. Formally, the input-output observations are given by

$$(x_k, y_k) = \left(x_k, y_k^f / F_{o,k}\right), k = 1, ..., K, \tag{5.1}$$

where $F_{o,k}$ is a stochastic efficiency measure, defined in (2.2), and $y_k^f \in IsoqP(x_k)$ is the unobservable frontier output for the k:th firm. The efficiencies are assumed to be drawn from the same distribution, i.e., $F_{o,k} \sim G_{F_o}\ \forall\ k$, where G_{F_o} denotes the common (unknown) distribution.

This DGP-model represents the idea that, conditioned on the inputs and the output proportions, the stochastic elements in the production process are completely represented by the random output efficiency measures.

The main idea in the simulation is to mimic the DGP. In short, the procedure in each resample is as follows: Conditioned on observed inputs and output proportions, the resample data are constructed in two steps. First, the frontier outputs are estimated. Second, bootstrap pseudo-outputs are generated by replicating the DGP in (5.1) using the estimated frontier outputs and pseudo-efficiencies drawn from some estimate of the distribution G_{F_o}. In this paper we use the simplest (naive) non-smoothed and non-parametric bootstrap where the empirical distribution of the estimated efficiencies is used to generate the pseudo-efficiencies.

The bootstrap Monte Carlo algorithm is given by the following steps:

1. Let the transformed input-output vectors be given by $\left(x_k, \widehat{y}_k^f\right) = \left(x_k, \widehat{F}_{o,k} \cdot y_k\right)$, $k = 1, ..., K$, where the transformed output vectors are efficient in the sense that $\widehat{y}_k^f \widehat{F}_{o,k} \cdot y_k \in Isoq\widehat{P}(x) \Leftrightarrow \widehat{F}_o\left(\widehat{y}_k^f, x_k\right) = 1$.

2. Resample, with replacement, K efficiency measures from $\widehat{F}_o(\cdot\,|CRS) = \left(\widehat{F}_{o,1}(\cdot\,|CRS), \ldots, \widehat{F}_{o,K}(\cdot\,|CRS)\right)$. Let the vector $\delta^* = (\delta_1^*, \ldots, \delta_K^*)$ denote the K resampled efficiency measures.

3. Let the bootstrap pseudo-cases be given by $(x_k^*, y_k^*) = \left(x_k, (\delta_k^*)^{-1} \widehat{y}_k^f\right)$, $k = 1, ..., K$.

4. The bootstrap output efficiencies, under CRS, are obtained from LP-problems similar to (2.3) as

$$\widehat{F}_o^{*b}\left(y_k^{*b}, x_k\,|CRS\right) = \max_z \left\{\theta : \theta y_k^{*b} \leq Y^{*b} z, x_k \geq Xz,\ z \in R_+^K\right\}, k = 1, \ldots, K. \tag{5.2}$$

The bootstrap efficiency measures for the NIRS and VRS-technologies are obtained similarly as in (2.4) and (2.5), respectively.

5. Repeat step (2) - (4) B times[3] to create B bootstrap K-vectors
$\widehat{F}_o^{*b} = \left(\widehat{F}_{o,1}^{*b}, \ldots, \widehat{F}_{o,K}^{*b}\right), b = 1, ..., B.$

Using the bootstrap CRS-, NIRS- and VRS-efficiency measures the bootstrap scale efficiency measures $\widehat{S}_{o1}^{*b}$ and $\widehat{S}_{o2}^{*b}$ are given by

$$\widehat{S}_{o1}^{*b} = \frac{\widehat{F}_o^{*b}\left(y^{*b}, x \,|CRS\right)}{\widehat{F}_o^{*b}\left(y^{*b}, x \,|VRS\right)}, b = 1, ..., B, \tag{5.3}$$

and

$$\widehat{S}_{o2}^{*b} = \frac{\widehat{F}_o^{*b}\left(y^{*b}, x \,|CRS\right)}{\widehat{F}_o^{*b}\left(y^{*b}, x \,|NIRS\right)}, b = 1, ..., B. \tag{5.4}$$

The bootstrapped bounds of the scale elasticity are given by

$$\widehat{\varepsilon}_o^{-}\left(\widehat{y}^{*fb}, x\right) = \Delta^{+}\widehat{F}_o^{*b}\left(\widehat{y}^{*fb}, x\right)/\delta, b = 1, ..., B, \tag{5.5}$$

and

$$\widehat{\varepsilon}_o^{+}\left(\widehat{y}^{*fb}, x\right) = \Delta^{-}\widehat{F}_o^{*b}\left(\widehat{y}^{*fb}, x\right)/\delta, b = 1, ..., B, \tag{5.6}$$

where $\widehat{y}^{*fb} = \widehat{F}_o^{*b}\left(y^{*b}, x\right) \cdot y^{*b} \in Isoq\widehat{P}^{*b}(x)$ is the bootstrap frontier output.

If data from more than one period are available the dynamic structure of the scale properties is kept in the bootstrap if the same randomly drawn firms are used in each time period to resample the scale measures. This would allow for tests of firm-specific intertemporal changes in scale properties.

Note that the CRS-technology is used as reference technology in the Monte Carlo simulation algorithm. The pseudo-output data are generated as deviations off the isoquant of the CRS-output set, using the CRS-efficiency measures to create the inefficiencies. Another approach would be to use the VRS-output set as reference set. One argument against this is that it would lead to a too small variability in the bootstrapped efficiency estimates.

5.1. Some remarks on the bootstrap procedure

The DGP in (5.1) underlying the bootstrap is similar to earlier work on bootstrapping of DEA-models. Löthgren and Tambour (1995), Simar and Wilson (1995), Wilson and Simar (1995) as well as some other work cited in Grosskopf (1996) specify this DGP.

Given the DGP, a fundamental issue is whether to condition the bootstrap efficiency estimates on the original data or not? Our approach is based on the resampled

[3]The number of replications is set to $B = 1000$. Efron and Tibshirani (1993), p. 275, recommend at least this number of simulation replicates in order to make the variability of the boundaries of the bootstrap confidence intervals "acceptably" low.

data both for the bootstrap estimation of the frontiers and the efficiencies. Most earlier work use the resampled data to estimate the bootstrap frontiers but condition the bootstrap efficiency estimates on the original observations. The key difference is in step 4 in the algorithm, where a conditioned bootstrap efficiency estimate is obtained by substituting the bootstrap output data y_k^{*b} by the original output y_k.

The sampling distribution of the DEA estimator is a function of realizations of two stochastic events. First, the observations are generated by a stochastic mechanism as specified in the DGP. Second, the efficiency estimators are functions of the frontier estimates. A valid resampling procedure should take this double stochastic feature into consideration. In our resampling algorithm the bootstrap frontier estimates and the bootstrap efficiency estimates are based on the resampled data in the same manner as the original estimates are based on the original data. In this way both the stochastic DGP and the sampling distribution of the frontier estimates will be incorporated in the bootstrap efficiency estimates in (5.2). The empirical distributions of the bootstrap efficiencies approximate the common distribution of the efficiencies, conditional on the inputs and the output proportions. This procedure ignore the original efficiency estimates when estimating sampling distributions of the firm-specific efficiency estimator. The firm-specific estimates are however taken into account in the bootstrap BC_a confidence intervals.

The DGP in (5.1) models the production frontier as deterministic in the sense that the frontier is fixed and not affected by random events. The stochastic element is the random inefficiencies off this frontier. The DGP specifies that the observed outputs always lie within the output set, i.e., the efficiency measures are truncated below at one. Our approach guarantees that this condition is fulfilled in the sense that $\widehat{F}_o^{*b} \geq 1, \forall b$. This property is not necessarily satisfied if the bootstrap efficiency estimates are conditioned on the original observations as in Simar and Wilson since in this case it is possible that $\widehat{F}_o^{*b} < 1$ for some b.

We choose, by arguments of simplicity, to use the simple non-parametric non-smoothed resampling scheme. If the bootstrap efficiency estimates are conditioned on the original observations, Wilson and Simar (1995) note that the simple bootstrap gives inconsistent estimates of the sampling distribution of the efficiency estimates. This inconsistency argument cannot, however, be raised against the unconditional bootstrap procedure in this paper since our approach is not conditioned on the original observations. Simar and Wilson propose the use of a smoothed bootstrap to obtain a consistent procedure. The proposed smoothing procedure can of course be incorporated into our bootstrap algorithm as well, although at the cost of increased complexity.

6. Confidence Intervals[4]

The percentile method, see e.g., Efron and Tibshirani (1993), is the most straightforward method to obtain bootstrap confidence intervals. The percentile method is based on the empirical cumulative distribution function (CDF) of $\widehat{S}_{o,k}^{*b}$, $b = 1, ..., B$, defined as $\widehat{G}_k(s) = \frac{1}{B}\sum_{b=1}^{B} I\left(\widehat{S}_{o,k}^{*b} \leq s\right)$, for any real value s, where $I(\cdot)$ denotes the indicator function. A $(1-2\alpha)$ equal-tail confidence interval for $S_{o,k}$ is given by the interval

$$\left(\widehat{S}_{o,k}^{*}(\alpha), \widehat{S}_{o,k}^{*}(1-\alpha)\right), k = 1, ..., K, \tag{6.1}$$

where $\widehat{S}_{o,k}^{*}(\alpha)$ is the α:th quantile of $\widehat{G}_k$, i.e., $\widehat{S}_{o,k}^{*}(\alpha) = \widehat{G}_k^{-1}(\alpha)$. The quantiles of $\widehat{G}_k$ are given by the $[(B+1)\alpha]$:th and the $[(B+1)(1-\alpha)]$:th ordered values of $\widehat{S}_{o,k}^{*b}$, $b = 1, ..., B$, respectively, where $[r]$denotes the integer part of any real value r.

An extension of the simple percentile method is the bias-corrected and accelerated (BC_a) method (Efron and Tibshirani (1993)). This method is a modified percentile method with $(1-2\alpha)$ confidence intervals given by

$$\left(\widehat{S}_{o,k}^{*}(\alpha_{1,k}), \widehat{S}_{o,k}^{*}(\alpha_{2,k})\right), k = 1, ..., K, \tag{6.2}$$

where the two quantiles $\alpha_{1,k}$ and $\alpha_{2,k}$ are given by

$$\alpha_{1,k} = \Phi\left(\widehat{z}_{0,k} + \frac{\widehat{z}_{0,k} + z(\alpha)}{1 - \widehat{a}_k\left(\widehat{z}_{0,k} + z(\alpha)\right)}\right), \tag{6.3}$$

and

$$\alpha_{2,k} = \Phi\left(\widehat{z}_{0,k} + \frac{\widehat{z}_{0,k} + z(1-\alpha)}{1 - \widehat{a}_k\left(\widehat{z}_{0,k} + z(1-\alpha)\right)}\right). \tag{6.4}$$

$\Phi(\cdot)$ denotes the standard normal CDF and $z(\cdot)$ denotes the α:th quantile of the standard normal distribution. The bias-adjustment term $\widehat{z}_{0,k}$ is given by

$$\widehat{z}_{0,k} = \Phi^{-1}\left(\widehat{G}_k\left(\widehat{S}_{o,k}\right)\right) = \Phi^{-1}\left(\frac{1}{B}\sum_{b=1}^{B} I\left(\widehat{S}_{o,k}^{*b} \leq \widehat{S}_{o,k}\right)\right). \tag{6.5}$$

Following Efron (1990), the acceleration adjustment term is estimated as

$$\widehat{a}_k = \frac{\sum_{k'=1}^{K} \widehat{u}_{k,k'}^3}{6\left(\sum_{k'=1}^{K} \widehat{u}_{k,k'}^2\right)^{3/2}}, \tag{6.6}$$

[4]This section is explicitly concerned with confidence intervals for the scale efficiency. Intervals for the scale elasticity are obtained analogously.

where $\widehat{u}_{k,k'}$, $k, k' = 1, ..., K$, denote estimates of the empirical influence functions. If both $\widehat{a}_k = 0$ and $\widehat{z}_0 = 0$, the resulting interval is the standard percentile interval in (6.1).

The BC_a-method is based on implicit transformation functions that are bias-correcting and variance stabilizing. The method gives confidence intervals with coverage accuracy of $O(K^{-1})$, i.e., $\Pr\left\{\widehat{S}^*_{o,k}(\alpha_{1,k}) \leq S_{o,k} \leq \widehat{S}^*_{o,k}(\alpha_{1,k})\right\} = 1 - 2\alpha + c/K$, for any constant c, $k = 1, ..., K$. This is in contrast to the accuracy of $O\left(K^{-1/2}\right)$ for the percentile method (Efron and Tibshirani (1993), p. 187). This property motivates the use of the BC_a-intervals.

Estimates of the empirical influence functions $u_{k,k'}$ are obtained using a first order approximation of the statistic. Conditional on the original sample, the bootstrapped scale efficiency is expanded as

$$\widehat{S}^{*b}_{o,k} \approx \widehat{S}_{o,k} + \frac{1}{K}\sum_{k'=1}^{K} f^{*b}_{k'} u_{k,k'}, k = 1, ..., K, b = 1, ..., B, \tag{6.7}$$

where $f^{*b}_{k'}$ denotes the resample frequency of case k' in the b:th bootstrap replication and $u_{k,k'}$ is the empirical influence functions satisfying the restriction $\sum_{k'=1}^{K} u_{k,k'} = 0$, $k = 1, ..., K$. This restriction follows from the fact that if the original sample is obtained in a resample the bootstrap scale efficiency equals the original estimate. The empirical influence functions $u_{k,k'}$ can be estimated by a standard restricted OLS-estimator given by

$$\widehat{u}_k = \widehat{u}^{OLS}_k - (F^{*\prime}F^*)^{-1} 1_K \left(1'_K (F^{*\prime}F^*)^{-1} 1_K\right)^{-1} 1'_K \widehat{u}^{OLS}_k, k = 1, ..., K, \tag{6.8}$$

where $\widehat{u}^{OLS}_k = K(F^*F'^*)^{-1} F^{*\prime}\left(\widehat{S}^*_{o,k} - \widehat{S}_{o,k} 1_B\right)$, $k = 1, ..., K$, is the unrestricted OLS-estimate of the influence functions. $\widehat{u}_k = (\widehat{u}_{k,1}, \ldots, \widehat{u}_{k,K})$ is the K-vector of estimated empirical influence functions. $\widehat{S}^*_{o,k}$ is a B-vector of bootstrapped scale efficiency measures, and $F^* = \left(f^{*1}, \ldots, f^{*b}, \ldots, f^{*B}\right)'$ is a $(B \times K)$-matrix of resampling frequencies.

7. Empirical application

7.1. Data

The data consist of primal production data from 29 Swedish public eye-care (ophthalmology) departments operating in 1993. The 29 departments represent 85 percent of the total number of (public) eye-care departments in Sweden. All departments are located at hospitals owned and organized by the County Councils which provide almost all inpatient care in Sweden. Five of the departments are located at large (teaching) hospitals, 19 at middle size hospitals and 5 at small hospitals. The data are taken from

a database compiled by SPRI (the Swedish Institute for Health Services Development) in collaboration with the Swedish Ophthalmology Society.

The input variables contain two measures on labor inputs, defined in terms of full time equivalent months for two physician categories (specialized and other physicians). Good measures on capital use in health care provision (at least in Sweden) are notoriously difficult to obtain. Although one could argue that capital costs amounts to a small fraction of the total cost in health care provision (approximately 5 percent, Federation of County Councils (1993)) we choose not to ignore it. The number of available beds is therefore included as a crude proxy for capital input.

Three common ophthalmology procedures (cataract surgery, operations of glaucoma and squint) and number of visits are used as output proxies. The far most common procedure in ophthalmology is which ranged between some 260 and 1500 procedures for the departments in our data. Descriptive statistics of the data are given in Table 1.

TABLE 1 IN ABOUT HERE

7.2. Empirical results[5]

The results are divided into two categories. The first is the original DEA-estimates. The next is the bootstrap results, which are used to test scale efficiency and scale elasticity. All results are given in Table 2.

7.2.1. Original DEA-results

The scale efficiency results are shown in column 2. Columns 4 and 5 contains "lower" and "upper" bounds for the elasticity of scale, using $\delta = 0.01$ in the approximations described in section 3.3.

TABLE 2 IN ABOUT HERE

Ten departments were scale efficient. Eight departments were operating in an IRS region and the remaining 11 departments in a DRS region, according to the original results. We note that none of the departments located at large or small hospitals were scale efficient. Furthermore, the original DEA-results show that the five departments located at large hospitals operated in a region of decreasing returns, whereas all the departments at small hospitals operated in a region of increasing returns to scale.

The scale elasticity approximation results give 11 cases with lower bound less than one (but greater than zero) and eight approximations (upper bound) greater than one. We note that the scale elasticity results confirm the conclusions from the

[5]All results are obtained using PC versions of LINGO and GAUSS.

scale efficiency method. That is, all departments that have an elasticity lower bound less than one are categorized as operating in a DRS region using the scale efficiency method. Likewise, all departments with an elasticity upper bound estimate greater than one are estimated as operating in an IRS region. In seven cases the approximation is zero. This result corresponds to cases where the efficiency measure does not change as the inputs are increased. We conclude that those departments are either located at a kink to the left of a flat segment (i.e., $\hat{F}_o(\cdot|VRS) = 1$, e.g., observation D in *Figure 1*) or under a flat segment of the VRS technology surface (i.e., $\hat{F}_o(\cdot|VRS) > 1$, e.g., observation E in *Figure 1*). The scale elasticity approximation columns contain a number of missing values. The missing values occur because no (optimal) solution could be found for those cases.

7.2.2. Bootstrap results

First, hypotheses regarding scale efficiency are tested by one-sided confidence intervals using the nested hypotheses test procedure described in section 4.[6] The results obtained by the percentile method are not presented since *all* departments are scale efficient according to this method. Second, we compute one-sided 95% confidence intervals for the lower bound of the scale elasticity approximations. The upper bound ($\widehat{\varepsilon_o}$) is not bootstrapped due to the possibility of infeasible solutions in computing the approximation when inputs are decreased. Hence, the presented bootstrap confidence intervals for the approximated scale elasticities are intervals for the lower bound of the scale elasticity.

The first hypothesis $H_{01} : S_{o2} = 1$, is rejected, and hence DRS is concluded, in 11 cases for the BC_a-method. In the next step, the scale efficiency hypothesis cannot be rejected in seven cases for the BC_a method, which implies that scale efficiency is concluded. For five departments the hypothesis $H_{02} : S_{o1} = 1|S_{o2} = 1$ can be rejected using the BC_a-method. Hence five cases with increasing returns to scale is concluded. Note that no hypotheses can be rejected for either department using the percentile method and consequently scale efficiency is concluded for all departments. The hypothesis of either scale efficiency or IRS is rejected in favor of DRS for some departments in the BC_a-method (nos. 3, 6, 17, 23 and 27). In those cases the original results indicate either scale efficiency (3 and 6) or increasing returns to scale (17, 23, 27). Note that DRS is concluded for departments nos. 17, 23 and 27 although the original results indicate IRS. Next, in three cases the second hypothesis can not be rejected (dep. nos. 5, 8 and 9 are labeled as scale efficient), whereas the original results categorize them as scale inefficient.

Turning to the bootstrap results for the scale elasticity, note that since we bootstrap the lower bound of the scale elasticity approximation, these results can only be used to test a hypothesis of CRS/DRS against the IRS alternative. If the bootstrap confidence interval lower bound is greater than one, the hypothesis of a scale elasticity equal to,

[6] In these tests $\alpha = 0.025$.

or less than, one can be rejected. If, on the other hand, the bootstrap lower bound is below one the CRS/DRS-hypothesis cannot be rejected. It cannot, however, be concluded that the firm is operating under DRS since the upper bound may not be below one. The BC_a-method gives a lower bound significantly greater than one for five departments (nos. 2, 17, 23, 26 and 28), whereas this does not occur in any case using the percentile method. We note that three of the five departments at small hospitals have significant increasing returns to scale. The BC_a columns contain three cases with missing values. This occurs either when $\hat{F}_o^{*b} \leq \hat{F}_o \, \forall \, b$ or $\hat{F}_o^{*b} \geq \hat{F}_o \, \forall \, b$, which implies that the bias-correction term is not defined.

The proposed scale elasticity approximation is to be considered as somewhat tentative. It is, however, straightforward to compute the approximation, at least in the case of increases in inputs. In the case of decreases in inputs, the method is sensitive since infeasible solutions can occur in some cases. This result is perhaps not so surprising, given the formulation of the linear program used in the approximation.

8. Summary and conclusions

The purpose of the paper is to develop an easily implemented bootstrap procedure that allows testing different hypotheses of returns to scale properties for individual firms. Based on a model of the data generating process underlying the observations on inputs and outputs the bootstrap offers a possibility to statistically test the results from the original DEA-model. Scale efficiency is tested using a nested hypothesis approach whereas the scale elasticity testing is more straightforward. The test may, as in our application, lead to the rejection of a hypothesis of scale efficiency for a firm, even if the original results indicate that the firm operates at a scale efficient size.

An empirical application of the scale estimations and the bootstrap is provided based on primal production data from a sample of Swedish eye-care departments. About 40 percent of the departments were categorized as scale efficient according to the original results. The bootstrap results differ in some cases in that some departments are categorized as operating under decreasing returns to scale (DRS) in the scale efficiency test and increasing returns to scale (IRS) in the scale elasticity test. Using the nested scale efficiency test procedure, DRS and subsequently IRS was rejected in favor of scale efficiency for about 24 percent of the departments using the BC_a-method. Thus, for about a third of the departments, the original scale efficiency results are altered using the bootstrap test. Using the percentile method, however, the hypothesis of scale efficiency can not be rejected for any of the departments. We find that the original scale elasticity results lead to the same conclusions as the scale efficiency results. Using the bootstrap the categorization is similar, although some cases with scale elasticity larger than one in the original results are not significant.

9. References

Banker, R. D., (1984), "Estimating Most Productive Scale Size Using Data Envelopment Analysis", *European Journal of Operational Research*, Vol. 17, 35 - 44.

Banker, R. D., (1996), "Hypothesis Tests Using Data Envelopment Analysis", *Journal of Productivity Analysis*, Vol. 7, 139 - 159.

Banker, R. D., Charnes, A. and Cooper, W. W., (1984), "Some Models for Estimating Technical and Scale Inefficiencies in Data Envelopment Analysis", *Management Science*, Vol. 30, 1078 - 1092.

Banker, R. D. and Thrall, R.M., (1992), "Estimation of Returns to Scale using Data Envelopment Analysis", *European Journal of Operational Research*, Vol. 62, 74 - 84.

Charnes, A., Cooper, W. W. and Rhodes, E. (1978), "Measuring the Efficiency of Decision Making Units", *European Journal of Operational Research*, Vol. 2, No. 6, 429 - 444.

Efron, B., (1990), "More Efficient Bootstrap Computations", *Journal of the American Statistical Association*, Vol. 85, No. 409, 79 - 89.

Efron, B. and Tibshirani, R. J., (1993), "*An Introduction to the Bootstrap*", Monographs on Statistics and Applied Probability, No. 57, Chapman and Hall, New York, London.

Farrell, J., (1957), "The Measurement of Productive Efficiency", *Journal of the Royal Statistical Society*", Series A (General), Vol. 120, Part III, 253 - 281.

Federation of County Councils (1993). " *Yearly Statistics for County Councils 1994*". Federation of County Councils, Stockholm. (Statistisk rsbok för landsting 1994)

Frisch, R. (1965), "*Theory of Production*", D Riedel, Dordrecht.

Färe, R., (1988), "*Fundamentals of Production Theory*", Lecture Notes in Economics and Mathematical Systems, Vol. 311, Springer-Verlag, Berlin Heidelberg.

Färe, R., Grosskopf, S. and Lovell C. A. K., (1988), "Scale Elasticity and Scale Efficiency", *Journal of Institutional and Theoretical Economics*, Vol. 144, 721 - 729.

Färe, R., Grosskopf, S. and Lovell C. A. K., (1994), "*Production Frontiers*", Cambridge University Press, Cambridge.

Forsund, F., (1996), "On the Calculation of the Scale Elasticity in DEA Models", *Journal of Productivity Analysis*, Vol. 7, No. 2/3, 283 - 302.

Grosskopf, S., (1996), "Statistical Inference and Nonparametric Efficiency: A Selective Survey", *Journal of Productivity Analysis,* Vol 7, No. 2/3, 161-176.

Löthgren, M. and Tambour, M., (1995), "Bootstrapping DEA-based Efficiency Measures and Malmquist Indices - A Study of Swedish Eye-Care Service Provision", Working Paper Series in Economic and Finance, No. 78, Stockholm School of Economics.

Löthgren, M. and Tambour, M., (1996), "Alternative Approaches to Estimate Returns to Scale in DEA-Models", Working Paper Series in Economic and Finance, No. 90, Stockholm School of Economics.

Shephard, R.W., (1970), "*The Theory of Cost and Production Functions*", Princetopn University Press, Princeton.

Simar, L., (1996), "Aspects of Statistical Analysis in DEA-Type Frontier Models", *Journal of Productivity Analysis*, Vol. 7, 177 - 185.

Simar, L. and Wilson, P. W., (1995), "Sensitivity Analysis of Efficiency Scores: How to Bootstrap in Nonparametric Frontier Models", Discussion Paper 9503, Institut de Statistique, Universite Catholique de Louvain.

Wilson, P. W. and Simar L., (1995), "Bootstrap Estimation for Nonparametric Efficiency Estimates", Unpublished Working Paper, July 1995.

Figures

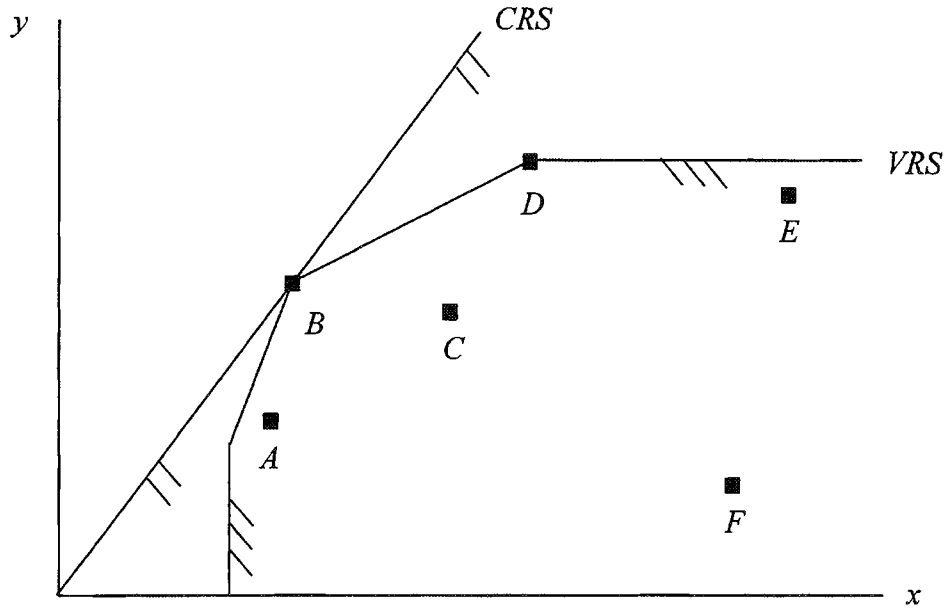

Figure 1: Single input- single output technology

Tables

Table 1: Descriptive statistics, inputs and outputs.

	Specialist physician	Other physicians	Number of beds	Cataract	Glaucoma	Squint	Visits
Max	180	78	22	1,509	212	172	43,830
Min	31	0	1	257	11	8	8,940
Mean (arithmetic)	89	23	9	783	55	44	23,855
Median	84	22	7	819	43	44	22,913
St dev	37	18	6	324	43	31	9,065

Table 2: Scale efficiency, original results (2), test results BC_a- (3) method. Scale elasticity approximation, lower (4) and upper (5) bound, original results. Test results, Percentile (6, 7) and BC_a- (8, 9) method. 1993-data.

		Scale efficiency[a]		Scale elasticity					
	Type[b]	Original	BC_a	Original		Percentile		BC_a	
				$\hat{\varepsilon}_o^-$	$\hat{\varepsilon}_o^+$	lower	upper	lower	upper
Dep	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
1	L	D	D	0.743	0.743	0	1.017	0.22	1.1
2	M	I	I	2.037	.	0.39	2.884	**1.785**	**3.164**
3	M	S	D	0.245	2.087	0.132	2.147	0.062	1.031
4	M	S	.	0	.	0	0	.	.
5	L	D	S	0	0.769	0	0.522	0	0.598
6	M	S	D	0	1.315	0	0.831	0	0.721
7	M	D	D	0.759	0.759	0.619	1.056	0.379	0.855
8	M	I	S	1.103	1.103	0.631	1.316	0.96	1.535
9	S	I	S	1.214	1.214	0.618	7.051	0.634	7.31
10	M	D	D	0.559	0.559	0.645	1.133	0.514	0.514
11	M	S	S	0.251	1.514	0.398	1.069	0.092	0.144
12	M	S	S	0.051	.	0.168	0.601	.	.
13	L	D	D	0	0.811	0	0.55	0	0.718
14	M	S	S	0.965	1.04	0.339	1.097	0.533	1.239
15	L	D	.	0.041	0.042	0	0.766	0	0.892
16	M	D	.	0.36	0.513	0.197	0.961	0.023	0.604
17	M	I	D	1.052	1.419	0.077	0.91	**1.248**	**1.605**
18	M	S	S	0.14	1.083	0.238	1.038	0.081	0.081
19	M	D	I	0	0.857	0.2	1.01	.	.
20	M	S	.	0	1.325	0	0.098	0	0.138
21	L	D	.	0	0.112	0	0.437	0	0.563
22	M	D	D	0.681	0.704	0	0.905	0.324	1.004
23	S	I	D	1.665	1.666	0.422	1.535	**1.729**	**1.898**
24	M	S	I	0.568	2.087	0.182	0.816	0.443	1.054
25	M	D	D	0.615	0.637	0.115	1.062	0.107	1.059
26	S	I	I	1.283	.	0.548	1.431	**1.042**	**1.593**
27	S	I	D	1.074	1.637	0.608	2.066	0.672	2.276
28	S	I	.	1.623	.	0.839	2.234	**1.063**	**2.642**
29	M	S	I	0.683	1.424	0	1.069	0.334	1.23

[a] S = Scale efficient, I = Increasing returns to scale, D = Decreasing returns to scale.

[b] Refers to the type of hospital each department is located at. L = large, M = medium size, S = small hospital.

III

Computationally Efficient Double Bootstrap Variance Estimation*

Sune Karlsson and Mickael Löthgren

Department of Economic Statistics
Stockholm School of Economics
P.O. Box 6501
S-113 83 Stockholm
Sweden
E-mail: stsk@hhs.se & stml@hhs.se
Fax: +46-8-348161.

Working Paper Series in Economics and Finance No. 151
January 1997

Abstract

The double bootstrap provides a useful tool for bootstrapping approximately pivotal quantities by using an "inner" bootstrap loop to estimate the variance. When the estimators are computationally intensive, the double bootstrap may become infeasible. We propose the use of a new variance estimator for the nonparametric bootstrap which effectively removes the requirement to perform the inner loop of the double bootstrap. Simulation results indicate that the proposed estimator produce bootstrap-t confidence intervals with coverage accuracy which replicates the coverage accuracy for the standard double bootstrap.

Key-Words: Bootstrap-t, Confidence intervals, Influence function, Nonparametric Bootstrap.
JEL-Classification: C14, C15.

*We thank Tor Jacobson and participants at the $(EC)^2$ conference in Florence, December 13 - 14, 1996, for helpful comments. The work of the first author was supported by a grant from the Swedish Research Council for the Humanities and Social sciences (HSFR).

1. Introduction

The bootstrap is frequently used to construct confidence intervals or conduct hypothesis tests in situations where the small sample distribution of the estimator is unknown or the distributional assumptions underlying the analysis are questionable. The accuracy of these procedures can often be improved by considering a studentized version of the statistic. It is, for example, well known that the coverage accuracy of one-sided bootstrap-t intervals is $O(n^{-1})$ compared to $O(n^{-1/2})$ for simple (non-pivotal) percentile intervals.

Studentizing, however, requires an estimate of the variance of the estimator in each bootstrap resample. When no (good) estimator of the variance is available, a bootstrap estimate of the variance can be obtained by doing a bootstrap on the bootstrap resample - a double bootstrap. With B_1 resamples in the outer bootstrap loop and B_2 resamples in the inner loop this requires calculations of order $B_1 B_2$ and can be extremely time consuming.

We propose an alternative double bootstrap-type variance estimator for use with the nonparametric bootstrap. The new estimator only requires calculations of order B_1, thus reducing the computational requirements considerably.

The variance estimates are easily obtained using standard regression methods or Jackknife estimators applied to the outer bootstrap resample replicates of the estimator. Simulation evidence are presented that indicate that bootstrap-t confidence intervals calculated using the proposed variance estimator have coverage accuracy that replicates results obtained from the standard double bootstrap variance estimator.

The paper unfolds as follows: In section 2 the bootstrap method is presented along with bootstrap confidence intervals. Section 3 presents the linear approximation of the estimator that underlies the new variance estimator presented in Section 4. Section 5 presents some theoretical properties of the estimator. Simulation results using several simulation designs and sample sizes are given in section 6. Section 7 gives a summary and some concluding remarks.

2. The Bootstrap Method

Consider an iid sample $\chi_n = \{X_1, \ldots, X_n\}$ of n observations on a univariate (or multivariate) random variable X with distribution F. Write the estimator of θ, the parameter of interest, as a function of the random sample as $\widehat{\theta}_n = \widehat{\theta}(\chi_n)$. In the nonparametric bootstrap a bootstrap resample $\chi_n^* = \{X_1^*, \ldots, X_n^*\}$ is a sample of size n drawn randomly, with replacement, from the sample χ_n. The observations in the resample are iid $X_i^* \sim \widehat{F}_n$, where $\widehat{F}_n$ denotes the empirical distribution function of the original sample. $\widehat{F}_n$ is given by $\widehat{F}_n(t) = \sum_{i=1}^n I(X_i \leq t)/n$, where $I(\cdot)$

denotes the standard indicator function. Denote repeated bootstrap resamples by χ_n^{*b}, $b = 1, ..., B_1$. Applying the estimator on a resample gives the bootstrap parameter estimate $\widehat{\theta}_n^{*b} = \widehat{\theta}\left(\chi_n^{*b}\right)$ as a function of the bootstrap resample.

2.1. Bootstrap Confidence Intervals

Under suitable regularity conditions, the empirical (bootstrap) distribution of $\left\{\widehat{\theta}_n^{*b},\ b = 1,\ \ldots, B_1\right\}$ converges to $F\left(\widehat{\theta}_n\right)$, the distribution of the estimator. This give rise to the simple bootstrap percentile confidence intervals for the parameter.

An approximate $(1 - 2\alpha)$ equal-tail double sided bootstrap percentile confidence interval for the parameter is given by $\left(\widehat{\theta}_n^{(\alpha)},\ \widehat{\theta}_n^{(1-\alpha)}\right)$, where $\widehat{\theta}_n^{(\alpha)}$ is the α-quantile of the B_1 resampled bootstrap estimates $\left\{\widehat{\theta}_n^{*b},\ b = 1,\ \ldots, B_1\right\}$ obtained as the solution to $\sum_{b=1}^{B_1} I\left(\widehat{\theta}_n^{*b} \leq \widehat{\theta}_n^{(\alpha)}\right) / (B_1 + 1) = \alpha$.

Bootstrap-t confidence intervals for the parameter are based on the distribution of the studentized (pivotal) statistic $T_n = n^{1/2}\left(\widehat{\theta}_n - \theta\right) / \widehat{\sigma}_n$, where $\widehat{\sigma}_n^2$ is an estimate of the variance of $n^{1/2}\widehat{\theta}_n$. In the bootstrap this, usually unknown, distribution is approximated by the distribution of the bootstrap version of T_n. The bootstrap statistic is given by $T_n^* = n^{1/2}\left(\widehat{\theta}_n^* - \widehat{\theta}_n\right) / \widehat{\sigma}_n^*$, where $\widehat{\sigma}_n^{*2}$ is an estimate of the variance of $n^{1/2}\widehat{\theta}_n^*$. Simulation based bootstrap estimates of the quantiles of T_n can be obtained from the empirical distribution of the B_1 resampled statistics $\left\{T_n^{*b} = n^{1/2}\left(\widehat{\theta}_n^{*b} - \widehat{\theta}_n\right) / \widehat{\sigma}_n^{*b},\ b = 1,\ \ldots, B_1\right\}$.

A $(1 - 2\alpha)$ equal-tail double-sided bootstrap-t confidence interval for the parameter is given by

$$\left(\widehat{\theta}_n - n_n^{-1/2}\widehat{\sigma}_n\widehat{t}_n^{(1-\alpha)},\ \widehat{\theta}_n - n_n^{-1/2}\widehat{\sigma}_n\widehat{t}_n^{(\alpha)}\right), \tag{2.1}$$

where the bootstrap quantile estimates of T_n are given by $\widehat{t}_n^{(\alpha)}$ such that $\sum_{b=1}^{B_1} I\left(T_n^{*b} \leq \widehat{t}_n^{(\alpha)}\right) / (B_1 + 1) = \alpha$.

For one-sided intervals and symmetric double-sided intervals, the coverage accuracy of the bootstrap-t intervals is of higher order than the percentile intervals. The coverage accuracy of one-sided bootstrap-t intervals is $O(n^{-1})$ compared to $O(n^{-1/2})$ for simple (non-pivotal) percentile intervals. Symmetric double-sided bootstrap-t intervals have coverage accuracy of $O(n^{-2})$ compared to $O(n^{-1})$ for percentile intervals. Furthermore, for equal tail double sided intervals the bootstrap-t and the percentile intervals both give the same coverage accuracy of $O(n^{-1})$. See Hall (1992, section 3.5 and 3.6) for further discussion on these issues.

In order to use the bootstrap-t method both the estimate $\widehat{\sigma}_n$ as well as repeated estimates of the variance $\widehat{\sigma}_n^{*2,b}$ are needed. If no analytical expressions exists for the variance estimate the bootstrap offers an alternative to produce both the estimate $\widehat{\sigma}_n$ as well as the estimates $\widehat{\sigma}_n^{*2,b}$.

2.2. Bootstrap variance estimation

Given the estimate $\hat{\theta}_n$, the bootstrap estimate of the variance $\sigma_n^2 = Var\left(n_n^{1/2}\widehat{\theta}\right)$ is given by the empirical variance of the B_1 bootstrapped estimates, i.e., $\widehat{\sigma}_n^2 = n\sum_{b=1}^{B_1}\left(\widehat{\theta}_n^{*b} - \overline{\widehat{\theta}}_n^*\right)^2/(B_1-1)$, where $\overline{\widehat{\theta}}^* = \sum_{b=1}^{B_1}\widehat{\theta}_n^{*b}/B_1$ is the average of the bootstrapped estimates.

The standard double bootstrap estimate of the variance $\sigma_n^{*2} = Var\left(n^{1/2}\widehat{\theta}_n^*\right)$, denoted $\widehat{\sigma}_n^{*2}$, is obtained analogously as the bootstrap estimate $\widehat{\sigma}_n^2$. The double bootstrap sample is a resample $\chi_n^{**} = \{X_1^{**}, \ldots, X_n^{**}\}$ of size n drawn randomly, with replacement, from the outer bootstrap resample χ_n^*. That is, $X_i^{**} \sim \widehat{F}_n^*$, where $\widehat{F}_n^*$ is the empirical distribution of the sample χ_n^*. The double bootstrap estimate of the parameter is given by $\widehat{\theta}_n^{**} = \widehat{\theta}(\chi_n^{**})$. B_2 resamples are drawn from χ_n^*, and double bootstrap estimates $\widehat{\theta}_n^{**b_2}$, $b_2 = 1, ..., B_2$, are calculated for each resample. The estimate of the bootstrap variance is then obtained as the empirical variance of the B_2 bootstrapped estimates, i.e., resampling from the resample b we have the estimate of the variance of $n^{1/2}\widehat{\theta}_n^{*b}$ as $\widehat{\sigma}_n^{*2,b} = n\sum_{b_2=1}^{B_2}\left(\widehat{\theta}_n^{**b_2} - \overline{\widehat{\theta}}_n^{**}\right)^2/(B_2-1)$, where $\overline{\widehat{\theta}}_n^{**} = \sum_{b_2=1}^{B_2}\widehat{\theta}_n^{**b_2}/B_2$ is the average of the double bootstrap estimates.

If the estimator is complex and involves, for example, iterations and optimization code the bootstrap-t confidence intervals may be infeasible to obtain due to time constraints. As an alternative to the double bootstrap estimator of the variance we propose a method that considerably cuts down the computer time needed for obtaining the desired variance estimates. The amount of calculations are cut down to B_1 instead of the B_1B_2 calculations required in a double bootstrap. Under this alternative approach bootstrap-t confidence intervals can be obtained using only the outer B_1 bootstrap results, avoiding the requirement for a double bootstrap.

3. Linear Approximations

The variance estimator we propose is based on a linear approximation of the estimator. Following Efron (1982) and Efron (1990) the bootstrap estimator can

be written in a linear ANOVA form as

$$\widehat{\theta}_n^* = \mu + \mathbf{p}^{*\prime}\mathbf{a} + u\,, \tag{3.1}$$

where $\mu = E\left\{\widehat{\theta}_n^* | \chi_n\right\}$ and $a_i = n\left[E\left\{\widehat{\theta}_n^* | X_1^* = x_i\right\} - \mu\right]$.

The random element is the resampling vector $\mathbf{p}^* = (p_1^*, \ldots, p_n^*)'$, where $p_i^* = \#\left\{X_j^* = X_i\right\}/n$ and $\sum_{i=1}^n p_i^* = 1$. The resampling vector $\mathbf{p}^*$ follows a rescaled multinomial distribution and is distributed as $\mathbf{p}^* \sim \left(\mathbf{p}^0, \frac{1}{n}\left(\frac{1}{n}\mathbf{I}_n - \mathbf{p}^0\mathbf{p}^{0\prime}\right)\right)$, where $\mathbf{p}^0 = \frac{1}{n}\mathbf{1}_n$ represents the original sample. $\mathbf{a}$ is called the bootstrap influence function and represents the linear influence of the resampling vector $\mathbf{p}^*$ on the bootstrap estimate $\widehat{\theta}_n^*$. The remainder term u represents all terms of order two and higher of $\mathbf{p}^*$.

Efron (1990) state the following properties of the decomposition in (3.1):

P.1. The elements of $\mathbf{a}$ sum to zero, i.e. $\mathbf{a}'\mathbf{1}_n = 0$.

P.2. The term u contains higher order terms of the resampling vector $\mathbf{p}^*$ and is uncorrelated with any linear function of $\mathbf{p}^*$, i.e. $E\left\{(\mathbf{c}'\mathbf{p}^*)\, u\right\} = 0$ for any n-vector $\mathbf{c}$. In particular, u is uncorrelated with $\mathbf{p}^{*\prime}\mathbf{a}$.

Using these properties the ideal ($B_1 \to \infty$) bootstrap estimate of the variance of $n^{1/2}\hat{\theta}_n$ is given by

$$\sigma_{BOOT}^2 = nVar\left(\widehat{\theta}_n^*\right) = n\mathbf{a}'Var\left(\mathbf{p}^*\right)\mathbf{a} + nVar\left(u\right) = \frac{\mathbf{a}'\mathbf{a}}{n} + O\left(n^{-1}\right) \tag{3.2}$$

The last equality follows from the fact that $Var\left(u\right) = O\left(n^{-2}\right)$ (Efron 1982, p. 23). The variance in (3.2) is similar to the delta method approximation of the variance, see e.g., Shao & Tu (1995).

For the double bootstrap an analogous decomposition to (3.1) exists. This can be written as

$$\widehat{\theta}_n^{**} = \mu^* + \mathbf{p}^{**\prime}\mathbf{a}^* + u^*, \tag{3.3}$$

where the double bootstrap influence vector $\mathbf{a}^*$, analogously to property P.1, satisfies $\mathbf{p}^{*\prime}\mathbf{a}^* = 0$. The resampling vector $\mathbf{p}^{**}$ has expectation, $E(\mathbf{p}^{**}) = \mathbf{p}^*$ and variance $Var(\mathbf{p}^{**}) = \frac{1}{n}\left(diag\left(\mathbf{p}^*\right) - \mathbf{p}^*\mathbf{p}^{*\prime}\right)$.

Following similar arguments as above, the ideal ($B_2 \to \infty$) double bootstrap estimate of the variance of the bootstrap replicate is given by

$$\sigma_{BOOT}^{*2} = nVar\left(\widehat{\theta}_n^{**}\right) = n\mathbf{a}^{*\prime}Var\left(\mathbf{p}^{**}\right)\mathbf{a}^* + nVar\left(u^*\right) = \mathbf{a}^{*\prime}diag\left(\mathbf{p}^*\right)\mathbf{a}^* + O\left(n^{-1}\right). \tag{3.4}$$

4. A New Double Bootstrap Variance Estimator

If the double bootstrap influence function $\mathbf{a}^*$ is known, the double bootstrap variance σ^{*2}_{BOOT} could be obtained directly from (3.4). As an alternative, an estimate of the double bootstrap influence function could be obtained from the linear approximation in (3.3) and double bootstrap replicates of the estimator. This requires, however, that a full double bootstrap is performed.

As an alternative, we propose an estimator of the bootstrap variance based on the variance expression (3.4) that does not requires a full double bootstrap procedure. The idea is simply to substitute the bootstrap influence vector $\mathbf{a}$ for the double bootstrap influence vector $\mathbf{a}^*$ in (3.4), that is,

$$\sigma^{*2}_{BOOT} \approx \mathbf{a}' diag\,(\mathbf{p}^*)\,\mathbf{a} = \sum_{i=1}^{n} p_i^* a_i^2. \tag{4.1}$$

Since the bootstrap influence function $\mathbf{a}$ is unknown some estimate is needed in (4.1). Plugging in an estimate $\widehat{\mathbf{a}}$ of $\mathbf{a}$, the proposed variance estimator is given by

$$v^* = \sum_{i=1}^{n} p_i^* \widehat{a}_i^2. \tag{4.2}$$

The estimator (4.2) is of the same structure as the variance expression in (3.2), but here the resampling vector $\mathbf{p}^*$ is used instead of $\mathbf{p}^0$ in the weighting of the bootstrap influence function $\mathbf{a}$. Using $\mathbf{p}^0$, we obtain the corresponding estimate of the variance of $n_n^{1/2}\widehat{\theta}_n$ as $v = \widehat{\mathbf{a}}'\widehat{\mathbf{a}}/n$.

The estimators of $\mathbf{a}$ discussed below at most require that the bootstrap table $\mathbf{P}^*$, containing the resampling vectors $\mathbf{p}^{*b}$, and the bootstrap estimates $\widehat{\theta}_n^{*b}$ are available. The estimator v^* can thus be calculated without any reference to the double bootstrap data and the inner bootstrap loop is not required.

4.1. Estimation of the bootstrap influence function

We present two estimators of the bootstrap influence function $\mathbf{a}$. They differ mainly in memory and computational requirements. The OLS-estimator requires that the bootstrap table is saved, increasing the memory requirements of the algorithm, and is computed after performing the outer bootstrap loop. The Jackknife estimator, on the other hand, can be computed before the outer bootstrap loop and allows on-line calculation of the variance estimator. Compared to the OLS-estimator it does require that the estimator $\widehat{\theta}_n$ is computed an additional n times, which can be time consuming in some applications.

4.1.1. OLS-estimation

Write the B_1 bootstrap replicates of the statistic in matrix form as

$$\widehat{\boldsymbol{\theta}}^* = \mu \mathbf{1}_{B_1} + \mathbf{P}^* \mathbf{a} + \mathbf{u}, \tag{4.3}$$

where $\mathbf{P}^*$ is a $(B_1 \times n)$ matrix of resampling vectors $\mathbf{p}^{*b}$, $\widehat{\boldsymbol{\theta}}^*$ is the B_1-vector of bootstrapped statistics and $\mathbf{u}$ is the B_1-vector of orthogonal error terms in the regression model.

Following, for example, Efron (1990) and Hesterberg (1995) the OLS-estimator of $\mathbf{a}$ and μ is obtained as

$$\widehat{\mu} = \frac{\mathbf{1}_n' \widehat{\mathbf{d}}}{n} \text{ and } \widehat{\mathbf{a}}_{OLS} = \widehat{\mathbf{d}} - \mathbf{1}_n \widehat{\mu}, \tag{4.4}$$

where $\widehat{\mathbf{d}} = \left(\mathbf{P}^{*\prime}\mathbf{P}^*\right)^{-1} \mathbf{P}^{*\prime} \boldsymbol{\theta}\left(\mathbf{P}^*\right)$ is the OLS estimator of $\mathbf{d}$ in the approximation $\boldsymbol{\theta}\left(\mathbf{P}^*\right) = \mathbf{P}^*\mathbf{d} + \mathbf{u}$ obtained by deleting μ in (4.3).

4.1.2. Jackknife estimation

The Jackknife estimate of $\mathbf{a}$, given by $\widehat{\mathbf{a}}_J$, with elements

$$\widehat{a}_{J,i} = (n-1)\left(\overline{\theta}_{\cdot} - \widehat{\theta}\left(\mathbf{p}_{-i}^0\right)\right), i = 1, ..., n, \tag{4.5}$$

where $\mathbf{p}_{-i}^0 = (1,\ 1,\ ...,\ 1,\ 0,\ 1,\ ...,\ 1)/(n-1)$, and $\overline{\theta}_{\cdot} = \sum_{i=1}^n \widehat{\theta}\left(\mathbf{p}_{-i}^0\right)/n$ is the average of the n delete-1 jackknife estimates of the statistic.

This way of constructing the jackknife estimate guarantees that the restriction $\mathbf{1}_n' \widehat{\mathbf{a}} = 0$ is satisfied.

5. Theoretical Properties

The improved coverage accuracy of bootstrap-t confidence intervals discussed in section 2.1 requires that the studentized statistic is asymptotically pivotal, i.e. that the asymptotic distribution is free of nuisance parameter (Beran 1987, Hall 1992).

When discussing the asymptotic properties of the studentized statistic, $T_n = n^{1/2}\left(\widehat{\theta}_n - \theta\right)/\sqrt{v}$, the influence function of the estimator $\widehat{\theta}$ is useful. Given that the parameter θ can be written as a functional of the distribution function, i.e., $\theta = \theta(F)$, the influence function $IF(x; F)$ of $\theta(F)$ is defined as

$$IF(x; F) = \lim_{\varepsilon \to 0} \frac{\theta\left((1-\varepsilon)F + \varepsilon\delta_x\right) - \theta(F)}{\varepsilon} \tag{5.1}$$

where δ_x is a unit probability mass on the point x. The influence function is also known as the von Mises derivative.

The influence function can be used to derive a von Mises expansion of the functional statistic, see, e.g., Fernholz (1983, chapter 2). Given the functional representation of the parameter, the estimator can be defined as the functional evaluated under the empirical distribution, i.e., $\widehat{\theta}_n = \theta\left(\widehat{F}_n\right)$. It is well known that a linear von Mises expansion of the estimator $\widehat{\theta}_n$ can be expressed as

$$\widehat{\theta}_n = \theta(F) + \frac{1}{n}\sum_{i=1}^{n} IF(x_i; F) + Rem\left(\widehat{F}_n - F\right). \tag{5.2}$$

The remainder term in the expansion, $Rem\left(\widehat{F}_n - F\right)$, contains higher order terms of the functional statistic.

The bootstrap linear approximation in (3.1) is based on the von Mises expansion. Using a bootstrap analogy and substituting the bootstrap empirical distribution $\widehat{F}_n^*$ for the empirical distribution $\widehat{F}_n$ and the empirical for the true distribution F in (5.2) results in the bootstrap linear approximation. This reveals a close connection between the influence function (5.1) and the bootstrap influence function. The bootstrap influence function is given by $a_i = IF\left(x_i, \widehat{F}_n\right)$ and we have from Efron (1982, p. 24) that $a_i \underset{n}{\rightarrow} IF(x_i, F)$.

Under quite general regularity conditions on the differentiability of the functional statistic $\theta(F)$, the estimator $\widehat{\theta}_n$ is asymptotically normally distributed as

$$\sqrt{n}\left(\widehat{\theta}_n - \theta\right) \xrightarrow{d} N(0, V(\widehat{\theta})) \tag{5.3}$$

with asymptotic variance $V(\widehat{\theta}) = \int IF(x;F)^2 dF < \infty$. Fernholz (1983) gives different sets of sufficient regularity conditions, the most restrictive being that a functional induced by $\theta(F)$ is Hadamard differentiable (Fernholz (1983, Theorem 4.4.2)).

Fernholz (1983, chapter 5) shows that the regularity condition holds for a wide range of estimators such as maximum likelihood type estimators (M-estimators), linear functions of order statistics (L-estimators) and statistical estimators based on rank statistics (R-estimators). Using the asymptotic normality the following theorem establishes a similar asymptotic distribution result for the studentized statistic T_n obtained using our proposed variance estimator v.

Theorem 5.1. *The studentized statistic,* $T_n = n^{1/2}\left(\widehat{\theta}_n - \theta\right)/\sqrt{v}$, *where* $v = \frac{1}{n}\sum_{i=1}^{n} \widehat{a}_i^2$, *is asymptotically distributed as* $N(0,1)$ *if* $\sqrt{n}\left(\widehat{\theta}_n - \theta\right) \xrightarrow{d} N(0, V(\widehat{\theta}))$. *This result holds for both the OLS- and the Jackknife estimator* $\widehat{\mathbf{a}}$ *of the bootstrap influence function* $\mathbf{a}$.

Proof. From property P.2 of the linear approximation it follows that the OLS estimator of the bootstrap influence vector is root-B_1 consistent. Since $a_i \underset{n}{\rightarrow} IF(x_i, F)$, we have $\widehat{a}_{OLS,i} \underset{n}{\rightarrow} IF(x_i, F)$, if B_1 tends to infinity faster than n.

Let $\varepsilon = 1/n$ and consider the distribution functions, $\widehat{F}_{-i}$, with observation i deleted. We can then write the Jackknife estimator of the influence vector element $\widehat{a}_{J,i}$ as,

$$\widehat{a}_{J,i} = (n-1)\left(\overline{\theta}_{\cdot} - \widehat{\theta}\left(\mathbf{p}^0_{-i}\right)\right) = \frac{n-1}{n}\frac{\theta\left((1-1/n)\,\widehat{F}_{-i} + 1/n\delta_{x_i}\right) - \theta\left(\widehat{F}_{-i}\right)}{1/n}, \tag{5.4}$$

which is approximately equal to the ratio in (5.1). From this it is clear that $\widehat{a}_{J,i} \underset{n}{\rightarrow} IF(x_i, F)$ and the Jackknife estimator of the bootstrap influence function is consistent.

That is, both the OLS- and the Jackknife estimator produce consistent estimates of the bootstrap influence function vector $\mathbf{a}$. It follows that, $v = \frac{1}{n}\sum_{i=1}^{n}\widehat{a}_i^2 \underset{n}{\rightarrow} V(\widehat{\theta}) = \int IF(x;F)^2 dF$ and $T_n \xrightarrow{d} N(0,1)$. ■

Hence, if the statistic satisfies the regularity conditions necessary for (5.3) to hold, the proposed variance estimator v gives an asymptotically normally distributed studentized statistic. Furthermore, the analysis in Beran (1987, section 4) applies when Theorem 5.1 holds. That is, the bootstrap-t confidence intervals using v and v^* have improved coverage accuracy compared to confidence intervals based on asymptotic theory and the simple percentile bootstrap confidence intervals. In addition to giving (asymptotically) superior coverage, the use of the proposed bootstrap-t intervals is virtually costless compared to percentile intervals and asymptotic theory intervals. The improved coverage accuracy of the bootstrap-t intervals can, of course, be achieved using any consistent estimator of $V(\widehat{\theta})$. In particular we could use the standard bootstrap estimate, but this requires that the inner bootstrap loop is actually carried out in order to obtain the double bootstrap variance estimate $\widehat{\sigma}_n^{*2}$, increasing the computational effort from $O(B_1)$ to $O(B_1 B_2)$.

6. Small Sample Properties

The analysis in the previous section tell us that the bootstrap-t confidence intervals using the proposed variance estimator perform as well as the as the standard bootstrap variance estimator. That is, the intervals have the same asymptotic coverage accuracy, $O\left(n^{-1}\right)$ for one-sided and equal-tail double-sided intervals and $O(n^{-2})$ for symmetric double-sided intervals. The small sample properties might, however, differ substantially and it is not clear which type of interval performs

best. To investigate the small sample properties we conduct a series of small Monte Carlo studies[1].

In each of the examples 1000 samples are generated from the underlying population distribution and bootstrap-t confidence intervals are calculated for the parameters of interest using the double bootstrap estimate of the standard error and the estimator v^* suggested in this paper. The number of bootstrap replications are 999 for the outer loop (B_1) and 100 for the inner, double bootstrap, loop (B_2). More precisely, the algorithm is as follows:

1. For $\alpha = \{0.025, 0.05, 0.1, 0.9, 0.95, 0.975\}$ quantiles $\widehat{t}_n^{(\alpha)}$ for the studentized statistic $T_n^* = n^{1/2}\left(\hat{\theta}_n^* - \hat{\theta}_n\right)/s_n^*$, where $s_n^{*2} = \{v_n^*, \hat{\sigma}_n^{*2}\}$, are obtained for each of the 1000 replicates as the $\alpha(B_1 + 1)$ order statistic of the set of bootstrap replicates $\left\{T_n^{*b}, b = 1, \ldots, B_1\right\}$.

2. Confidence intervals are obtained as in (2.1). The variance estimate used is consistent with the method used to obtain the variance estimates s_n^{*2} in the quantile estimation. That is, v_n and $\widehat{\sigma}_n^2$, respectively.

The variance estimates v and v^* were calculated using both the OLS and Jackknife estimates of the influence vector $\mathbf{a}$. The coverage of the intervals are virtually identical for the two estimators and we only report the results for the intervals where v and v^* are based on the OLS-estimates of $\mathbf{a}$.

The examples are simple in the sense that the estimators are easy to calculate, thus facilitating the Monte Carlo simulations. They are, however, realistic in the sense that they consider features of the distributions of estimators, i.e. symmetric, skewed and fat-tailed distributions which are common in applied work.

Example 6.1. *Symmetric mixture of normals. Samples of* $n = 10, 30$ *and* 100 *are generated where* X_i *is drawn from a* $N(1,1)$ *with probability* 0.6 *and a* $N(1,4)$ *with probability* 0.4. *Bootstrap-t confidence intervals for the mean and variance are calculated based on the natural estimators* $\overline{x} = \sum x_i/n$ *and* $s^2 = \sum(x_i - \overline{x})^2/(n-1)$. *The observed coverage levels of the confidence intervals are reported in Table 6.1.*

Both the double bootstrap (DB) and the confidence intervals based on the estimator v^* (OLS) work very well for the mean. This is not surprising since the mean is a linear statistic. The OLS intervals have lower coverage than the DB intervals for small n but for large n the coverage are about the same.

The picture is different for the variance. Here the lower one-sided confidence intervals have too high coverage and the upper one-sided confidence intervals have

[1] The simuations are performed using GAUSS version 3.2.15. The program code can be found at http://www.hhs.se/research/wpecofi/97/wp0151.htm.

Table 6.1: Coverage accuracy for bootstrap-t confidence intervals. Symmetric mixture of normals (Example 1)

Method	n	One-sided intervals: Percentile						Double-sided intervals: Level		
		0.025	0.05	0.10	0.90	0.95	0.975	0.95	0.90	0.80
					Mean					
DB	10	0.973	0.941	0.890	0.889	0.944	0.971	0.944	0.885	0.779
	30	0.977	0.944	0.893	0.899	0.952	0.975	0.952	0.896	0.792
	100	0.976	0.948	0.905	0.898	0.954	0.978	0.954	0.902	0.803
OLS	10	0.940	0.919	0.867	0.866	0.915	0.942	0.882	0.834	0.733
	30	0.969	0.937	0.887	0.892	0.943	0.963	0.932	0.880	0.779
	100	0.971	0.947	0.896	0.889	0.952	0.997	0.948	0.899	0.785
					Variance					
DB	10	0.940	0.889	0.804	0.906	0.944	0.971	0.911	0.833	0.710
	30	0.940	0.865	0.765	0.951	0.975	0.990	0.930	0.840	0.716
	100	0.852	0.776	0.648	0.982	0.994	0.997	0.849	0.770	0.630
OLS	10	0.908	0.847	0.760	0.854	0.876	0.883	0.791	0.723	0.614
	30	0.905	0.841	0.738	0.945	0.966	0.977	0.882	0.807	0.683
	100	0.830	0.761	0.636	0.980	0.992	0.996	0.826	0.753	0.616

too low coverage. This can be attributed to the variance having a skew distribution. The confidence intervals based on the double bootstrap have slightly better coverage than the intervals based on v^* but the differences are small. The OLS intervals for both the mean and the variance have lower coverage accuracy than the double bootstrap intervals. But this difference diminishes with an increased sample size.

Example 6.2. *Skew mixture of normals. This is identical to Example 1 except that X_i is drawn from a $N(1,1)$ with probability 0.6 and a $N(3,4)$ with probability 0.4. The results are given in Table 6.2.*

In this example the distribution of the mean is skewed but the confidence intervals for the mean perform as in *Example 1.* For the one-sided intervals a clear convergence tendency to the nominal level can be seen for both the DB and the OLS intervals. Both types of intervals undercover for small n even if this is most clear for the OLS intervals. The double-sided DB intervals have the nominal coverage level even for small samples, whereas the OLS intervals undercovers for $n = 10$ but converge quickly and for $n = 30$ the coverage is close to the nominal level.

Table 6.2: Coverage accuracy for bootstrap-t confidence intervals. Skew mixture of normals (Example 2)

Method	n	One-sided intervals: Percentile						Double-sided intervals: Level		
		0.025	0.05	0.1	0.9	0.95	0.975	0.95	0.9	0.8
					Mean					
DB	10	0.981	0.964	0.914	0.891	0.932	0.960	0.941	0.896	0.805
	30	0.986	0.962	0.914	0.903	0.957	0.973	0.959	0.919	0.817
	100	0.979	0.949	0.900	0.918	0.961	0.980	0.959	0.910	0.818
OLS	10	0.963	0.943	0.896	0.868	0.906	0.929	0.892	0.849	0.764
	30	0.978	0.955	0.906	0.887	0.948	0.967	0.945	0.903	0.793
	100	0.974	0.948	0.897	0.914	0.957	0.978	0.952	0.905	0.811
					Variance					
DB	10	0.965	0.924	0.853	0.906	0.939	0.969	0.934	0.863	0.759
	30	0.980	0.958	0.908	0.914	0.949	0.970	0.950	0.907	0.822
	100	0.987	0.955	0.892	0.901	0.953	0.970	0.957	0.908	0.793
OLS	10	0.937	0.893	0.826	0.840	0.870	0.884	0.821	0.763	0.666
	30	0.966	0.934	0.876	0.904	0.935	0.954	0.920	0.869	0.780
	100	0.974	0.937	0.884	0.899	0.951	0.968	0.942	0.888	0.783

For the variance, the intervals do not show the same pattern as in *Example 1.* The coverage for both DB and OLS one-sided intervals are near the nominal coverage level. For small n both interval types undercovers for both lower and upper intervals. For the double sided intervals, the same pattern can be seen although the convergence seems slower for the OLS intervals. For $n = 100$ DB gives the nominal level, but the OLS intervals performs slightly worse and undercovers.

Example 6.3. *Exponential Distribution. Samples of* $n = 10, 30$ *and* 100 *are generated from an exponential distribution with parameter* $\lambda = 1$. *For each sample we estimate the mean, variance and* λ *(as* $1/\overline{x}$*) and calculate confidence intervals. The observed coverage levels of these confidence intervals are reported in Table 6.3.*

As in *Example 2* the sample mean has a skewed distribution and the results for the confidence intervals for the mean are similar with slightly worse coverage for the upper one-sided intervals. The performance for the double-sided intervals are similar. Both DB and OLS undercover but the difference to the nominal level is more pronounced for OLS. A clear convergence tendency to the nominal levels can be seen for both DB and OLS.

Table 6.3: Coverage accuracy for bootstrap-t confidence intervals. Exponential distribution (Example 3)

Method	n	One-sided intervals: Percentile						Double-sided: intervals: Level		
		0.025	0.05	0.1	0.9	0.95	0.975	0.95	0.9	0.8
					Mean					
DB	10	0.987	0.976	0.933	0.850	0.911	0.953	0.940	0.887	0.783
	30	0.983	0.969	0.913	0.908	0.941	0.968	0.951	0.910	0.821
	100	0.975	0.944	0.894	0.882	0.939	0.963	0.938	0.883	0.776
OLS	10	0.978	0.954	0.918	0.811	0.857	0.889	0.867	0.811	0.729
	30	0.976	0.963	0.908	0.901	0.937	0.961	0.937	0.900	0.809
	100	0.970	0.943	0.892	0.880	0.932	0.957	0.927	0.875	0.772
					Variance					
DB	10	0.997	0.983	0.943	0.846	0.907	0.951	0.948	0.890	0.789
	30	0.988	0.971	0.918	0.910	0.941	0.970	0.958	0.912	0.828
	100	0.976	0.947	0.895	0.883	0.937	0.963	0.939	0.884	0.778
OLS	10	0.999	0.992	0.976	0.741	0.782	0.798	0.797	0.774	0.717
	30	1.000	0.988	0.960	0.872	0.904	0.926	0.926	0.892	0.832
	100	0.994	0.969	0.917	0.861	0.897	0.935	0.929	0.866	0.778
					λ					
DB	10	0.959	0.923	0.861	0.925	0.969	0.984	0.943	0.892	0.786
	30	0.973	0.943	0.910	0.913	0.968	0.981	0.954	0.911	0.823
	100	0.963	0.939	0.878	0.896	0.943	0.974	0.937	0.882	0.774
OLS	10	0.995	0.986	0.964	0.790	0.817	0.832	0.827	0.803	0.754
	30	1.000	0.989	0.967	0.818	0.866	0.881	0.881	0.855	0.785
	100	0.995	0.980	0.936	0.843	0.886	0.913	0.908	0.866	0.779

For the variance the DB intervals do better than the double bootstrap for one-sided intervals. For the lower (upper) one-sided intervals both intervals over (under) covers for small samples but the DB intervals converge to the nominal levels, whereas the performance of the OLS intervals seems to depend on the level of the interval. The smaller the level (0.95 and 0.9) the better the performance. This indicate that one-sided OLS intervals based on skewed statistics perform worse in the tails. The double sided intervals under covers for both methods and all sample sizes. As in other cases, the OLS intervals perform worse than DB, but the difference diminishes. For $n = 100$ the performance of the two methods are approximately equal.

The behavior of the OLS intervals for λ is about the same as for the variance. For the DB intervals, on the other hand, the behavior of the one-sided intervals

is the mirror image as compared to the variance intervals. A convergence to the nominal levels is seen but the lower (upper) one-sided intervals under (over) covers. The double sided intervals for both DB and OLS reveals that both intervals under covers for small n, but the difference to the nominal level decreases with increased n. The performance of DB is slightly better than for OLS but for smaller nominal levels (0.90 or 0.80) the difference is not significant.

7. Summary and Concluding Remarks

This paper has presented a new variance estimator which can be used instead of the computationally intensive nonparametric double bootstrap variance estimator. The estimator can be used to obtain bootstrap-t confidence intervals. This type of interval is based on a studentized version of the statistic where an estimate of the bootstrap variance is needed. If no analytic expression exists for the variance estimates, the proposed estimator gives an efficient way to obtain bootstrap-t confidence intervals avoiding the need to perform a full nested double bootstrap. With B_1 resamples in the outer bootstrap loop and B_2 resamples in the inner loop a double bootstrap requires calculations of order $B_1 B_2$. This can be extremely time consuming for complex estimators. Our alternative double bootstrap-type variance estimator only requires calculations of order B_1, thus reducing the computational requirements considerably.

The variance estimator is given by a weighted average of squared influence functions, with weights given by the resample frequencies in the bootstrap resample. The influence functions are given by the first order terms in a Taylor-like expansion of the estimator. Variance estimates are easily obtained using standard OLS or Jackknife estimators applied to the outer bootstrap resample replicates of the estimator.

Simulation results indicate that the bootstrap-t confidence intervals based on the proposed variance estimator perform as well as the double bootstrap based intervals. Generally, the intervals have slightly lower coverage accuracy than the double bootstrap intervals, but the difference decreases with increasing sample size. In general, the new estimator performs well when the double bootstrap does well and fails when the double bootstrap fails.

We note that both the double bootstrap intervals and the intervals obtained using our variance estimator perform poorly in situations where the estimator has a skewed distribution. It is our experience that the bootstrap-t does not handle this type of situation very well. This failure may however be remedied by using additional prepivoting methods as discussed in Beran (1987) and Vinod (1995). The additional prepivoting would require an extra bootstrap level and further increased computational burden. The proposed variance estimator could,

however, once again remove the need to perform the innermost bootstrap loop. We leave the study of the effectiveness of this approach to future research.

References

Beran, R. (1987), 'Prepivoting to reduce level error of confidence sets', *Biometrika* **74**, 457–468.

Efron, B. (1982), *The Jackknife, the Bootstrap and other Resampling Plans*, CBMS-NSF Regional Conference Series in Applied Mathematics, Society for Industrial and Applied Mathematics.

Efron, B. (1990), 'More efficient bootstrap computations', *Journal of the American Statistical Association* **85**, 78–89.

Fernholz, L. T. (1983), *von Mises Calculus for Functional Statistics*, Vol. 19 of *Lecture Notes in Statistics*, Springer-Verlag.

Hall, P. (1992), *The Bootstrap and Edgeworth Expansion*, Springer-Verlag.

Hesterberg, T. (1995), 'Tail-specific linear approximations for efficient bootstrap simulations', *Journal of Computational and Graphical Statistics* **4**, 113–133.

Shao, J. & Tu, D. (1995), *The Jackknife and Bootstrap*, Springer-Verlag.

Vinod, H. D. (1995), 'Double bootstrap for shrinkage estimators', *Journal of Econometrics* **68**, 287 – 302.

IV

Generalized Stochastic Frontier Production Models

by

Mickael Löthgren*

Stockholm School of Economics
Working Paper Series in Economics and Finance No. 149
December 1996

Abstract

This paper proposes a generalization of the single output stochastic production frontier model. The generalization allows estimation of production frontiers and distance functions (technical efficiency) for multiple output technologies using cross section or panel data on inputs and outputs.
Key Words: Distance function, Panel data, Stochastic frontier models, Technical efficiency.
JEL-Classification: C21, C23, D24.

1 Introduction

The stochastic production frontier model introduced by Aigner et al. (1977) deals with estimation of production frontiers and technical efficiency using cross section firm data. Schmidt and Sickles (1984) generalized the model specification to panel data frontier production models. The basic frontier models have been extended in various ways regarding both specification and estimation. See, e.g., Greene (1993) for a recent survey of the frontier model literature.

The established frontier production models are specified for single output technologies. For general multiple output technologies the standard approach to frontier and efficiency estimation is to impose behavioral assumptions of either cost minimization, revenue- or profit maximization and estimate a frontier cost, revenue or profit function. This may, however, be infeasible if cost and/or price data are unavailable or unreliable as is often the case in for example public production. In this case, the standard frontier models cannot be used to estimate frontiers and efficiency

*Department of Economic Statistics, P.O. Box 6501, S-113 83 Stockholm, Sweden. e-mail: stml@hhs.se. Fax: +46 - 8 - 34 81 61. I thank Tor Jacobson, Magnus Tambour and Anders Westlund for helpful comments.

without imposing some (arbitrary) aggregation of the outputs (or inputs) to a single dimension. This drawback with the established frontier models is addressed in this paper.

The purpose is to propose a generalization of the stochastic production frontier model to incorporate general multiple output technologies. The generalized frontier model allows estimation of production frontiers and efficiency using data on inputs and outputs without the need to rely on behavioral assumptions of the firms. The proposed generalization is based on the foundations on stochastic production theory developed in Krug (1976). This work extended the axiomatic production theory developed by Shephard (1970), where distance functions are used as scalar valued representations of general technologies. The output distance function represents a measure of the radial distance from the observed output to the production frontier and can be interpreted as a radial measure of technical efficiency.

The generalized production frontier model is based on a polar coordinate representation of the output vector. A stochastic ray frontier production function is defined where the Euclidean norm of the output vector is specified as a function of the inputs and the direction of the output vector, represented by the polar coordinate angles. The stochastic ray production function allows estimation of production frontiers and firm-specific distance functions (technical efficiencies) using estimation techniques developed for the standard (single output) production frontier models.

The paper unfolds as follows: Section 2 gives the basic stochastic production theory and definitions of the stochastic output distance function and the generalized ray production function. In Section 3 the generalized production frontier model is derived. Section 4 concludes the paper with a summary.

2 Production Theory

Consider a production technology where inputs $x \in R_+^d$ are used to produce outputs $y \in R_+^p$. Following the foundations on stochastic production theory developed by Krug (1976) let $(\Omega, \mathcal{F}, \mathcal{P})$ denote a probability space. Ω denotes a collection of states of the world, $\mathcal{F}$ is a Borel field and $\mathcal{P}$ is a probability measure on the $\sigma-$algebra $(\Omega, \mathcal{F})$. The stochastic technology can be described by the stochastic output set $P(x, \omega)$ defined as

$$P(x, \omega) = \{y \in R_+^p : x \text{ can produce } y \text{ at state of the world } \omega \in \Omega\}. \quad (1)$$

The technology is assumed to satisfy a set of basic assumptions/axioms discussed in Krug (1976). Shephard (1970) and Färe (1988) present similar axioms for non-stochastic technologies. Convexity of $P(x, \omega)$ for all x and $\omega \in \Omega$ and disposability (monotonicity) of inputs and outputs are the most common axioms. Free output disposability states that $y \in P(x, \omega), y' \leq y \Rightarrow y' \in P(x, \omega)$ for all $\omega \in \Omega$ and weak disposability states that $y \in P(x, \omega), \gamma \in [0, 1] \Rightarrow \gamma y \in P(x, \omega)$ for all $\omega \in \Omega$.

Provided the stochastic output set $P(x,\omega)$ is compact (closed and bounded) and nonempty the well-known stochastic production function for the single output technology is defined as

$$\mathrm{f}(x,\omega) = \max\{y \in R_+ : y \in P(x,\omega)\}. \tag{2}$$

If the outputs are disposable the output set is given by an interval on the real line, i.e., $P(x,\omega) = \{y \in R_+ : 0 \leq y \leq \mathrm{f}(x,\omega)\}$.

The key idea in the generalization of the production function to multiple output technologies is to define an output set conditioned on both the inputs x and the direction of the ray through the observed outputs y. This conditioning will allow a representation of the multi output technology similar to the production function (2). A natural approach to do this is to write the output vector in polar coordinate form as

$$y = \iota \cdot m(\theta), \tag{3}$$

where $\iota = \|y\| = \left(\sum_{i=1}^{p}(y_i^2)\right)^{1/2}$ denotes the Euclidean norm of the outputs y and $m_i(\theta) = \cos\theta_i \prod_{j=0}^{i-1} \sin\theta_j$, $i = 1,...,p$, $\theta \in \left[0, \frac{\pi}{2}\right]^{p-1}$, $\sin\theta_0 = \cos\theta_p = 1$, is a transformation function of the angles θ in $(p-1)$-dimensional space (Mardia, Kent and Bibby (1979)). The function $m(\cdot)$ represents the transformation of the angle vector θ to the output mix $m(\theta) = y/\iota$ with norm $\|m(\theta)\| = 1$.

Using the polar coordinate representation in (3) we define a stochastic ray production function as

$$\mathbf{f}(x,\theta,\omega) = \max\{\iota \in R_+ : \iota \cdot m(\theta) \in P(x,\omega), \omega \in \Omega\}. \tag{4}$$

This function is analogous to the single output production function in (2) and gives the maximum norm of the attainable outputs with direction θ given inputs x. If the outputs are, at least, weakly disposable a stochastic ray output set $P(x,\theta,\omega)$ can be defined as

$$P(x,\theta,\omega) = \{y \in R_+^p : y = \iota \cdot m(\theta), 0 \leq \iota \leq \mathbf{f}(x,\theta,\omega)\}. \tag{5}$$

This set gives the set of feasible outputs along the ray with direction θ given inputs x and the state of the world ω.

2.1 The distance function

A scalar valued representation of the technology is given by the stochastic output distance function defined as

$$D_o(x,y,\omega) = \inf\left\{\delta \in R_{++} : \frac{y}{\delta} \in P(x,\omega), \omega \in \Omega\right\}. \tag{6}$$

This function provides a complete representation of the technology, see e.g., Shephard (1970) and Färe (1988). It measures the radial distance from the outputs y to the isoquant of $P(x,\omega)$ defined as $IsoqP(x,\omega) =$

$\{y : y \in P(x,\omega), \lambda y \notin P(x,\omega), \lambda > 1\}$. Using the polar coordinate representation of the outputs the distance function can be expressed as

$$D_o(x,y,\omega) = \frac{\|y\|}{\|y^f\|} = \frac{\|\iota \cdot m(\theta)\|}{\|\mathbf{f}(x,\theta,\omega)\, m(\theta)\|} = \frac{\iota}{\mathbf{f}(x,\theta,\omega)}, \tag{7}$$

where $y^f = \mathbf{f}(x,\theta,\omega)\, m(\theta) \in IsoqP(x,\omega)$ is the stochastic frontier output on the isoquant along the ray through y.

3 Frontier Models

Let the random variable $V : \Omega \to R$, defined by $v = V(\omega)$, where V is $\mathcal{F}$-measurable, represent the state of nature ω effect on the ray frontier function $\mathbf{f}(x,\theta,\omega)$. Following the standard approach in stochastic frontier models a multiplicative separability is imposed on the ray production function $\mathbf{f}(x,\theta,\omega)$ which is specified as separable in a function of the inputs and ray direction and (a function of) the random variable v as

$$\mathbf{f}(x,\theta,\omega) = f(x,\theta)\, g(v). \tag{8}$$

Furthermore, let the random variable $U : \Omega \to R_+$, defined by $u = U(\omega)$, where U is $\mathcal{F}$-measurable, represent the state of nature ω effect on the distance from the observed ray norm of the outputs y to the ray frontier function. Imposing multiplicative separability, the observed ray norm ι can be written as a separable function of the frontier ray function and (a function of) the random variable u as

$$\iota = f(x,\theta)\, g(v)\, h(-u). \tag{9}$$

Under the specification in (8) the frontier outputs $y^f \in IsoqP(x,\omega)$ is given by

$$y^f = f(x,\theta)\, g(v)\, m(\theta), \tag{10}$$

whereas the observed outputs from the specification in (9) can be written as

$$y = y^f h(-u) = f(x,\theta)\, g(v)\, m(\theta)\, h(-u). \tag{11}$$

Note that this ray frontier specification implies that both the state of nature effect, $g(v)$, and the inefficiency effect, $h(-u)$, on the outputs are radial. I.e., all output dimensions are affected by the same radial factors, $g(v)$ and $h(-u)$, respectively.

The separable multiplicative specification for the ray frontier function (8) and the observed ray (9) implies that the stochastic distance function can be written solely as a function of the random variable u as

$$D_o(x,y,\omega) = \frac{f(x,\theta)\, g(v)\, h(-u)}{f(x,\theta)\, g(v)} = h(-u). \tag{12}$$

We follow the standard approach and specify the functions $g(\cdot)$ and $h(\cdot)$ as exponential, i.e., $g(v) = \exp(v)$ and $h(-u) = \exp(-u)$. Using this and taking logarithms of (9), gives an error component generalized frontier model given by

$$\ln \iota = \ln f(x, \theta) + v - u. \tag{13}$$

$u \geq 0$ represents the efficiency in terms of the distance from $\ln l$ to the stochastic ray frontier given by $(\ln f(x, \theta) + v)$. The distance function is given by $D_o(x, y, \omega) = \exp(-u)$.

The generalized frontier model (13) resembles the standard stochastic error component frontier production model introduced by Aigner et al. (1977). The dependent variable is the norm of the output vector instead of the single output, as in the Aigner et al. model. Furthermore, the independent variables are given by both the inputs x and the ray direction θ. For a single output technology the generalized model (9) simplifies to the Aigner et al. model.

3.1 Panel data models

Let i, $i = 1, ..., N$, index a sample of firms and let t, $t = 1, ..., T$, index time periods. Using this firm and time index in (13) gives the error component panel data model

$$\ln \iota_{it} = \ln f(x_{it}, \theta_{it}) + v_{it} - u_{it}, \tag{14}$$

where $\iota_{it} = \|y_{it}\|$. $u_{it} \geq 0$ represents the efficiency of the i:th firm in time period t in terms of the distance from $\ln \iota_{it}$ to the stochastic ray frontier given by $(\ln f(x_{it}, \theta_{it}) + v_{it})$. The stochastic distance function is given by $D_{o,it} = \exp(-u_{it})$. v_{it} is assumed to be $iid\ N(0, \sigma_v^2)$ random noise. For a single output technology with time independent firm specific efficiencies the generalized model (14) simplifies to the standard frontier firm effects panel data model introduced by Schmidt and Sickles (1984)

$$\ln y_{it} = \ln f(x_{it}) + v_{it} - u_i. \tag{15}$$

3.2 Estimation

Generalized frontier models for multiple output technologies specified like (14) can be estimated using techniques developed for single output frontier models like (15). Estimation based on cross section or panel data (fixed effect or random effect models) are feasible. Greene (1993) reviews several estimation approaches based on parametric specification of the ray frontier function $f(x_{it}, \theta_{it})$ and various structures on the firm- and time specific distance functions (u_{it}). Kneip and Simar (1996) present general nonparametric kernel estimators of stochastic frontier models for single output production. The nonparametric estimators offer flexibility and can also be applied to generalized multiple output models like (14).

4 Summary

This paper presents a generalization of the single output stochastic frontier model. Based on a polar coordinate representation of the outputs a stochastic ray frontier production function can be defined where the Euclidean norm of the output vector is specified as a function of the inputs and the direction of the output vector.

In the standard approach to frontier and efficiency estimation for general multiple output technologies firm behavioral assumptions of cost minimization, revenue- or profit maximization are imposed and frontier cost, revenue or profit functions are estimated. The generalized ray frontier function offers an alternative to this approach in that stochastic production frontiers and firm-specific distance functions (technical efficiency) can be estimated for general multiple output technologies using data on inputs and outputs. No behavioral assumptions or cost or price data are needed. Estimation can easily be performed using techniques developed for the single output frontier models.

5 References

Aigner, D., Lovell, C. A. K. and P. Schmidt, 1977, Formulation and Estimation of Stochastic Frontier Production Function Models, Journal of Econometrics 6, 21-37.

Färe, R., 1988, Fundamentals of Production Theory, Lecture Notes in Economics and Mathematical Systems 311 (Springer-Verlag, Berlin Heidelberg).

Greene, W. H., 1993, The Econometric Approach to Efficiency Analysis, in: H. O. Fried, C. A. K. Lovell and S. S. Schmidt, eds., The Measurement of Productive Efficiency - Techniques and Applications (Oxford University Press, New York).

Kneip, A. and L. Simar, 1996, A General Framework for Frontier Estimation with Panel Data, Journal of Productivity Analysis 7, 187-212.

Krug, E., 1976, Stochastic Production Correspondences, Mathematical Systems in Economics 24 (Verlag Anton Hain, Meisenhaim am Glan).

Mardia, K. V., Kent, J. T. and J. M. Bibby, 1979, Multivariate Analysis (Academic Press, London).

Schmidt, P. and R. C. Sickles, 1984, Production Frontiers and Panel Data, Journal of Business and Economic Statistics 2, 367-374.

Shephard, R., 1970, Theory of Cost and Production Functions (Princeton University Press, New Jersey).

V

A Multiple Output Stochastic Ray Frontier Production Model

Mickael Löthgren*

Stockholm School of Economics

Working Paper Series in Economics and Finance No. 158
February 1997

Abstract

This paper proposes an approach to specify and estimate multiple input, multiple output production frontiers and technical efficiency using a stochastic ray frontier production model. A possible model extension is to incorporate a technical efficiency effects model to allow estimation of the effects of various explanatory variables on technical efficiency. An empirical application using Swedish health care data reveals a significant positive effect on technical efficiency of an "internal market" reform while the effect on the production frontier is negative. Technical change is found to be positive while technical efficiency has decreased over time.
Key Words: Composed error model, Distance function, Panel data, Stochastic ray frontier model, Technical change, Technical efficiency.
JEL-Classification: C12, C13, C21, C23, D24.

---

*Department of Economic Statistics, P.O. Box 6501, S-113 83 Stockholm, Sweden. E-mail: stml@hhs.se. Fax: +46 - 8 - 34 81 61. I thank Tor Jacobson, Sune Karlsson and Anders Westlund for helpful comments and Magnus Tambour for helpful comments and for providing the data.

1. Introduction

Farrell (1957) and Aigner and Chu (1968) pioneered the work on estimation of frontier production functions and technical efficiency. The frontier production function is defined by the maximum output obtainable from a given input bundle. Technical efficiency is defined by the amount observed output deviates from the production frontier and is accounted for by a truncated error term added to the production frontier function. The stochastic production frontier model introduced by Aigner, Lovell, and Schmidt (1977) and Meeusen and van den Broeck (1977) extended the original deterministic model to account for stochastic effects on the production frontier. The basic models have been extended in several directions as presented in surveys by, e.g., Schmidt (1986) and Greene (1993).

The stochastic frontier model is a scalar valued representation of the technology where a single output (input) is specified by a function of multiple inputs (outputs). The standard approach to estimate frontier functions for multiple input multiple output technologies is to estimate dual frontier cost- or profit functions where the scalar valued cost or profit is used as dependent variable. This requires that input- and output prices are available and that the behavioral assumptions of cost minimization or profit maximization are valid for the production under study. These requirements restrict the frontier model applicability since there are situations where input- and output prices are unavailable or the behavioral assumptions are invalid.

This paper addresses the issue of specification and estimation of production frontiers and technical efficiency for multiple input multiple output production technologies. A new stochastic ray frontier production model is presented that is based on the definition of a ray production function. The ray production function is a scalar valued representation of the multiple input multiple output technology specifying the Euclidean norm of the frontier output vector as a function of inputs and output mix, represented by the output polar coordinate angles. For a single output technology, the output norm equals the output level and the ray model simplifies to the standard single output model. A stochastic ray frontier model is obtained by introducing a composed error term in the same way as is done in the single output model. The structure imposed by the model is that the error terms are restricted to affect the output norm multiplicatively implying that observed outputs are given by radial rescaling of frontier outputs. The derived technical efficiency measure represents the radial distance from the output to the production frontier corresponding to the definition of technical efficiency in Farrell (1957) and

the output distance function by Shephard (1970).

The ray frontier model offers an alternative to the approach presented in Kumbhakar (1996) where the technology is represented using separate scalar valued input and output functions. Kumbhakar (1996, p. 228) notes that this approach is, however, not very promising due to the limited availability of multiple output production functions. Furthermore, the imposed separability does not allow modelling of interaction effects between inputs and outputs in the production process. The proposed ray frontier model allows, on the other hand, for possible interactions of inputs and output mix on the frontier output norm. A possible extension of the ray frontier model is to incorporate and extend the technical efficiency effects model in Battese and Coelli (1995) to multiple input multiple output technologies. In this model technical efficiency is specified as a separate function of a set of appropriate explanatory variables. This model extension allows simultaneous identification and estimation of the ray frontier function, technical efficiency and the technical efficiency effects function.

The paper includes an empirical application using panel data from the Swedish public health care sector. The data consist of a panel of 26 Swedish local government county councils over the period $1989-1994$. During this period five county councils initiated an "internal market" organizational reform with the primary objective to increase technical efficiency. The data have been studied previously by Tambour and Rehnberg (1997) where the Swedish health care sector and the "internal market" reform is presented in detail. Tambour and Rehnberg (1997) use the DEA method to analyze the reform effect on technical efficiency. In this paper a linear technical efficiency effects stochastic ray frontier model is used to estimate the reform effect on both the production frontier and technical efficiency. The model is estimated by maximum likelihood estimation using the program FRONTIER written by T. Coelli. The results indicate that the reform has had a positive effect on technical efficiency while the effect on the production frontier has been negative. The technical change in health care production is clearly positive while technical efficiency has tended to decrease over time.

The paper unfolds as follows: Section 2 presents the ray production function and the stochastic ray frontier composed error model. The panel data model and the distributional assumptions underlying the maximum likelihood estimation are presented in Section 3. Section 4 provides an empirical application and Section 5 ends the paper with a summary and some concluding remarks.

2. The Ray Frontier Production Function

Consider a production technology where multiple inputs $x \in R_+^d$ are used to produce multiple outputs $y \in R_+^p$. The technology can be described by the output set $P(x)$ defined as

$$P(x) = \{y \in R_+^p : x \text{ can produce } y\}. \tag{2.1}$$

The technology is assumed to satisfy a set of basic axioms discussed in Shephard (1970) and Färe (1988). Convexity of $P(x)$ for all x and disposability (monotonicity) of inputs and outputs are the most important axioms. Provided the output set $P(x)$ is compact (closed and bounded) and nonempty the standard production function for the single output technology is defined as

$$\mathrm{f}(x) = \max\{y \in R_+ : y \in P(x)\}. \tag{2.2}$$

A key step to obtain a scalar valued representation of the multiple output technology is to use a polar coordinate representation of the output vector which can be written as

$$y = \iota \cdot m(\theta), \tag{2.3}$$

where $\iota = \|y\| = \left(\sum_{i=1}^p (y_i^2)\right)^{1/2}$ denotes the Euclidean norm of the output vector y. The function $m : \left[0, \frac{\pi}{2}\right]^{p-1} \to [0,1]^p$, defined by $m_i(\theta) = \cos\theta_i \prod_{j=0}^{i-1} \sin\theta_j$, $i = 1, ..., p$, $\theta \in \left[0, \frac{\pi}{2}\right]^{p-1}, \sin\theta_0 = \cos\theta_p = 1$ (Mardia, Kent, and Bibby (1979)) represents a transformation of the polar coordinate angle vector $\theta \in \left[0, \frac{\pi}{2}\right]^{p-1}$ to the output mix vector $m(\theta) = y/\iota$. The polar coordinate angles θ are easily obtained from the observed outputs by inverting the transformation function $m(\theta)$ recursively for each dimension $i = 1, ..., p$, respectively. The solution for θ can written as

$$\theta_i = \cos^{-1}\left(y_i/\iota \prod_{j=0}^{i-1} \sin\theta_j\right), i = 1, ..., p, \tag{2.4}$$

where $\sin\theta_0 = \cos\theta_p = 1$. The first angle is given by $\theta_1 = \cos^{-1}(y_1/\iota)$. This is used in the calculation of the second angle which is given by $\theta_2 = \cos^{-1}(y_2/\iota\sin\theta_1)$. The remaining angles $\theta_i, i = 3, ..., p-1$, are obtained by continuing the recursion.

The polar coordinate representation can be used to define a ray production function analogous to the single output production function (2.2) as

$$\mathbf{f}(x, \theta) = \max\{\iota \in R_+ : \iota \cdot m(\theta) \in P(x)\}. \tag{2.5}$$

This function gives the maximum norm of the attainable outputs given inputs x and output mix represented by the polar coordinate angles θ.

The ray function inherits some properties from the output set. For instance, free input disposability of the output set, stating that $P(x') \subseteq P(x''), \forall\, x'' \geq x'$, translates to a positive input monotonicity property of the ray function $\mathbf{f}(x, \theta)$, i.e., $\mathbf{f}(x'', \theta) \geq \mathbf{f}(x', \theta), \forall\, x'' \geq x'$. The curvature of the production frontier can be derived from the partial derivatives of the ray function with respect to the polar coordinate angles, $\partial \mathbf{f}(x, \theta) / \partial \theta_i, i = 1, ..., p-1$. These derivatives reveal the change in the output norm when the output mix is changed along the production frontier, given the level of the inputs. For a technology with three outputs, the first angle θ_1 represents the angle from the y_1 axis toward the plane spanned by the y_2 and y_3 axis. The angle θ_2 represents the angle between y_2 and y_3 in the $y_2 - y_3$ plane. Theses two angles and the ray norm ι identifies the output vector y by the representation in (2.3). The partial derivative $\partial \mathbf{f}(x, \theta) / \partial \theta_1$ thus represents the change of the frontier output norm for changes in the output mix along the output frontier with fixed proportions between y_2 and y_3. The derivative $\partial \mathbf{f}(x, \theta) / \partial \theta_2$ represents, on the other hand, the frontier output norm change due to changes in the output mix, with the level of the first output dimension y_1 held constant.

A multiple output scale elasticity measure can be defined in terms of partial derivatives of the ray function as

$$\varepsilon = \frac{\nabla_x \mathbf{f}(x, \theta) \cdot x}{\mathbf{f}(x, \theta)}, \tag{2.6}$$

where $\nabla_x \mathbf{f}(x, \theta)$ denotes the gradient of the ray function. This scale measure gives the ratio of the proportional change in the frontier output norm due to a proportional change in the inputs, conditioned on the output direction as represented by the polar coordinate output angles θ and the input level. Since a scale change in the frontier output norm corresponds to a proportional change in all frontier output dimensions, the scale elasticity measure in (2.6) can be interpreted as a local measure of the frontier elasticity of scale. For a single output technology (2.6) simplifies to the well known single output scale elasticity measure $\varepsilon = \nabla_x \mathrm{f}(x) \cdot x / \mathrm{f}(x)$.

2.1. A stochastic ray frontier production function

By introducing a composed error term similarly as in the original single output model by Aigner, Lovell, and Schmidt (1977) and Meeusen and van den Broeck (1977) a stochastic ray frontier production function model can be specified as

$$\iota = f(x, \theta) \exp(v - u). \tag{2.7}$$

The observed ray norm ι is specified as a function of a deterministic ray function $f(x, \theta)$ and a composed error term $\varepsilon = v - u$ where the symmetric random variable v accounts for random events affecting the production frontier and the truncated random variable u, $u \geq 0$, accounts for random events affecting the technical efficiency of the firm.

The stochastic frontier norm is given by $\iota^f = f(x, \theta) \exp(v)$. The technical efficiency is modeled as affecting the output radially. The observed output is obtained from a radial rescaling of the frontier output level $y^f = \iota^f m(\theta)$ and equals $y = y^f \exp(-u)$, where the output reduction factor defines the technical efficiency measure $TE = \exp(-u)$. This measure represents the radial distance from the output vector to the frontier of the output set and corresponds to the (output) technical efficiency in Farrell (1957). It is also equivalent to the output distance function defined in Shephard (1970) as

$$D_o(y, x) = \min \left\{ \mu : \frac{y}{\mu} \in P(x) \right\}. \tag{2.8}$$

Under the specification above, the distance function is given by the ratio of the frontier norm to the observed norm, i.e. $D_o(y, x) = \iota^f / \iota = \exp(-u)$. Technical efficiency is thus completely represented by the random variable u.

3. The Model

A linear panel data stochastic ray frontier production model is obtained by imposing a linear functional form of the ray function and taking logarithms of (2.7) as

$$\ln \iota_{it} = \beta_0 + z_{it}' \beta + v_{it} - u_{it}, i = 1, ..., N, t = 1, ..., T, \tag{3.1}$$

where i indexes firms and t indexes time periods. z_{it} is a K-vector of (transformations of) the inputs and the output angles and other firm and time specific variables including time. $u_{it} \geq 0$ represents the efficiency of firm i at time period t as the distance from the log norm $\ln \iota_{it}$ to the stochastic logarithmic ray frontier given by $(\ln f(x_{it}, \theta_{it}) + v_{it})$. If $T = 1$ (3.1) simplifies to a linear cross section ray frontier model.

The model (3.1) can be rewritten in standard panel data form as

$$\ln \iota_{it} = \alpha_{it} + z'_{it}\beta + v_{it}, i = 1, ..., N, t = 1, ..., T, \tag{3.2}$$

where $\alpha_{it} = \beta_0 - u_{it}$ denotes the firm- and time specific intercept. For a single output technology, where $\iota_{it} = y_{it}$, with time invariant technical efficiencies (3.2) simplifies to the firm specific intercept production frontier panel data model introduced by Schmidt and Sickles (1984).

The parameters β_0 and β in (3.1) and (3.2) can be estimated by OLS or maximum likelihood estimation. Similarly as for the single output stochastic frontier models the OLS estimator of the intercept β_0 will be inconsistent, due to the truncated error term u in the composed error $(v - u)$. However, the intercept could be consistently estimated using a correction of the OLS estimator, known as COLS obtained from the original OLS intercept estimator and higher order (i.e., second and third) moments of the OLS residuals. The slope parameters can be estimated consistently using OLS. See e.g., Greene (1993) for a discussion of the consistency of the OLS and the COLS estimation methods.

In this paper estimation is performed by MLE and the following assumptions are used:

1. The random noise terms v_{it} are assumed IID normally distributed as $v_{it} \sim N(0, \sigma_v^2)$.

2. The inputs x_{it} and polar coordinate angles θ_{it} are independent of $v_{i't'}$ for $i, i' = 1, ...N$, $t, t' = 1, ...T$.

3. Following Battese and Coelli (1995), the efficiency terms u_{it} are modelled as ID (independently distributed) non-negative random firm- and time specific efficiencies with a truncated (at zero) normal distribution with time varying mean, i.e., $u_{it} \sim |N(m_{it}, \sigma_u^2)|$. The firm- and time varying mean m_{it} is specified as a linear function $m_{it} = z_{it}^{*\prime}\delta$, where z_{it}^* is an m-vector of exogenous variables associated with the technical efficiencies.

   Alternatively interpreted, the efficiency terms are given by

$$u_{it} = m_{it} + \omega_{it}, \tag{3.3}$$

   where ω_{it} are unobservable ID random variables $\omega_{it} \sim N(0, \sigma_w^2)$, $\omega_{it} \geq -m_{it}$. The truncation $\omega_{it} \geq -m_{it}$ guarantees that the efficiency terms are non-negative. The efficiency terms u_{it} (or equivalently ω_{it}) are furthermore assumed independent of $v_{i't'}$, for $i, i' = 1, ...N$, $t, t' = 1, ...T$.

An advantage of the technical efficiency effects model is that a time effect can be incorporated in both the ray frontier function and in a specification of time varying technical efficiency. This allows estimation of both (exogenous) technical change of the production frontier and the linear time varying structure of technical efficiency. Battese and Coelli (1995) and the references cited therein provide a discussion of estimation of the model.

In this paper we concentrate on estimation of the frontier function and the technical efficiency effects model. Firm specific technical efficiencies (distance functions) can be estimated using the approach by Jondrow, Lovell, Materov, and Schmidt (1982) generalized by Battese and Coelli (1988) to panel data models. The technical efficiencies are predicted by the conditional expectation of the truncated error term u, conditioned on composed error realizations as

$$\widehat{TE}_{it} = E\left(\exp\left(-u_{it}\right)|\widehat{\varepsilon}_{it}\right) = E\left(\exp\left(-m_{it} - \omega_{it}\right)|\widehat{\varepsilon}_{it}\right). \tag{3.4}$$

4. Empirical Application

4.1. Data

The data consist of a panel data set of 26 Swedish County Councils providing health care over the period 1989 – 1994. Data are highly aggregated relating to health care production of short term surgical care and internal medicine. As proxies for inputs in the health care technology we use the cost of production ($COST$) in MSEK (1990 years prices) and the number of available beds ($BEDS$). Cost is used since data on labor use are not easily accessible for short term care. The cost measure does not reflect expenses for capital services. Therefore, the number of beds is used as a crude proxy for capital input in the health care production. The proxies for outputs are the number of operations ($OPER$), the number of admissions ($ADMI$) and the number of visits to physicians ($VISIT$). Tambour and Rehnberg (1997) provide a full discussion of the data definitions and data sources. During the observed time period, five of the 26 county councils initiated the internal market reform. One in 1990, two in 1991 and 1992, respectively.

Table 1 gives some aggregate descriptive statistics of the variables where the polar coordinate representation of the outputs are presented is addition to the original input and output data. The output norm ι is defined as the Euclidean norm of the vector $(OPER, ADMI, VISIT)$. The polar coordinates θ_1 and θ_2 are obtained from the calculation described in (2.4).

Table 1: Descriptive statistics of inputs and outputs. $N = 26$, $T = 6$.

	Mean	Std. Dev.	Min.	Max.
Input:				
$COST$	1797	1593	286	9868
$BEDS$	1167	865	186	5893
Output:				
$OPER$	26880	23039	3324	146785
$ADMI$	54533	40415	9745	249751
$VISIT$	348	315	62	1890
ι	60917	46367	10353	287570
θ_1	1.128	0.058	0.905	1.267
θ_2	0.0062	0.0012	0.0033	0.0091

4.2. Results

The following loglinear stochastic ray frontier production function is estimated

$$\ln \iota_{it} = \beta_0 + \beta_1 \ln COST_{it} + \beta_2 \ln BEDS_{it} + \beta_3 \ln \theta_{1,it} + \beta_4 \ln \theta_{2,it} + \beta_5 D_{IM,it} + \beta_6 t + u_{it} + v_{it}, i = 1, ..., N, t = 1, ...T, \tag{4.1}$$

where $D_{IM,it}$ is a dummy variable representing the "internal markets" reform. $D_{IM,it} = 0$ before the reform has been introduced and $D_{IM,it} = 1$ thereafter. The time trend parameter β_6 represents the exogenous neutral technical change in the ray frontier. The technical efficiency effects model is specified as

$$u_{it} = \delta_0 + \delta_1 D_{IM,it} + \delta_2 t + w_{it}. \tag{4.2}$$

where $w_{it} \sim N\left(0, \sigma_w^2\right), w_{it} \geq -\left(\delta_0 + \delta_1 D_{IM,it} + \delta_2 t\right)$. The inclusion of the market dummy in both the frontier function and in the technical efficiency model allows identification of the organizational effect on the production frontier, represented by the shift parameter β_5 and the effect on technical efficiency, represented by the parameter δ_1. The parameter δ_2 represents the linear effect of time on the technical inefficiency.

Estimation is performed using the MLE Fortran program FRONTIER, version 4.1, by Tim Coelli, Centre for Efficiency and Productivity Analysis University of

New England, Armidale. Coelli (1992, 1996) present the program that is publicly available for downloading at the website:
http://www.une.edu.au/econometrics/cepa.htm.

The estimation results are given below (standard errors in parentheses):

Stochastic Ray Frontier:

$$\ln \iota = \underset{(0.173)}{3.06} + \underset{(0.044)}{0.042} \ln COST + \underset{(0.047)}{0.970} \ln BEDS$$

$$- \underset{(0.130)}{0.089} \ln \theta_1 - \underset{(0.027)}{0.141} \ln \theta_2 - \underset{(0.032)}{0.183} D_{IM} + \underset{(0.0092)}{0.106} t$$

Technical Inefficiency:

$$u_{it} = \underset{(0.032)}{0.115} - \underset{(0.048)}{0.359} D_{IM} + \underset{(0.0096)}{0.054} t$$

Variance Parameters: $\widehat{\sigma}^2 = 0.00536$, $\widehat{\gamma} = 0.999$

Loglikelihood: 201.66

The positive signs of the estimated input (elasticity) coefficients in the stochastic ray frontier model are as expected in accordance with the input monotonicity property. The cost input is however not significant. Thus, increased costs does not increase in the output norm significantly. The returns to scale estimate in this linear model is given by 1.01, indicating existence of constant returns to scale in health care production. The estimates of the polar angles are both negative with a significant estimate for the second angle. This indicates that changes in the frontier output mix when visits are substituted for admissions, with the levels of operations held fixed, lead to an increase in number of visits less than the decrease in admissions. The estimated coefficients on the internal market dummy and the time trend are clearly significant. The negative sign on the market dummy indicates that the organizational reform has shifted the production frontier inwards, decreasing the maximum output producible given inputs and output mix. This result could be explained by increasing (administrative) costs due to the internal market reform not accompanied by increased output levels. The trend parameter estimate is positive, indicating a positive technical change in the health care production frontier during the observed time period. A possible explanation of this is that new medical technology and health care treatment methods have had

a significant positive effect on the production frontier. For the technical inefficiency, the negative estimate of the market dummy indicates that the reform has had a significant positive effect on the technical efficiency. This result confirms the finding in Tambour and Rehnberg (1997), indicating that one of the primary objectives for implementing the reform is fulfilled. The positive estimate of the time trend indicates, on the other hand, that the technical inefficiency of the observed county councils tended to increase over time throughout the observed period. This result could be explained by a heterogeneity among the observed county councils in implementing new technology and new treatment methods. With a positive technical change, as it is estimated here, the units exploiting the new technology are pushing the production frontier outwards. This implies that the distance to the frontier for the units that are more reluctant in implementing the new technology is increasing. In the model applied here, this increased distance is identified as increased technical inefficiency.

In summary, the results indicate that the organizational reform has had a somewhat mixed effect on the health care production. The reform has been successful in the primary motivation of increasing technical efficiency. At the same time the reform has had a cost in the sense of a negative production frontier shift. One interpretation of this is that the estimated frontier shift provides a measure of the adjustment cost involved in implementing the reform. This result should however be carefully interpreted due to the limited length of the observed time period. As more time periods become available, the persistence of this adjustment cost can be assessed. It is possible that the long term effect of the reform is positive, both regarding the technical efficiency effect and the production frontier effect. Nevertheless, the results indicate a clear adjustment cost - persistent or not - of the internal market reform.

Likelihood ratio tests of various parameter restrictions in the ray frontier function and the efficiency effects function are presented in *Table 2*. The likelihood ratio test statistic is given by $\lambda = 2\left(\ln L_1 - \ln L_0\right)$, where $\ln L_0$ and $\ln L_1$ denotes the loglikelihood value under the null hypothesis H_0 and the alternative H_1, respectively.

The variance parameter $\gamma = \sigma_u^2/\sigma^2$, where $\sigma^2 = \sigma_u^2 + \sigma_v^2$ is the variance of the composed error, represents overall technical efficiency in the model. The hypothesis $H_0 : \gamma = 0$ states that the technical efficiency variance is zero, or equivalently that all firm are technically efficient. Since the efficiency variance is non-negative this hypothesis is tested against the one-sided alternative $H_1 : \gamma > 0$. Coelli (1995) notes that the asymptotic distribution of the LR test statistic

for hypotheses including a zero variance parameter follows a mixed chi-square distribution under the null hypothesis. This is due to the fact that the variance parameter lies on the boundary of the parameter space under the null hypothesis. Critical values for these hypotheses are obtained from Table 1 in Kodde and Palm (1986), using $q+1$ degrees of freedom, where q is the number of parameters specified to be zero under the null which are not boundary values.

Table 2: Likelihood ratio test results (5% significance level)

Null Hypothesis	λ	Critical Value[1]	Decision
$H_0 : \gamma = 0$	16.68	7.05	Reject H_0
$H_0 : \gamma = \delta_0 = \delta_1 = \delta_2 = 0$	16.68	8.76	Reject H_0
$H_0 : \delta_1 = \delta_2 = 0$	14.64	5.89	Reject H_0
$H_0 : \beta_5 = \beta_6 = 0$	93.20	5.89	Reject H_0
$H_0 : \beta_3 = \beta_4 = 0$	31.82	5.89	Reject H_0
$H_0 : \beta_1 + \beta_2 = 1$	18.38	3.84	Reject H_0

The estimate of the variance parameter γ is close to one, indicating that technical efficiency variation is a main component of the estimated model. This is confirmed by the rejection of both $H_0 : \gamma = 0$ and $H_0 : \gamma = \delta_0 = \delta_1 = \delta_2 = 0$. The hypothesis $H_0 : \gamma = \delta_0 = \delta_1 = \delta_2 = 0$ represents a test of absence of stochastic technical inefficiency effects in the model as specified by (4.2). Under the hypothesis $H_0 : \gamma = 0$ the model reduces to a traditional regression model where the variables entering the technical effects model (4.2) instead are included in the stochastic frontier function. Since both the internal market dummy and the trend variable are already included in the frontier function (4.1) the parameters δ_0, δ_1 and δ_1 cannot be identified and separated from the frontier function parameters β_0, β_5 and β_6. Under the restricted model, the estimated reform effect on the production frontier is significant and positive. Since the hypothesis $H_0 : \gamma = 0$ is clearly rejected, this indicates the importance of including the reform variable in both the frontier function and in the technical effects model in order to separate the different effects of the reform on the production.

The hypothesis $H_0 : \delta_1 = \delta_2 = 0$, testing existence of reform- and time dependence of the technical efficiency distribution, can be rejected. Under this hypothesis the inefficiency terms follow a truncated half normal distribution $u_{it} \sim$

[1] For the hypotheses $H_0 : \gamma = 0$ and $H_0 : \gamma = \delta_0 = \delta_1 = \delta_2 = 0$ the critical values are obtained from Table 1 in Kodde and Palm (1986) using 3 and 4 degrees of freedom, respectively.

$|N(\mu_u, \sigma_u^2)|$, where the mean $\mu_u = \delta_0$. The hypothesis $H_0 : \beta_5 = \beta_6 = 0$ of the inclusion of the market dummy and the trend variable in the frontier model can be rejected, indicating that the internal market and the time effects jointly matter in the frontier function specification.

The curvature hypothesis $H_0 : \beta_3 = \beta_4 = 0$ is a test of curvature shape of the production frontier. This hypothesis can be rejected implying that we can reject that the output set is "ball shaped" with constant frontier norm for all output directions. Finally, the constant returns to scale hypothesis $H_0 : \beta_1 + \beta_2 = 1$ is rejected. This result seems somewhat strange since the point estimate of the returns to scale factor $\beta_1 + \beta_2$ is as close to one as 1.01. A possible explanation is that the finite sample distribution of the likelihood ratio test statistic differs from the asymptotic $\chi^2(1)$ distribution. A confirmation of this requires however simulation studies not yet performed.

5. Summary and Concluding Remarks

This paper presents the stochastic ray frontier production function model as a generalization of the single output frontier model allowing estimation of frontier functions and technical efficiency for multiple input multiple output technologies. The stochastic ray frontier function is based on a polar coordinate representation of the outputs and specifies the output norm as a function of inputs and polar coordinate output angles. The error terms affect the outputs radially and technical efficiency is defined by the radial distance from the observed output norm to the production frontier. The ray model can incorporate the technical efficiency effects model by Battese and Coelli (1995) allowing identification and estimation of various specifications of both the frontier and the technical efficiency effects models. A trend variable can be included to account for technical change in the production frontier as well as the temporal pattern of technical efficiency. Estimation of the parameters in the model can be performed by maximum likelihood estimation based on distributional assumptions on the composed error terms.

The model is applied to a panel of 26 Swedish county councils providing health care over the period $1989-1994$. During this period five county councils initiated an organizational reform with a main objective to increase technical efficiency. The estimation results indicate that the reform has had a significant positive effect on the technical efficiency. However, the reform has had a negative effect in terms of a negative production frontier shift that can be interpreted as the cost of introducing the reform. Furthermore, the technical change is positive, indicating

increased production possibilities in the health care production over the studied time period. Technical inefficiency has on the other hand increased over time.

As a concluding remark, we note the need for future work on the choice of appropriate functional form for the ray frontier function. The main purpose of the included empirical application using the simple linear ray frontier model is to illustrate estimation of the new ray model. A natural extension is to perform a model selection analysis using several competing flexible functional forms as is done in the single output context in Battese and Broca (1996).

References

AIGNER, D., C. A. K. LOVELL, AND P. SCHMIDT (1977): "Formulation and Estimation of Stochastic Frontier Production Function Models," *Journal of Econometrics*, 6, 21–37.

AIGNER, D. J., AND S. F. CHU (1968): "On Estimating the Industry Production Function," *American Economic Review*, 58, 824–839.

BATTESE, G. E., AND S. S. BROCA (1996): "Functional Forms of Stochastic Frontier Production Functions and Models for Technical Inefficiency Effects: A Comparative Study for Wheat Farmers in Pakistan," Centre for Efficiency and Productivity Analysis and Department of Econometrics, University of New England, Armidale.

BATTESE, G. E., AND T. COELLI (1988): "Prediction of Firm-Level Technical Efficiencies with a Generalized Frontier Production Function and Panel Data," *Journal of Econometrics*, 38, 387–399.

——— (1995): "A Model for Technical Inefficiency Effects in a Stochastic Frontier Production Function," *Empirical Economics*, 20, 325–332.

COELLI, T. (1992): "A Computer Program for Frontier Production Function Estimation," *Economics Letters*, 39, 29–32.

——— (1995): "Estimators and Hypothesis Tests for a Stochastic Frontier Function: A Monte Carlo Study," *Journal of Productivity Analysis*, 6, 247–268.

——— (1996): "A Guide to FRONTIER Version 4.1: A Computer Program for Stochastic Frontier Production and Cost Function Estimation," Centre for

Efficiency and Productivity Analysis (CEPA) Working Paper 96/07, University of New England, Armidale.

FÄRE, R. (1988): *Fundamentals of Production Theory*, vol. 311 of *Lecture Notes in Economics and Mathematical Systems*. Springer-Verlag, Berlin Heidelberg.

FARRELL, M. S. (1957): "The Measurement of Productive Efficiency," *Journal of the Royal Statistical Society, Series A*, 120, 253–281.

GREENE, W. H. (1993): "The Econometric Approach to Efficiency Analysis," in H. O. Fried, C. A. K. Lovell and S. S. Schmidt (eds.) The Measurement of Productive Efficiency, Oxford University Press, New York.

JONDROW, J., C. A. K. LOVELL, I. S. MATEROV, AND P. SCHMIDT (1982): "On the Estimation of Technical Inefficiency in the Stochastic Frontier Production Model," *Journal of Econometrics*, 19, 233–238.

KODDE, D. A., AND F. C. PALM (1986): "Wald Criteria for Jointly Testing Equality and Inequality Restrictions," *Econometrica*, 54, 1243 – 1248.

KUMBHAKAR, S. C. (1996): "Efficiency Measurement with Multiple Outputs and Multiple Inputs," *Journal of Productivity Analysis*, 7, 225–255.

MARDIA, K. V., J. T. KENT, AND J. M. BIBBY (1979): *Multivariate Analysis*. Academic Press, London.

MEEUSEN, W., AND J. VAN DEN BROECK (1977): "Efficiency Estimation from Cobb-Douglas Production Functions with Composed Error," *International Economic Review*, 18, 435–444.

SCHMIDT, P. (1986): "Frontier Production Functions," *Econometric Reviews*, 4, 289–328.

SCHMIDT, P., AND R. C. SICKLES (1984): "Production Frontiers and Panel Data," *Journal of Business and Economic Statistics*, 2, 367–374.

SHEPHARD, R. W. (1970): *The Theory of Cost and Production Functions*. Princeton University Press.

TAMBOUR, M., AND C. REHNBERG (1997): "Internal Markets and Performance in Swedish Health Care," Mimeo, Centre for Health Economics, Stockholm School of Economics.

VI

Productivity and Customer Satisfaction -A DEA Network Model*

Mickael Löthgren† and Magnus Tambour‡
Stockholm School of Economics
P.O. Box 6501
S-113 83 Stockholm
Sweden

Working Paper Series in Economics and Finance No. 140
December 1996

Abstract

This paper presents a network model incorporating customer satisfaction into efficiency and productivity measures. The network consists of a production node and a consumption node and offers flexibility in modelling the production and consumption process. Allocation of input resources to production and customer oriented activities is allowed. In the consumption process inputs and output characteristics result in customer satisfaction. The model solution identifies optimal allocation of resources between production and customer oriented activities. Data envelopment analysis estimators of the defined theoretical measures of efficiency and productivity are presented. An empirical application using data from a sample of Swedish pharmacies is included. Results from the network model and a direct productivity model indicate increased average productivity, although the productivity progress is somewhat lower in the network model.

Key-words: Customer satisfaction, Data envelopment analysis, Distance function, Malmquist productivity index, Network model.

JEL-Classification: D11, D12, D24.

The paper was presented at the second biannual Georgia Productivity Workshop, Athens, GA, November 1 - 3, 1996 and the Nordic Workshop in Productivity, Growth and Development, Göteborg, November 28 - 29, 1996.

*We thank Anders Westlund for helpful comments and suggestions and the National Corporation of Swedish Pharmacies for providing the data, in particular Karin Arrhenius and Jörgen Dolby.

†Department of Economic Statistics. *E-mail:* stml@hhs.se. *Fax:* +46-8-348161.

‡Centre for Health Economics. *E-mail:* hemt@hhs.se. *Fax:* +46-8-302115.

1. Introduction

Customer satisfaction has become an important performance indicator for both private and public firms as discussed in Fornell (1992). Customer satisfaction barometers can often be regarded as complementary to productivity measures since the latter seldom take customer perceived quality into consideration. If customer satisfaction is an explicit objective for a firm, the inclusion of such measures into productivity indices leads to more valid measures of productivity. The purpose of this paper is to present an approach to include customer satisfaction in productivity measures. The proposed model allows estimation of efficiency and productivity taking both qualitative and quantitative (productivity) aspects into account. The estimates can be obtained using data envelopment analysis (DEA).

DEA is a non-parametric linear programming (LP) approach to estimate production characteristics such as technical efficiency and productivity (see, e.g., Charnes, Cooper and Rhodes (1978) and Banker, Charnes and Cooper (1984)). An appealing property of the DEA-approach is that multiple-input, multiple-output technologies readily can be modeled without revenue or cost data. This is important in efficiency and productivity measurement in public production, where price or cost data are often unavailable.

We propose the use of a network model where the fundamental idea is that the production and consumption processes can be represented as separate nodes. This approach extends the network model in Färe and Grosskopf (1996), where a network model for productivity measurement with intertemporal products is specified.

The network approach is flexible in the sense that production and consumption can be jointly modeled along with a broad representation of customer satisfaction. The network model allows an allocation of resources between customer oriented activities (x^C) and traditional production (x^P). A subvector of the outputs (y^C), representing characteristics and quality attributes, is treated as an intermediate input in a consumption technology. Moreover, quality assessments (q) as well as the "ordinary" outputs (y^P) are considered as final exogenous production from the consumption and the production node, respectively. The quality assessments are here represented by data from Customer Satisfaction Barometers (CSB), see e.g., Fornell (1992). It can be noted that the idea of a consumer technology is similar to the model in Lancaster (1991) where it is assumed that "consumption is an activity in which goods, singly or in combination, are inputs and in which the output is a collection of characteristics", Lancaster (1991: p. 12).

The technology is represented by distance functions which also define the efficiency and productivity measures. The distance functions can be estimated by DEA using only primal production data without imposing a functional form for the technology. The LP solutions allow identification of optimal allocation of production resources. Furthermore, the optimal level of the characteristics and attributes subvector (y^C) can also be identified.

An alternative approach to incorporate customer satisfaction in productivity indices is given in Färe, Grosskopf and Roos (1996). Their approach is based on a preference indirect distance function defined as a distance function incorporating a utility restriction. Quality assessments are used as weights in a (linear) utility function defined over attributes (and outputs). One disadvantage with this approach when using DEA is that a functional form of the utility function has to be specified and that there has to be a one-to-one relation between the assessments and the outputs to allow an interpretation of the assessments as utility function weights. In the network model the utility restriction is exchanged by a distance function defined for the consumer technology and the relation between the assessment dimensions in the CSB and the outputs (and quality attributes) does not have to be one-to-one. This implies that a broad representation of consumer satisfaction can be incorporated in the network productivity model.

The paper includes an empirical application where the network model and a direct productivity model is estimated using a sample of Swedish pharmacies. The outlined model is an appropriate model of the pharmacy technology since a large proportion of labor time is devoted to customer related activities such as information and counselling.

The paper unfolds as follows: *Section* 2 presents the network model and the distance function representation of the technology. This section also defines the Malmquist productivity index for the network and the production node. DEA estimators of firm specific distance functions and Malmquist indices are presented in *Section* 3. *Section 4* presents the empirical application and *Section* 5 contains a summary and some concluding remarks.

2. The Network Model

We use a network technology that consists of two nodes as a representation of the production and consumption process. In the production node, firms use inputs $x^P \in R_+^N$ to produce outputs $y = \left(y^P, y^C\right) \in R_+^{M+J}$. The subvector y^P represents traditional, marketable (final) outputs whereas the subvector y^C represents non-marketable characteristics and attributes of the production. The characteristics and attributes are then considered as intermediate inputs, together with inputs $x^C \in R_+^N$ in the consumption node, resulting in customer quality assessments $q \in R_+^L$. The total inputs $x = x^P + x^C$ and the vector $\left(y^P, q\right)$ are treated as exogenous. The network model is illustrated in Figure 2.1 adapted from Färe and Grosskopf (1996).

2.1. The production technology (P-node)

The production technology is represented by an input distance function (Shephard (1970)) defined as:

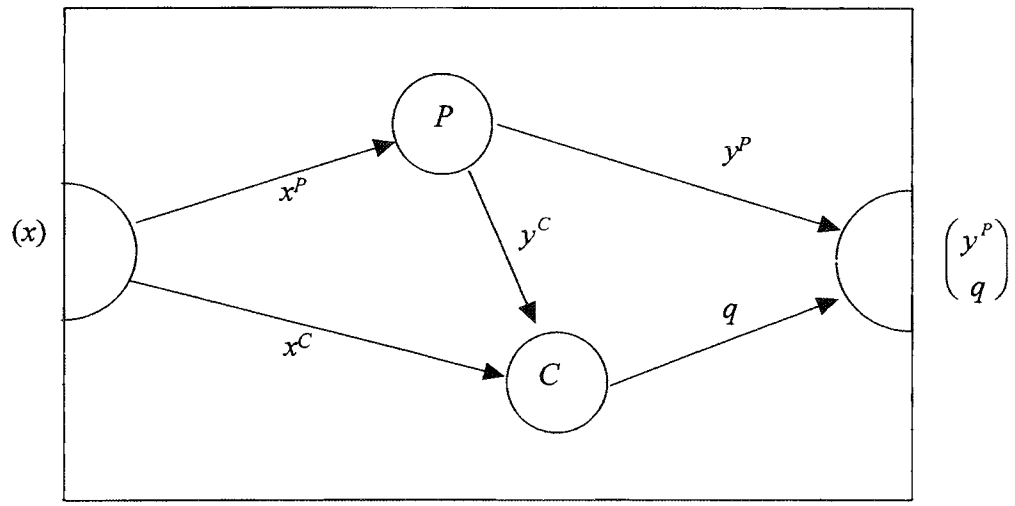

Figure 2.1: The Network Technology

$$D_i^{P,t}\left(x^P,\left(y^C,y^P\right)\right)=\max\left\{\lambda:\frac{x^P}{\lambda}\in L^t\left(y^C,y^P\right)\right\}, \tag{2.1}$$

where $L^t\left(y^C,y^P\right)=\left\{x^P:\ x^P \text{ can produce } \left(y^C,y^P\right) \text{ at time } t\right\}$ is the period t input requirement set. The production technology is assumed to satisfy a set of axioms. In short: (i) inactivity is allowed, (ii) "free lunch" is not allowed, (iii) strong disposability of inputs and outputs and (iv) the output set is a compact and convex set. See Färe, Grosskopf and Lovell (1994) for a presentation of these axioms. The input distance function takes on values larger than one if and only if $x^P\in L^t\left(y^C,y^P\right)$. Technical efficiency is achieved when the distance function equals one.

2.2. The network technology

Färe and Grosskopf (1996) define a network technology for the P-node. Here we specify a network model as joint production and consumption technology in an analogous manner. The network technology in Figure 2.1 can be represented by the network distance function

$$\mathcal{D}_i^t\left(x,\left(y^P,q\right)\right)=\max_{\widetilde{x}^P,\widetilde{x}^C,\widetilde{y}^C}\left\{\lambda:\frac{x}{\lambda}\in\mathcal{L}^t\left(\left(\widetilde{y}^C,y^P\right),q\right),x\geq\widetilde{x}^P+\widetilde{x}^C\right\}, \tag{2.2}$$

where $\mathcal{L}^t\left(\left(y^C,y^P\right),q\right)=\left\{x:\ x^P\in L^t\left(y^C,y^P\right),q\in Q^t\left(x^C,y^C\right),x\geq x^P+x^C\right\}$ is the network grand input set. $Q^t\left(x^C,y^C\right)=\left\{q:x^C \text{ and } y^C \text{ can give } q \text{ at time } t\right\}$ is the

quality assessment set in the consumption technology in time period t representing the set of quality assessments attainable from $x^C$and y^C.

The distance function measures the maximum proportional contraction of the inputs given that an output vector can be produced that still allows the quality assessments to be attained. The network distance function thus constitutes a measure of technical efficiency that incorporates consumer satisfaction represented by the quality assessments in the consumption node.

Note that the optimal values of the choice variables $\widetilde{x}^P, \widetilde{x}^C$ and $\widetilde{y}^C$ in (2.2) identify an optimal allocation of the inputs (x^P and x^C) and the output attributes/characteristics (y^C). It is of course possible that one or more elements in the input vector x only goes to one of the two nodes. This can easily be taken into consideration in the model. Since $x_n = x_n^P + x_n^C$, $n = 1, ..., N$, we simply set, for example, $x_{n'}^P = 0$, if input n' is entirely used in the C-node.

Two properties of the network distance function can be noted. $\mathcal{D}_i\left(x, \left(y^P, q\right)\right)$ is homogenous of degree 1 in inputs, i.e., $\mathcal{D}_i^t\left(\theta x, \left(y^P, q\right)\right) = \theta \mathcal{D}_i^t\left(x, \left(y^P, q\right)\right)$, for $\theta > 0$ and nondecreasing in q (and y^P), i.e., $\mathcal{D}_i^t\left(x, \left(\gamma y^P, \gamma q\right)\right) \leq \mathcal{D}_i^t\left(x, \left(y^P, q\right)\right)$, for $\gamma \geq 1$.

2.3. P-node productivity

Caves, Christensen and Diewert (1982) showed that productivity changes can be measured by Malmquist indices, defined in terms of ratios of distance functions. Following Färe, Grosskopf and Roos (1996) an input-based Malmquist productivity index can be defined for the production technology (P-Node) as

$$M_i^{t,t+1}\left(x^{P,t}, x^{P,t+1}, y^t, y^{t+1}\right) = \left[\frac{D_i^{P,t}\left(x^{P,t+1}, y^{t+1}\right)}{D_i^{P,t}\left(x^{P,t}, y^t\right)} \frac{D_i^{P,t+1}\left(x^{P,t+1}, y^{t+1}\right)}{D_i^{P,t+1}\left(x^{P,t}, y^t\right)}\right]^{\frac{1}{2}}. \tag{2.3}$$

Productivity improvement is signaled by a Malmquist index less than one and a negative change in productivity is signaled by an index greater than one. The Malmquist index can be decomposed into two components as in Färe, Grosskopf and Roos (1996)

$$M_i^{t,t+1}(\cdot) = \underbrace{\frac{D_i^{P,t+1}\left(x^{P,t+1}, y^{t+1}\right)}{D_i^{P,t}\left(x^{P,t}, y^t\right)}}_{E_i^{P,t,t+1}} \underbrace{\left[\frac{D_i^{P,t}\left(x^{P,t}, y^t\right)}{D_i^{P,t+1}\left(x^{P,t}, y^t\right)} \frac{D_i^{P,t}\left(x^{P,t+1}, y^{t+1}\right)}{D_i^{P,t+1}\left(x^{P,t+1}, y^{t+1}\right)}\right]^{\frac{1}{2}}}_{TC_i^{P,t,t+1}}. \tag{2.4}$$

The term outside the brackets, E_i, measures change in efficiency which can be interpreted as a "catching up" (to the production frontier) effect. The TC_i term measures change in technology in terms of shifts in the production frontier.

2.4. Network productivity

A network Malmquist productivity index based on the network distance functions can be defined analogously to the P-node index in (2.4) as

$$\mathcal{M}_i^{t,t+1}\left(x^t, x^{t+1}, y^{P,t}, y^{P,t+1}, q^t, q^{t+1}\right) = \underbrace{\frac{\mathcal{D}_i^{t+1}\left(x^{t+1}, y^{P,t+1}, q^{t+1}\right)}{\mathcal{D}_i^t\left(x^t, y^{P,t}, q^t\right)}}_{\mathcal{E}_i^{t,t+1}} \times \tag{2.5}$$

$$\underbrace{\left[\frac{\mathcal{D}_i^t\left(x^t, y^{P,t}, q^t\right)}{\mathcal{D}_i^{t+1}\left(x^t, y^{P,t}, q^t\right)} \frac{\mathcal{D}_i^t\left(x^{t+1}, y^{P,t+1}, q^{t+1}\right)}{\mathcal{D}_i^{t+1}\left(x^{t+1}, y^{P,t+1}, q^{t+1}\right)}\right]^{\frac{1}{2}}}_{\mathcal{TC}_i^{t,t+1}},$$

where $\mathcal{E}_i$ reflects the change in technical efficiency or "catching up" and $\mathcal{TC}_i$ reflects technology change as a shift in the network production frontier.

3. Data Envelopment Analysis

Using a sample of K firms the distance functions and the Malmquist productivity indices can be estimated using DEA. The data are collected in matrices: $X^P = (x_1^P, ..., x_K^P)$ and $X^C = (x_1^C, ..., x_K^C)$ which are $(N \times K)$ matrices of inputs, $Y = \begin{pmatrix} Y^P \\ Y^C \end{pmatrix} = \left(\begin{pmatrix} y_1^P \\ y_1^C \end{pmatrix}, ..., \begin{pmatrix} y_K^P \\ y_K^C \end{pmatrix}\right)$ which is a $((M+J) \times K)$ matrix of outputs and finally $Q = (q_1, ..., q_K)$ is an $(L \times K)$ matrix of quality assessments.

3.1. DEA estimators of distance functions

The estimate of the P-node input distance function (2.1), under constant returns to scale (CRS), is given by the solution to the linear program (c.f., Färe, Grosskopf and Roos (1996))

$$\left[\widehat{D}_i^{P,t}\left(x_k^{P,t}, y_k^t\right)\right]^{-1} = \min_z \left\{\theta : \theta x_k^{P,t} \geq X^{P,t} z, y_k^t \leq Y^t z, z \in R_+^K\right\}. \tag{3.1}$$

By adding an additional restriction on the intensity variables, z, alternative returns to scale restrictions can be imposed on the distance function estimate. The restriction $\sum_{k=1}^K z_k = 1$ imposes variable returns to scale (VRS) and the restriction $\sum_{k=1}^K z_k \leq 1$ imposes non increasing returns (see, e.g., Färe, Grosskopf and Lovell (1994)).

The network distance function (2.2), under CRS, is estimated by the following LP problem (c.f., Färe (1991), Färe and Grosskopf (1996))

$$\left[\widehat{\mathcal{D}}_i^t\left(x_k^t, y_k^{P,t}, q_k^t\right)\right]^{-1} =$$

$$\begin{aligned}\min_{\widetilde{x}^C,\widetilde{x}^P,\widetilde{y}^C,z_1,z_2} \{\lambda: \quad & \lambda x_k^t \geq \widetilde{x}^C + \widetilde{x}^P, \\ & \widetilde{x}^P \geq X^{P,t} z_1, y_k^{P,t} \leq Y^{P,t} z_1, \widetilde{y}^C \leq Y^{C,t} z_1, \\ & \widetilde{y}^C \geq Y^{C,t} z_2, \widetilde{x}^C \geq X^{C,t} z_2, q_k^t \leq Q^t z_2, \\ & z_1 \in R_+^K, z_2 \in R_+^K\Big\}.\end{aligned} \tag{3.2}$$

3.2. DEA estimators of the Malmquist productivity index

Given repeated observations on the sample of K firms, the Malmquist index can be estimated with DEA. In addition to (2.1) and (2.2), cross-period distance functions have to be estimated.

The P-node cross-period distance function $D_i^{P,t}\left(x^{P,t+1}, y^{t+1}\right)$ used in the Malmquist index in (2.3) is estimated analogous to (3.1) as

$$\left[\widehat{D}_i^{P,t}\left(x_k^{P,t+1}, y_k^{t+1}\right)\right]^{-1} = \min_z \left\{\theta : \theta x_k^{P,t+1} \geq X^{P,t} z, y_k^{t+1} \leq Y^t z, z \in R_+^K\right\}. \tag{3.3}$$

In (3.3) observations from period $t+1$ are evaluated against an estimate of the input requirement set $L^t\left(y_k^{t+1}\right)$, using observations from period t. The cross-period distance function $D_i^{P,t+1}\left(x^{P,t}, y^t\right)$ is estimated analogously, with the time indices t and $t+1$ interchanged.

Similarly, the network cross-period distance function $\mathcal{D}_i^{t+1}\left(x^t, y^{P,t}, q^t\right)$ used in (2.5) is estimated similarly as (3.2)

$$\left[\widehat{\mathcal{D}}_i^{t+1}\left(x_k^t, y_k^{P,t}, q_k^t\right)\right]^{-1} =$$

$$\begin{aligned}\min_{\widetilde{x}^C,\widetilde{x}^P,\widetilde{y}^C,z_1,z_2} \{\lambda: \quad & \lambda x_k^t \geq \widetilde{x}^C + \widetilde{x}^P, \\ & \widetilde{x}^P \geq X^{P,t+1} z_1, y_k^{P,t} \leq Y^{P,t+1} z_1, \widetilde{y}^C \leq Y^{C,t+1} z_1, \\ & \widetilde{y}^C \geq Y^{C,t+1} z_2, \widetilde{x}^C \geq X^{C,t+1} z_2, q_k^t \leq Q^{t+1} z_2, \\ & z_1 \in R_+^K, z_2 \in R_+^K\Big\}.\end{aligned} \tag{3.4}$$

The network cross-period distance function $\mathcal{D}_i^t\left(x^{t+1}, y^{P,t+1}, q^{t+1}\right)$ is estimated analogously, with the time indices $t+1$ and t interchanged.

4. Empirical Application

4.1. Data

An empirical application is included to illustrate the method and the types of conclusions that can be made from a comparison of a standard productivity model with the network model. The data consist of a sample of 31 Swedish pharmacies in 1993 and 1994. Färe, Grosskopf and Roos (1995, 1996) use similar type of data, but other subsamples of Swedish pharmacies. One novelty with the data set we use is that we have data available on the allocation of one of the two input dimension (the labor hours) to the two nodes. This implies that we can estimate the complete network model specification which allows differences in allocation of inputs to the P-node (traditional production) and the C-node (customer oriented). The allocation of inputs (labor hours) to the C- and P-node, is based on budget data in terms of ratios of total labor hours. Due to data limited availability we have to use the same allocation ratio (x^C/x^P) in both years.

The dataset include two inputs, three outputs, two quality attributes and five quality assessment dimensions. The following set of variables are used:

Two inputs (x^P) and (x^C):

LABOR-P and LABOR-C (hours of pharmacist and technical staff services in the P-node and C-node, respectively),

COSTO (the value of other inputs, SEK 1000)

Three outputs (y^P):

NUMPRE (number of outpatient prescriptions),

OTC (number of over the counter transactions),

OEXP (number of other types of expeditions).

Two output attributes (y^C):

OPEN (number of business hours open per week),

PRESERV (percent of prescriptions filled within one day).

Five CSB quality assessments (q):

AVAIL (availability of the pharmacy service),

PREMI (the pharmacy premises),

QPRESERV (service on prescription drugs),

PREFREE (service on prescription free drugs),

QUE (que service).

The five assessments dimensions represent impact transformed estimates of the different quality dimensions in the Customer Satisfaction Index (CSI). The data used in the estimation of the CSI are obtained from pharmacy specific customer satisfaction barometer surveys. The CSI is estimated using a Partial Least Square (PLS)-algorithm where the rating and the impact of the different quality dimensions on the CSI are estimated. In the analysis the transformation is simply a multiplication of the rating of each dimension of q with the corresponding impact estimate (see, e.g, Fornell (1992) for a discussion of the CSI-estimation procedure). Descriptive statistics of the data are given in Table 1.

Table 1
Descriptive statistics for the inputs, outputs, attributes and CSB-data

	Mean		**Min.**		**Max.**	
	1993	*1994*	*1993*	*1994*	*1993*	*1994*
Inputs:						
LABOR-P	3,754	3,863	408	426	11795	11703
LABOR-C	6,849	7,045	669	713	20969	20806
COSTO	588,40	670,70	148.10	174.40	1691.40	1804.90
Outputs:						
NUMPRE	46,763	45,197	10,313	9,074	129,640	134,260
OTC	52,584	51,361	9,438	8,938	174,601	173,567
OEXP	3995	8350	527	1184	14207	27329
Attributes:						
OPEN	44	44	26	27	62	62
PRESERV	86	89	0	0	98	98
CSB Data:						
AVAIL	36.7	39.3	7.4	8.0	130.5	84.0
PREMI	84.6	71.1	37.2	16.8	148.2	132.0
QPRESERV	54.6	44.3	0.0	0.0	133.5	136.0
PREFREE	55.5	41.1	8.9	0.0	109.2	77.4
QUE	36.9	44.6	15.6	17.2	62.4	106.6

From Table 1 we see that input use have increased on average. The average values of the quality attributes are nearly the same in both years. Data on outputs and quality assessments show no clear pattern although a high increase in other types of expeditions (OEXP) can be noted. This is caused by a change in technology for some pharmacies resulting in increased other types of expeditions partly accounted for by a decrease in outpatient prescriptions.

4.2. Results

We present (geometric) average results for the network Malmquist index and its components. Note that restrictions are imposed on the attributes $\left(y^C\right)$ since opening hours and the service level are bounded above by 168 and 100, respectively. Furthermore, results based on a "standard" model of the technology using the approach in Färe, Grosskopf and Roos (1995) are also presented. This (direct) model does not account for customer satisfaction. Quality attributes are, however, included as a subvector of the output vector, i.e., the output is given by the output from the P-node as $y = \left(y^P, y^C\right)$. The inputs are aggregated into a single input vector, consisting of the total (exogenous) inputs in the network model, i.e., the inputs $x = x^P + x^C$ are used.

Qualitative and quantitative comparisons of these results indicate how the inclusion of customer perceived quality (based on the network model) affect the estimated productivity scores. Implicit adjustment terms can be defined as the ratio of the network Malmquist index to the standard Malmquist index, i.e., $\widehat{AP} = \widehat{M}_i^{t,t+1}(\cdot) / \widehat{M}_i^{t,t+1}(\cdot)$. If $\widehat{AP} < 1$ (> 1), the effect of accounting for customer satisfaction is positive (negative) and productivity growth is higher (lower) when accounting for customer satisfaction.

Results from the two models are given in Table 2. Results on optimal values of the choice variables x^P, x^C and y^C in the estimation of the network distance functions (3.2) are given in Table 3.

Qualitatively, the two models give similar results in terms of sample averages. The network model, however, indicates a lower (average) productivity progress than the standard model. Both the efficiency and technical change components in the Malmquist index decomposition indicate progress in productivity.

The results show 13 (18) cases where the estimated productivity (change) is higher (lower) in the network model compared to the standard model. Although there are equally many cases with productivity progress, the pharmacies with increased productivity are not the same in the two models. Three pharmacies experienced increased productivity in the network model, but decreased productivity in the standard model and consequently there are three cases with the opposite results. We note that the models also give similar results in terms of rank comparisons. Six units are in the group with ten highest productivity scores in both models and seven units are in the group with ten lowest scores in both models.

Table 2
Summary results: Malmquist index, decompositions and implicit adjustment terms[1]

	Network Model	Standard Model	Adjustment terms
Productivity change:			
geometric mean	0.949	0.914	1.038
progress (counts)	21	21	13
regress (counts)	10	10	18
no change (counts)	0	0	0
Efficiency change:			
geometric mean	0.97	0.979	0.99
progress (counts)	16	18	16
regress (counts)	12	5	12
no change (counts)	3	8	3
Technical change:			
geometric mean	0.978	0.934	1.048
progress (counts)	20	21	17
regress (counts)	11	10	14
no change (counts)	0	0	0
# Efficient units	7 / 7	11 / 10	

On average the adjustment term is greater than one for the productivity and technical change, whereas it is less than (but very close to) one for the efficiency change component. We thus draw the conclusion that the inclusion of customer perceived quality lowers the estimated productivity and technical change for this sample.

The number of efficient units is less in the network model compared to the standard model, both in 1993 and 1994. The number of units with progress and regress in efficiency also differs between the two models.

Table 3
Descriptive statistics of optimal values of x^C, x^P and y^C

	Mean		**Min.**		**Max.**	
	1993	*1994*	*1993*	*1994*	*1993*	*1994*
Inputs:						
LABOR-P	4,294	5,238	409	506	26,876	23,608
LABOR-C	4,254	3,316	794	713	20,969	10,003
Attributes:						
OPEN	52	46	25	27	168	112
PRESERV	74	89	0	0	100	100

[1]In column 3 progress = # cases where $\widehat{AP} < 1$ and regress = # cases where $\widehat{AP} > 1$.

The results for optimal values of x^C, x^P and y^C in Table 3 reveal that the optimal levels of labor time allocated to customer activities are for both years, on average, less than the actual observed levels. The opposite results is obtained for labor time allocated to production. Equivalently, the ratio of optimal allocation to production activities over optimal allocation to customer oriented activities (i.e., x^P/x^C) is higher than the observed values. One interpretation of this is that more resources should be allocated to production activities and less to consumer activities.

For the output attributes the average results show that the optimal values of the open times exceed the observed values for both years. The optimal value of the service level is below the observed level in 1993 and equal to the observed level in 1994. An interpretation of this is that the pharmacies should substitute opening hours for service level. It can also be noted that the optimal values attain maximum values of both attributes in 1993 for one pharmacy.

5. Summary and Concluding Remarks

This paper proposes a network model which gives flexibility in the modeling of productivity and customer satisfaction. The model allows inputs to be allocated to customer oriented activities or traditional production activities. The inputs directed to customer activities, together with quality attributes of the production is assumed to give customer satisfaction, here represented by quality assessments from customer satisfaction barometer surveys. If these data are available, optimal allocation of the inputs to the production- and consumption nodes in the network can be estimated using DEA.

For firms with an objective to achieve a high degree of customer satisfaction, the approach gives valid measures of efficiency and productivity. This can be important for organizations with quality objectives and incentive structures based on productivity measures.

We present an empirical application using a sample of Swedish pharmacies. Similar qualitative results are obtained from the network model and a standard productivity model. In both models the estimated (average) productivity change is positive. The network model, however, shows a lower (average) productivity progress than the standard model. Both the efficiency and technical change components in the Malmquist index decomposition indicate progress in productivity. The technical change component accounts for most of the progress, although this is more clear in the standard model than in the network model. There are, however, some differences regarding which pharmacies are experiencing progress and regress in productivity in the different models. Furthermore, the results indicate that more resources should be allocated to the production activities and less to the consumer activities and that the pharmacies should substitute opening hours for service level.

To conclude, this paper has presented an approach to account for customer satisfaction in estimation of efficiency and productivity. For firms with customer satisfaction as a stated objective this type of model provides a way to obtain more valid measures

of efficiency and productivity than traditional approaches where customer perceived quality are often ignored. Furthermore, important results for management decisions regarding both allocation of resources and quality attribute prioritations are provided.

6. References

Banker, R. D., Charnes, A. and W. W. Cooper, (1984), "Some Models for Estimating Technical and Scale Inefficiencies in Data Envelopment Analysis", *Management Science*, Vol. 30, 1078-1092.

Caves, D., Christensen, L. and E. Diewert, (1982), "The Economic Theory of Index Numbers and the Measurement of Input, Output and Productivity", *Econometrica*, Vol. 50, 1393-1414.

Charnes, A., Cooper, W. W. and E. Rhodes, (1978), "Measuring the Efficiency of Decision Making Units", *European Journal of Operational Research*, Vol. 2, 429-444.

Fornell, C., (1992), "A National Customer Satisfaction Barometer: The Swedish Experience", *Journal of Marketing*, Vol. 56, 6-21.

Färe, R., (1991), "Measuring Farrell Efficiency for a Firm With Intermediate Inputs", *Academia Economic Papers*, Vol. 19, 329-340.

Färe, R., Grosskopf, S. and C. A. K. Lovell, (1994), "*Production Frontiers*", Cambridge University Press, Cambridge.

Färe, R. and S. Grosskopf, (1996), "Productivity and Intermediate Products: A Frontier Approach", *Economics Letters*, Vol. 50, 65-70.

Färe, R., Grosskopf, S. and P. Roos, (1995), "Productivity and Quality Changes in Swedish Pharmacies", *The International Journal of Production Economics*, Vol. 29, 137-144.

Färe, R., Grosskopf, S. and P. Roos, (1996), "Integrating Consumer Satisfaction into Productivity Indexes", Working Paper 1996:4, The Swedish Institute for Health Economics.

Lancaster, K., (1991), "*Modern Consumer Theory*", Edward Elgar Publishing, Hants.

Shephard, R. W., (1970), "*The Theory of Cost and Production Functions*", Princeton University Press, Princeton.