J Prod Anal (2011) 36:219–230
DOI 10.1007/s11123-011-0231-5

One-stage estimation of the effects of operational conditions and practices on productive performance: asymptotically normal and efficient, root-*n* consistent StoNEZD method

Andrew L. Johnson · Timo Kuosmanen

Published online: 8 June 2011

Abstract Understanding the effects of operational conditions and practices on productive efficiency can provide valuable economic and managerial insights. The conventional approach is to use a two-stage method where the efficiency estimates are regressed on contextual variables representing the operational conditions. The main problem of the two-stage approach is that it ignores the correlations between inputs and contextual variables. To address this shortcoming, we build on the recently developed regression interpretation of data envelopment analysis (DEA) to develop a new one-stage semi-nonparametric estimator that combines the nonparametric DEA-style frontier with a regression model of the contextual variables. The new method is referred to as stochastic semi-nonparametric envelopment of z variables data (StoNEZD). The StoNEZD estimator for the contextual variables is shown to be statistically consistent under less restrictive assumptions than those required by the two-stage DEA estimator. Further, the StoNEZD estimator is shown to be unbiased, asymptotically efficient, asymptotically normally distributed, and converge at the standard parametric rate of order $n^{-1/2}$. Therefore, the conventional methods of statistical testing and confidence intervals apply for asymptotic inference. Finite sample performance of the proposed estimators is examined through Monte Carlo simulations.

A. L. Johnson (✉)
Department of Industrial and Systems Engineering, Texas A&M University, 237K Zachry Engineering Center, College Station, TX 77843-3131, USA
e-mail: ajohnson@tamu.edu

T. Kuosmanen
School of Economics, Aalto University, Runeberginkatu 22-24, 00101 Helsinki, Finland
e-mail: timo.kuosmanen@aalto.fi

Keywords Data envelopment analysis · Two-stage method · Partial linear model

JEL Classification C14 · C51 · D24

1 Introduction

A firm is said to operate technically efficiently if it produces the maximum output for a given level of input use.[1] In practice, the level of technical efficiency depends on operational conditions, such as the external operational environment in which production occurs, the internal characteristics of the firms such as the type and vintage of technology, and the managerial practices. Following Banker and Natarajan (2008), we refer to the categorical, ordinal, interval or ratio scale data that characterize operational conditions and practices as *contextual variables*.

Understanding the effects of contextual variables on performance can provide valuable information for managers who develop business strategies or make decisions on operational practices, and for policy makers who may influence the external operating environment of firms through standards, regulations, taxes, subsidies, and other policy measures. Investigating the effects of contextual variables on efficiency has attracted a lot of deserved attention in the literature of productive efficiency analysis. One of the first discussions is by Hall and Winsten (1959), who hold a rather pessimistic view regarding estimating these effects, asserting that production processes in

[1] The term "firm" here refers to any production unit that transforms inputs to output, including both non-profit and for-profit organizations. The firm can refer to a facility or sub-division of a company or to an aggregate entity such as an industry, a region, or a country.

different contexts are not comparable. Succeeding studies develop models in which efficiency and the effects of the contextual variables are estimated either sequentially in two stages or simultaneously in a single stage. Timmer (1971) pioneered the two-stage parametric regression method where the frontier production function and the firm-level technical efficiency are first estimated by regressing the output on inputs, and the estimated efficiency measures are subsequently regressed on the contextual variables. However, it has been established that the second stage estimator is biased and inconsistent when the inputs are correlated with the contextual variables (Wang and Schmidt 2002). In the literature of stochastic frontier analysis (SFA) it is nowadays well-understood that ignoring the effect of contextual variables in the first stage regression results as omitted variable bias, which carries over to the second stage regression.

Ray (1988, 1991) was the first to apply the nonparametric data envelopment analysis (DEA) method in the first stage of the two-stage approach. In the two-stage DEA method, the efficient frontier and the firm-level efficiency scores are first estimated by DEA. In the second stage, the estimated DEA efficiency scores are regressed on contextual variables. In the second stage, a variety of regression techniques have been used, including the classic ordinary least squares (OLS) and the maximum likelihood (ML) based probit, logit, truncated regression, among others (e.g., Simar and Wilson 2007; Estelle et al. 2010).

The statistical properties of the two-stage DEA estimator have recently attracted critical debate. As an attempt to rationalize studies where two-stage DEA is used, Simar and Wilson (2007) develop a formal statistical model that can be estimated by using truncated regression in the second stage. Most importantly, they argue that the conventional statistical inferences are invalid in the second-stage regression, and propose to solve this problem by using the bootstrap method.[2] In a separate line of research, Banker and Natarajan (2008) attempt to validate the two-stage DEA approach by proving consistency of the second-stage OLS estimator of the contextual variables, under certain assumptions. Subsequent discussion has focused on the assumptions. Banker and Natarajan argue that their statistical model requires less restrictive assumptions than the model of Simar and Wilson (2007). Simar and Wilson (2010) respond by outlining a list of seven assumptions in the Banker and Natarajan paper that they find restrictive. Our parallel paper (Johnson and Kuosmanen 2010) elaborates the assumptions and the statistical properties of the two-stage estimators further. In that paper we emphasize the correlation of the contextual variables with the inputs as a major source of bias, analogous to the omitted variable bias of the two-stage parametric regression.

Noting that the parametric SFA approach has evolved from the two-stage approach to the one-stage method where the frontier and the effects of contextual variables are jointly estimated, the purpose of this paper is to show that similar development is possible in the axiomatic, semi-nonparametric approach. Relaxing some of the restrictive assumptions of the Banker and Natarajan (2008) model, we develop a new one-stage semi-nonparametric estimator that combines the axiomatic, DEA-style production frontier with the parametric regression of the contextual variables. Our estimator builds on the nonparametric least-squares interpretation of DEA developed in Kuosmanen and Johnson (2010) and the StoNED method of Kuosmanen and Kortelainen (2011) that combines the axiomatic, nonparametric DEA-style frontier with the probabilistic, SFA-style treatment of inefficiency and noise.[3] In this respect, the developments of this paper can be seen as an extension of the StoNED method to accommodate contextual variables. We therefore refer to the method proposed in this paper as stochastic semi-nonparametric envelopment of z variables data (StoNEZD).

The proposed StoNEZD estimator has several advantages for estimating the effects of contextual variables. We show that the StoNEZD-estimator is statistically consistent under more general assumptions than those required by the two-stage DEA. In particular, the StoNEZD-estimator remains consistent even when the noise term is unbounded and the contextual variables are correlated with inputs. Further, we show that the StoNEZD-estimator for the contextual variables is asymptotically efficient, asymptotically normally distributed, and converges at the standard parametric rate of order $n^{-1/2}$. These are the key properties to ensure that the conventional methods of statistical inference (such as t tests for the significance and the confidence intervals) are valid for asymptotic inference, despite the nonparametric specification of the frontier and the presence of an asymmetric, non-normal disturbance term.

To complement the asymptotic properties established in this paper, we examine the finite sample performance of the proposed estimator by means of Monte Carlo simulation. We replicate the data generation process (DGP) considered by Banker and Natarajan (2008). The results

[2] According to Simar and Wilson (2010), they do not advocate this method. Rather, their intention is to rationalize the work on two-stage DEA and to demonstrate one way to perform statistical inference.

[3] Kuosmanen (2006) originally abbreviated StoNED for stochastic nonparametric envelopment of data. At the request of a referee, Kuosmanen and Kortelainen (2011) changed the name as stochastic nonsmooth envelopment of data. The StoNED approach combines elements from both parametric and nonparametric frontier methods, so it could be best described as semi-nonparametric (see Chen 2007; footnote 1).

show that the StoNEZD-estimator yields systematically more precise estimates than two-stage DEA in virtually all scenarios considered, whether noise is present or not. Examining performance of the proposed estimation procedures via Monte Carlo simulations provides strong support for using the one-stage method for joint estimation of the frontier and the contextual variables.

The rest of the paper is organized as follows. Section 2 introduces the theoretical model to be estimated, together with the necessary notation and assumptions. Section 3 develops the StoNEZD method for joint estimation of the production frontier and the influence of the contextual variables. Section 4 describes the Monte Carlo simulation results to demonstrate the performance of the proposed methods compared to two-stage DEA. Finally, Sect. 5 provides concluding remarks and suggestions for future research. The proofs of theorems are provided in "Appendix".

2 Model and assumptions

Building on Banker and Natarajan's (2008) model, we consider a multiple input, single output, cross-sectional production model

$$y_i = \phi(\mathbf{x}_i) \cdot e^{\varepsilon_i}, \quad i = 1, \ldots, n \tag{1}$$

where y_i denotes the output of firm i, $\mathbf{x}_i \in \Re_+^m$ is a vector of inputs, $\phi : \Re_+^m \rightarrow \Re_+$ is a classic (increasing and concave) frontier production function, and ε_i is a composite disturbance term that captures all deviations from the best-practice production function. The disturbance term ε_i consists of three specific components according to the equation

$$\varepsilon_i = \mathbf{z}_i'\boldsymbol{\delta} - u_i + v_i, \tag{2}$$

where $\mathbf{z}_i \in \Re^r$ denotes a column vector of contextual variables that characterize the measured values of operational conditions and practices of firm i, and $\boldsymbol{\delta} = (\delta_1 \ldots \delta_r)'$ is a vector of unknown parameters (to be estimated), representing the average effect of contextual variables $\mathbf{z}_i$ on performance, $u_i \geq 0$ is a random inefficiency term of firm i that represents technical inefficiency, and v_i is a random disturbance term that represents the effects of omitted factors, measurement errors, and other stochastic noise.

Empirical data of a sample of n firm is characterized by the output vector $\mathbf{y}$, $n \times m$ input matrix $\mathbf{X}$, $n \times r$ matrix of contextual variables $\mathbf{Z}$, and use $\mathbf{0} = (0 \ldots 0)'$ and $\mathbf{1} = (1 \ldots 1)'$ with appropriate dimensions. All vectors are column vectors (unless otherwise indicated). We assume that no columns of the $\mathbf{X}$ and $\mathbf{Z}$ matrices are linearly dependent.

Introducing the contextual variables as a part of the composite disturbance term may suggest that we implicitly assume the contextual variables $\mathbf{z}$ influence technical efficiency, and not the frontier ϕ. According to this interpretation, $\mathbf{z}_i'\boldsymbol{\delta} - u_i$ can be seen as the overall efficiency of firm i, where the term $\mathbf{z}_i'\boldsymbol{\delta}$ represents technical inefficiency that is explained by the contextual variables, and the component u_i represents the proportion of inefficiency that remains unexplained. Equivalently, one could model the frontier as $\phi(\mathbf{x}) \cdot e^{\mathbf{z}'\boldsymbol{\delta}}$, assuming the contextual variables influence the technology and not the efficiency. Both interpretations are equally valid, and we have no particular reason to prefer one over another. We are interested in estimating the net effect of contextual variables on output y. Whether $\mathbf{z}$ variables influence the output through the technology or the level of efficiency (or both) is another question that falls beyond the scope of this paper. Of course, one can simply take either interpretation as the maintained assumption, but identifying the two effects from empirical data is very challenging.

For the parametric part of the model (i.e., $\mathbf{z}_i'\boldsymbol{\delta}$) we assume a linear functional form, noting that the elements of $\mathbf{z}_i$ can first be transformed by using suitable data transformation (e.g., exponential or logarithmic). Further, vector $\mathbf{z}_i$ may well include quadratic, cubic or higher-order polynomial transformations of the original data on operational conditions and practices. Moreover, elements of $\mathbf{z}_i$ may be binary dummy variables that can be used as numerical representations of categorical and/or ordinal data.

Distributional assumptions on the random inefficiency and noise terms v_i and u_i are not imposed; rather, we assume the existence of well-behaved probability density functions f_u and f_v. The function f_v is assumed to be symmetric with zero-mean, and f_u is left-truncated at zero (i.e., $f_u(u) = 0$ for all $u < 0$). We denote the expected inefficiency as $\mu = E(u_i) > 0$. Variables v_i and u_i are assumed to be independently distributed random variables that are uncorrelated with the input variables $\mathbf{x}_i$, the contextual variables $\mathbf{z}_i$, and with each other. However, we do not assume $\mathbf{z}_i$ and $\mathbf{x}_i$ to be uncorrelated.

The theoretical model specified above addresses one of Simar and Wilson's (2007) critiques by presenting a coherent DGP in the semi-nonparametric frontier estimation. The recent paper Simar and Wilson (2010) outlines seven assumptions regarding Banker and Natarajan's (2008) model that can be seen as restrictive: (1) The input variables and the contextual variables are independently distributed. (2) A bounded noise distribution (Gstach 1998). (3) The direction of the effect of the contextual variables is known a priori. (4) Homoskedasticity and homogeneity of the noise and inefficiency terms. (5) A single output. (6) The contextual variables have a monotonic influence on the frontier. (7) The contextual variables only affect the production possibility set, but not the level of inefficiency.

While the model considered in this paper is based on Banker and Natarajan (2008), assumptions (1), (2), and (3)

are not required. In these respects it is evident that our model is more general than that of Banker and Natarajan (2008). We do impose the homoskedasticity assumptions (4) to prove some of useful properties, but the most essential properties continue to hold in the heteroskedastic case (see Sect. 3). Regarding (5), we consider the single output case to maintain a direct contact with the SFA models, but the method can be applied in the multi-output setting by applying the directional distance functions (see Kuosmanen 2006; Kuosmanen and Johnson 2011, for further discussion). Points (6) and (7) were already mentioned above. We will comment on some of these points in more detail in Sect. 3.

Simar and Wilson (2007) examine a model where z variables influence the probability distribution of the random inefficiency term u. This is a major structural difference that makes Simar and Wilson's models difficult to compare to two-stage DEA or StoNEZD. Assuming the z variables affect the density of u is equally valid (or arbitrary) as assuming the z variables influence the frontier. Further, Simar and Wilson (2007) impose several additional assumptions such as no noise, smoothness of the true frontier, and truncated normality of the second-stage disturbance term. In these respects, the model considered in this paper is more general than Simar and Wilson's model. In particular, we find the assumption of no noise particularly restrictive.

To conclude, we stress that the main critique of two-stage DEA by Simar and Wilson (2007) concerns the statistical inferences in the second-stage regression: they state that the residuals in the second-stage regression are serially correlated in finite samples because the dependent variables (the DEA efficiency estimates) are correlated with each other, and that this correlation does not disappear quickly enough as the sample size increases. The one-stage method developed in the next section overcomes this shortcoming, as well as some other limitations of the two-stage DEA.

3 One-stage semi-nonparametric estimators with contextual variables

Consider the two-stage DEA method as a combination of two optimization problems: the first optimization problem computes the DEA efficiency estimates, and second is a least squares or ML problem to estimate the impacts of contextual variables on efficiency. However, the impact of the contextual variables is not taken into account in the first stage. In the absence of noise, the DEA estimator is known to be biased in finite samples, but the omitted contextual variables can further deteriorate the finite sample bias of DEA (see Johnson and Kuosmanen 2010, for a more detailed discussion). Clearly, it would be advantageous to integrate the two optimization problems into one problem that jointly estimates the frontier and the impacts of the contextual variables on efficiency. The equivalence result established by Kuosmanen and Johnson (2010) provides the means to integrate the two optimization problems using convex nonparametric least squares (CNLS: Hildreth 1954; Hanson and Pledger 1976; Groeneboom et al. 2001).

To solve the CNLS problem, Kuosmanen (2008) has shown that the set of continuous, monotonic increasing and globally concave regression functions can be equivalently represented by a family of piece-wise linear functions that are characterized using the Afriat inequalities (Afriat 1967, 1972). Starting with Eq. 1, we can take the natural log of both sides. Kuosmanen and Kortelainen (2011) develop the CNLS formulation for estimating the log-transformed model without contextual variables $\mathbf{z}$ (Sect. 4.3, Eq. 30). Conveniently, the contextual variables $\mathbf{z}$ can be directly incorporated to the objective function, with no effect on the system of Afriat inequalities imposed in the constraints. This yields the StoNEZD problem

$$\begin{aligned} &\min_{\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\delta},\hat{\phi}} \sum_{i=1}^{n} (\ln y_i - \ln \hat{\phi}_i - \mathbf{z}_i'\boldsymbol{\delta})^2 \\ &s.t. \\ &\hat{\phi}_i = \alpha_i + \mathbf{x}_i\boldsymbol{\beta}_i \;\forall i = 1,\ldots,n \\ &\alpha_i + \mathbf{x}_i\boldsymbol{\beta}_i \le \alpha_h + \mathbf{x}_i\boldsymbol{\beta}_h \;\forall h,i = 1,\ldots,n \\ &\boldsymbol{\beta}_i \ge \mathbf{0} \;\forall i = 1,\ldots,n \end{aligned} \tag{3}$$

Problem (3) is a nonlinear programming (NLP) problem that has a nonlinear objective function and a system of linear inequality constraints. It can be solved by using standard NLP algorithms and solvers such as MINOS, CONOPT, KNITRO or PATHNLP.

In Eq. 3, the first constraint estimates α_i and $\boldsymbol{\beta}_i$ for each observation; thus n different regression lines are estimated instead of fitting one regression line to the cloud of observed points as in OLS. Note that by replacing the Afriat inequalities and the constraint $\boldsymbol{\beta}_i \ge \mathbf{0}\;\forall i$ with $\alpha_i = \alpha_j, \boldsymbol{\beta}_i = \boldsymbol{\beta}_j\;\forall i,j = 1,\ldots,n$, the standard OLS problem is obtained. In problem (3), the n estimated lines can be interpreted as tangent lines to ϕ. The slope coefficients $\boldsymbol{\beta}_i$ represent the marginal products of inputs (i.e., the subgradients $\nabla\phi(\mathbf{x}_i)$). The second constraint imposes concavity by applying a system of Afriat inequalities (see Kuosmanen 2008 for details). The third constraint imposes monotonicity.

In problem (3), parameter vector $\boldsymbol{\delta}$ is common to all observations. The contextual variable extension can be seen as a restricted special case of the models presented in Kuosmanen and Johnson (2010) and Kuosmanen and Kortelainen (2011) where $\mathbf{Z}$ is a subset of $\mathbf{X}$ for which $\boldsymbol{\beta}_i = \boldsymbol{\beta}_j\;\forall i,j = 1,\ldots,n$ or with a change in notation $\boldsymbol{\delta}_i = \boldsymbol{\delta}_j\;\forall i,j = 1,\ldots,n$. Regardless of how the CNLS residuals are further decomposed to inefficiency and noise

terms, this CNLS method can be used for estimating the shape of the production function and the effects of contextual variables simultaneously in a one-stage convex programming problem.

Kuosmanen and Johnson (2010) have shown (in the absence of contextual variables $\mathbf{z}$) that restricting the deviations of the observed outputs y_i from the estimated frontier $\hat{\phi}_i$ as non-positive yields the standard DEA frontier (under variable returns to scale).[4] Similarly, we could add the sign constraint $\ln y_i - \ln \hat{\phi}_i - \mathbf{z}_i'\boldsymbol{\delta} \leq 0 \; \forall i$ in the StoNEZD problem (3). The resulting one-stage DEA formulation and its properties are examined in detail the parallel paper Johnson and Kuosmanen (2010).

The StoNEZD estimator (3) that does not restrict the sign of the least squares residuals has certain advantages to the sign-constrained DEA estimator that become evident by examing the statistical properties. The finite sample bias of the one-stage and two-stage DEA estimator for the effects of contextual variables is elaborated in Johnson and Kuosmanen (2010). Regarding the potential bias of the StoNEZD, we can show the following.

Theorem 1 *If the data-generating process satisfies the maintained assumptions stated in Sect. 2, the StoNEZD-estimator for the coefficients of the contextual variables* $(\hat{\boldsymbol{\delta}}^S)$ *is statistically unbiased*:

$$E(\hat{\boldsymbol{\delta}}^S) = \boldsymbol{\delta}. \tag{4}$$

Importantly, the StoNEZD-estimator is unbiased even when contextual variables are correlated with inputs, irrespective of whether stochastic noise is present or not. Further, the sign of $\boldsymbol{\delta}$ does not need to be specified in advance.

Regarding the asymptotic properties, Hanson and Pledger (1976) have formally proven consistency of the original CNLS estimator, and Groeneboom et al. (2001) have established its asymptotic distribution and the rates of convergence. The following theorems extend the known asymptotic results of the CNLS estimator to the StoNEZD coefficients of the contextual variables.

Theorem 2 *If the following conditions are satisfied*

(i) *sequence,* $\{(y_i, \mathbf{x}_i, \mathbf{z}_i), \quad i = 1, \ldots, n\}$ *is a random sample of independent observations,*

(ii) $\lim_{n\to\infty} \mathbf{Z}'\mathbf{Z}/n$ *is a positive definite matrix,*

(iii) *the inefficiency terms* $\mathbf{u}$ *and the noise terms* $\mathbf{v}$ *are identically and independently distributed (i.i.d.) random variables with* $\mathrm{Var}(\mathbf{u}) = \sigma_u^2\mathbf{I}$ *and* $\mathrm{Var}(\mathbf{v}) = \sigma_v^2\mathbf{I}$,

then the StoNEZD-estimator for the coefficients of the contextual variables $(\hat{\boldsymbol{\delta}}^S)$ *is statistically consistent and asymptotically normally distributed according to:*

$$\hat{\boldsymbol{\delta}}^S \sim_a N(\boldsymbol{\delta}, (\sigma_v^2 + \sigma_u^2)(\mathbf{Z}'\mathbf{Z})^{-1}). \tag{5}$$

The proof of consistency of the StoNEZD-estimator allows for the standard normally distributed noise. In contrast, Banker and Natarajan's (2008) proof of the consistency of the two-stage DEA estimator critically depends on a truncated distribution of noise. Further, we do not need to assume the sign of $\boldsymbol{\delta}$ in advance. Compared to the assumptions stated in Theorem 1 of Kuosmanen and Johnson (2010), we have only imposed the additional assumption (iii) that the inefficiency and noise terms are identically and independently distributed. It is worth noting that the assumption (iii) is needed for deriving the limiting distribution, but relaxing assumption (iii) would not influence the consistency. Therefore, the StoNEZD-estimator $\hat{\boldsymbol{\delta}}^S$ is consistent under more general assumptions than the two-stage DEA estimators.

The previous paragraph regarding assumption (iii) relates to the issue of heteroskedasticity, noted by Simar and Wilson in their point (4). The classic OLS estimator is known to be unbiased and consistent under heteroskedasticity, but more efficient estimators are available if some specific model of heteroskedasticity is assumed. It is straightforward to verify that the same properties carry over to the StoNEZD-estimator $\hat{\boldsymbol{\delta}}^S$: the proofs of unbiasedness and consistency in Theorems 1 and 2 do not critically depend on the homoskedasticity assumption. Specifically, the estimator of the nonparametric frontier $\hat{\phi}^S$ can be biased under heteroskedasticity, but this does not affect the unbiasedness of $\hat{\boldsymbol{\delta}}^S$ if the bias of $\hat{\phi}^S$ is uncorrelated with the contextual variables $\mathbf{z}$. Of course, if heteroskedasticity is due to contextual variables, or inputs that are highly correlated with the contextual variables, then the bias will carry over to $\hat{\boldsymbol{\delta}}^S$ (see the proof of Theorem 1 in "Appendix" for details). To test for the presence of significant heteroskedasticity in the StoNEZD residuals, the standard White test (White 1980), can be applied. Note that the White test is based on the least squares residuals, and it does not assume any particular regression function or specification of heteroskedasticity.

The asymptotic normality established in Theorem 2 is a very useful property in practice. This property allows us to apply the conventional methods of statistical inference from the linear regression analysis despite the presence of a

[4] While the DEA problems are usually solved by running a separate LP problem for each firm, there is a stream of DEA papers that integrate the n LP problems into one LP problem to ensure that the multiplier weights of some inputs/outputs are the same across all firms (e.g., Li and Ng 1995; Kuosmanen et al. 2006).

nonparametric frontier ϕ and an asymmetric inefficiency term u. For example, we may test the statistical significance of the contextual variables by the standard t test, derive confidence intervals for the $\boldsymbol{\delta}$ parameters, and test hypotheses on $\boldsymbol{\delta}$. This kind of statistical analyses allows us to identify statistically significant operational practices or production attributes, or test more specific hypotheses regarding the effects of the context. It is important to emphasize that the conventional standard errors are biased and inconsistent under heteroskedasticity. This is another important reason why heteroskedasticity is a concern in the present setting. If the White test indicates significant heteroskedasticity, the robust standard errors can be used, as suggested in White (1980) and MacKinnon and White (1985).

It is worth emphasizing that asymptotic normality does not depend on any distributional assumptions on the inefficiency or the noise terms, but stems from the central limit theorem. The conventional inference procedures are valid in very large samples, but inferences in the finite samples require that the $\hat{\boldsymbol{\delta}}^S$ converge quickly enough. Therefore, the following theorem is important.

Theorem 3 *If the conditions stated in Theorem 2 are satisfied and the skewness of the inefficiency terms u_i is finite such that $E(|u_i - \mu|^3) = \gamma < \infty$, then the StoNEZD-estimator for the coefficients of the contextual variables ($\hat{\boldsymbol{\delta}}^S$) converges to the true $\boldsymbol{\delta}$ at the standard parametric rate on the order of $n^{-1/2}$. Specifically, there exist a positive constant C such that for all n,*

$$|\hat{\boldsymbol{\delta}}^S - \boldsymbol{\delta}| \leq \frac{C\gamma}{\sqrt{n}(\sigma_v + \tilde{\sigma}_u)^3}\mathbf{1}. \tag{6}$$

Under the very mild and natural additional assumption of a finite absolute skewness of the inefficiency term, $\hat{\boldsymbol{\delta}}^S$ converge at the same rate as the usual parametric estimators. This important result ensures that the conventional methods of statistical inference are valid even in finite samples. Further, many of the tools available in econometrics such as hypothesis testing, confidence interval construction, and point estimates depend on the underlying estimated parameter being asymptotically normal. This addresses Simar and Wilson's (2007) primary concern regarding two-stage DEA. Further, the previous formal theorems show that the StoNEZD estimator addresses many of the restrictive assumptions of the two-stage DEA method, as identified by Simar and Wilson (2010).

Although the StoNEZD-estimator $\hat{\boldsymbol{\delta}}^S$ shares many of the attractive properties of the classic OLS estimator, efficiency of $\hat{\boldsymbol{\delta}}^S$ cannot be ensured in finite samples due to the very general specification of the nonparametric frontier ϕ. However, the following result on asymptotic efficiency can be established:

Theorem 4 *If the conditions stated in Theorem 2 are satisfied, then the StoNEZD-estimator for the coefficients of the contextual variables ($\hat{\boldsymbol{\delta}}^S$) is asymptotically efficient. That is,*

$$\text{AsyVar}(\hat{\boldsymbol{\delta}}^S) \leq \text{AsyVar}(\hat{\boldsymbol{\delta}}) \tag{7}$$

for any other consistent, asymptotically normally distributed estimator $\hat{\boldsymbol{\delta}}$.

This result shows that the StoNEZD-estimator has the smallest variance and the smallest mean squared error (MSE) of any potential estimator that permits conventional statistical inferences in large samples. Of course, in small samples other estimators may also be competitive. Therefore, it is interesting to examine the finite sample performance of alternative estimators within the controlled environment of Monte Carlo simulations show in Sect. 4.

Before we proceed to the Monte Carlo study, the estimation of the inefficiency terms u_i deserves further elaboration. The residuals of the CNLS model can be recovered as $\varepsilon_i^S = \ln y_i - \ln \hat{\phi}_i^S - \mathbf{z}_i'\hat{\boldsymbol{\delta}}^S$. These residuals will capture the combined effect of the inefficiency term u_i, the noise terms v_i, and the expected inefficiency μ, specifically, $\mu - u_i + v_i$. Note that $E(\mu\mathbf{1} - \mathbf{u} + \mathbf{v}) = \mathbf{0}$. The firm-specific inefficiency terms u_i remain unidentified under the general maintained assumptions imposed in Sect. 2; some further assumptions must be imposed to disentangle u_i.

In cross sectional data, we may try to identify the firm-specific inefficiency terms by imposing some distributional assumptions on both $\mathbf{v}$ and $\mathbf{u}$. In the parametric literature, the classic SFA model by Aigner et al. (1977) (see also Meeusen and Van den Broek 1977) assumes that noise terms $\mathbf{v}$ are normally distributed (specifically, $\mathbf{v} \sim N(\mathbf{0}, \sigma_v^2\mathbf{I})$) and the inefficiency terms $\mathbf{u}$ are half-normally distributed (i.e., $\mathbf{u} \sim |N(\mathbf{0}, \sigma_u^2\mathbf{I})|$). In the present semi-nonparametric setting, Kuosmanen (2006) and Kuosmanen and Kortelainen (2011) have proposed to estimate the parameters σ_v, σ_u based on the CNLS residuals.

4 Monte Carlo simulations

This section examines the finite sample performance of the methods StoNEZD and two-stage DEA in the controlled environment of Monte Carlo simulations. We first describe the DGP used in the simulations and the performance statistics, and then report and discuss the results.

4.1 Design of experiments

To ensure comparability with earlier Monte Carlo evidence, we replicate the main features of the DGP used by Banker and Natarajan (2008). Their model can be summarized as

$$y_i = (x_i^3 - 12x_i^2 + 48x_i - 37) \cdot \exp(z_i\delta - u_i + v_i), \quad i = 1, \ldots, n \tag{8}$$

The true production function is a third-order polynomial $\phi(x) = (x^3 - 12x^2 + 48x - 37)$ of a single input variable x. This function is continuous, monotonic increasing, and concave over the relevant range of inputs used in the simulations.

The inputs x are randomly sampled from $Uni[1,4]$, the noise terms v are drawn from $N(0, \sigma_v^2)$, and the inefficiency terms u from $|N(0, \sigma_u^2)|$. Random variables v and u are drawn independently from x and each other. Following Wang and Schmidt (2002), we assume that the contextual variables are dependent on the input levels, and are generated according to equation

$$z = \rho[(x-1)/3] + w\sqrt{1-\rho^2} \tag{9}$$

where parameter ρ is some pre-specified correlation coefficient and w is an independent random number drawn from $Uni[0,1]$. The values of $\rho \in \{-0.8, -0.4, 0.0, +0.4, +0.8\}$ are considered. Note that when $\rho = 0$, the contextual variable z has the same distribution as in the baseline model of Banker and Natarajan.

Performance of the estimator is evaluated in terms of the root mean squared deviation (RMSD) criterion, defined for the coefficient δ as

$$\mathrm{RMSD}(\hat{\delta}) = \frac{100}{\delta} \times \sqrt{\frac{1}{M}\sum_{t=1}^{M}(\hat{\delta}_t - \delta)^2} \tag{10}$$

where δ is the true coefficient (initially set as $\delta = -0.2$) and M is the number of simulation trials. The RMSD is always greater than or equal to zero: the small values of RMSD indicate good performance. Each scenario was replicated 100 times (i.e., $M = 100$).

4.2 Results

Considered first the Banker and Natarajan's base case where the parameter values are $\delta = -0.2$, $\sigma_u = 0.15$, $\sigma_v = 0.04$, $n = 100$. Table 1 reports the RMSD statistics for the two methods under varying sample sizes n and levels of the correlation coefficient ρ. Two-stage DEA is implemented as regressing the output-oriented variable returns to scale DEA efficiency estimates on contextual variable z by OLS.

Table 1 Performance (by RMSD) of alternative methods in estimating the effect of contextual variable (δ) varying the number of firms (n) and the correlation of x and z (ρ). Baseline scenario

Scenario	Correlation (ρ)	Estimation method	
		Two-stage DEA	StoNEZD
$n = 50$	−0.8	87.5	40.9
	−0.4	44.2	27.7
	0.0	27.7	25.7
	0.4	27.7	27.9
	0.8	50.4	41.7
$n = 100$	−0.8	74.0	29.1
	−0.4	32.2	19.5
	0.0	18.9	17.9
	0.4	24.1	17.7
	0.8	57.6	30.1
$n = 200$	−0.8	83.0	17.4
	−0.4	30.2	11.3
	0.0	10.8	10.3
	0.4	12.5	11.2
	0.8	36.5	15.3

Base case: $\delta = -0.2$, $\sigma_u = 0.15$, $\sigma_v = 0.04$, $M = 100$

The results in Table 1 show a major advantage for the StoNEZD-estimator. The difference relative to two-stage DEA is notable especially when the contextual variable is correlated with the input. Large absolute value of ρ has a negative effect on the performance of the StoNEZD methods, but the effect is much smaller than in the case of two-stage DEA.

We next repeated the simulations under different parameter values and examine how the signal and noise influence the performance of the two methods. Let us first examine the impact of decreasing the coefficient of contextual variable to $\delta = -0.4$. This means that the contextual variable z has a larger (negative) effect on firms' performance, which should be easier for our estimators to capture. Table 2 reports the analogous results to Table 1 for this new scenario. The main differences can be observed in the precision of the estimated δ coefficients. Performance of two-stage DEA does not change much, but StoNEZD clearly improves in precision when the signal in δ increases.

Table 3 considers decreasing the coefficient of contextual variable to $\delta = -0.6$. The improvement in the precision of the StoNEZD estimator is notable for all correlation levels; the errors in terms of RSMD are less than ten percent.

The noise term can also be expected to have a major impact on the precision of the estimators. We next consider

Table 2 Performance (by RMSD) of alternative methods in estimating the effect of contextual variable (δ) varying the number of firms (n) and the correlation of x and z (ρ). Double signal scenario

Scenario	Correlation (ρ)	Estimation method	
		Two-stage DEA	StoNEZD
$n = 50$	−0.8	79.2	25.6
	−0.4	34.8	17.2
	0.0	17.3	15.8
	0.4	20.2	17.2
	0.8	45.7	25.5
$n = 100$	−0.8	78.0	12.8
	−0.4	30.5	8.52
	0.0	11.3	9.71
	0.4	13.4	8.41
	0.8	51.1	15.0
$n = 200$	−0.8	72.0	10.0
	−0.4	24.7	6.73
	0.0	7.13	6.10
	0.4	12.5	6.60
	0.8	44.6	10.0

Case: $\delta = -0.4$, $\sigma_u = 0.15$, $\sigma_v = 0.04$, $M = 100$

Table 3 Performance (by RMSD) of alternative methods in estimating the effect of contextual variable (δ) varying the number of firms (n) and the correlation of x and z (ρ). Triple signal scenario

Scenario	Correlation (ρ)	Estimation method	
		Two-stage DEA	StoNEZD
$n = 50$	−0.8	71.5	13.8
	−0.4	28.9	9.60
	0.0	12.6	8.96
	0.4	15.7	9.73
	0.8	44.6	15.0
$n = 100$	−0.8	66.7	9.73
	−0.4	23.3	6.81
	0.0	7.66	6.81
	0.4	15.6	6.95
	0.8	46.9	9.28
$n = 200$	−0.8	67.9	5.78
	−0.4	20.6	3.77
	0.0	4.28	3.46
	0.4	11.1	3.77
	0.8	39.1	5.95

Case: $\delta = -0.6$, $\sigma_u = 0.15$, $\sigma_v = 0.04$, $M = 100$

a heavy noise scenario, reported in Table 4 where σ_v is increased to value $\sigma_v = 0.09$, keeping the other parameter values at their baseline levels. This scenario is expected to favor the StoNEZD estimator, which explicitly attempts to estimate and account for the noise. Indeed, the StoNEZD

Table 4 Performance (by RMSD) of alternative methods in estimating the effect of contextual variable (δ) varying the number of firms (n) and the correlation of x and z (ρ). Heavy noise scenario

Scenario	Correlation (ρ)	Estimation method	
		Two-stage DEA	StoNEZD
$n = 100$	−0.8	86.0	34.3
	−0.4	38.0	23.3
	0.0	25.1	24.7
	0.4	28.5	26.3
	0.8	50.6	35.8
$n = 200$	−0.8	70.9	17.7
	−0.4	27.2	11.8
	0.0	13.8	12.5
	0.4	17.6	12.0
	0.8	46.2	18.3

Case: $\delta = -0.2$, $\sigma_u = 0.15$, $\sigma_v = 0.09$, $M = 100$

Table 5 Performance (by RMSD) of alternative methods in estimating the inefficiency term (u) and the effect of contextual variable (δ) under different correlation of x and z (ρ). No noise scenario

Scenario	Correlation (ρ)	Estimation method	
		Two-stage DEA	StoNEZD
$n = 100$	−0.8	68.5	26.2
	−0.4	33.9	18.5
	0.0	17.2	16.6
	0.4	26.1	17.5
	0.8	61.4	26.8

Case: $\delta = -0.2$, $\sigma_u = 0.15$, $\sigma_v = 0.00$, $M = 100$

estimator has the lowest RMSD at all values of ρ, proving robust to the correlation of the input and the contextual variable. The two-stage DEA method performs well when ρ is equal to zero or takes a small positive value, but it deteriorates dramatically when ρ is negative. As expected, both methods suffer from the increased noise in the estimation of the parameter δ, but StoNEZD methods prove superior at all levels of ρ in this scenario.

Finally, consider a deterministic setting where the standard deviation of the noise term is set to zero (i.e., $\sigma_v = 0$), keeping other parameter values the same as in the base scenario of Table 1. The results of the no noise scenario are reported in Table 5. Both methods improve their performance compared to the base scenario. However, even though the DGP does not involve any noise, the StoNEZD estimator that assumes a positive noise term yields more precise inefficiency estimates than the two-stage estimator that explicitly assumes away the possibility of noise.

In conclusion, our results confirm the correlation between the input and the contextual variable increases the

small sample bias of DEA: this explains why performance of two-stage DEA deteriorates when z and x are correlated as well as the asymmetry in the performance under positive versus negative correlation. The simultaneous estimation method developed in this study—StoNEZD—proves more robust to the correlation of x and z. For both methods, the correlation of x and z causes the estimates for the effect of the contextual variable to deteriorate. In terms of the overall performance under wide variety of conditions, the StoNEZD method appears as a more dependable and robust alternative.

5 Conclusions

Estimating the effects of contextual variables is important because the results can provide valuable information regarding the relationship between characteristics of operational conditions and practices to output levels or performance. This information can help to identify ways to improve performance, taking production analysis from a method to develop static rankings of performance to a methodology that can prescribe performance improvement strategies.

Previously analysts had to choose between the fully parametric one-stage SFA or the two-stage DEA that requires other restrictive assumptions. Wang and Schmidt (2002) described the omitted variable bias in the classic two-stage approach to analyzing the context of production within parametric models. However, until this point a parallel line of development seemed impossible within the axiomatic DEA framework. The least-squares interpretation of DEA developed in Kuosmanen and Johnson (2010) provides new optimism. We can now approach DEA as a special case of nonparametric regression subject to axiomatic shape constraints and a sign-constraint for the residuals (i.e., inefficiency).

This paper introduces a new one-stage estimation strategy that jointly estimates an axiomatic production function and a linear regression model of characterizing the context of production. The new semi-nonparametric regression based estimator, referred to as the StoNEZD, produces estimates of the influence of the context that are unbiased, asymptotically efficient and normally distributed, and converge at the standard parametric rate despite the nonparametric frontier function. This justifies the use of conventional methods for inference such as t tests and confidence intervals.

To complement the asymptotic results, Monte Carlo simulations were conducted to investigate the finite sample performance of the developed method. The evidence from these simulations shows that the proposed approach outperforms conventional two-stage DEA in almost all scenarios. The advantage of the proposed method is particularly evident when the contextual variables are correlated with inputs.

We believe our method can provide insight into the relationship between estimating efficiency and modeling the context of production. Modeling these two components will assist managers who develop business strategies or determine operational practices, and policy-makers charged with developing regulations. Our approach is applicable to a variety of industries and levels of analysis.

Appendix: Proofs of theorems

Theorem 1 *If the data-generating process satisfies the maintained assumptions stated in Sect. 2, the* StoNEZD-*estimator for the coefficients of the contextual variables* $(\hat{\boldsymbol{\delta}}^S)$ *is statistically unbiased:*

$$E(\hat{\boldsymbol{\delta}}^S) = \boldsymbol{\delta}.$$

Proof Denote by $\phi(\mathbf{X}) = (\phi(\mathbf{x}_1)\ldots\phi(\mathbf{x}_n))'$ the vector of true frontier outputs and by $\hat{\phi}^S(\mathbf{X}) = (\hat{\phi}^S_1\ldots\hat{\phi}^S_n)'$ the vector of the estimated outputs according to the nonparametric part of the StoNEZD model. Consider the sub-problem of (3) that determines the optimal coefficients $\hat{\boldsymbol{\delta}}^S$, taking $\hat{\phi}^S(\mathbf{X})$ as given. The linear regression equation governing the z variables can be stated as

$$\ln \mathbf{y} - \ln \hat{\phi}^S(\mathbf{X}) = \mathbf{Z}\boldsymbol{\delta} + \boldsymbol{\varepsilon}.$$

The optimal solution to the sub-problem must satisfy the well-known first-order conditions of the standard OLS estimator (e.g., Greene 2007, Chap. 3), which imply the following closed form solution:

$$\hat{\boldsymbol{\delta}}^S = [(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'](\ln \mathbf{y} - \ln \hat{\phi}^S(\mathbf{X})).$$

The assumptions that the inefficiency term u and the noise v are statistically indepdent of the contextual variables implies that the composite error term $\varepsilon = \mathbf{v} - \mathbf{u}$ satisfies

$$E(\mathbf{Z}'\boldsymbol{\varepsilon}) = E[\mathbf{Z}'(\ln \mathbf{y} - \ln \phi(\mathbf{X}) - \mathbf{Z}\boldsymbol{\delta})] = \mathbf{0}.$$

Reorganizing this equality yields

$$E[\mathbf{Z}'(\ln \mathbf{y} - \ln \phi(\mathbf{X}))] = E[\mathbf{Z}'\mathbf{Z}\boldsymbol{\delta}].$$

Rewriting the closed form solution to $\hat{\boldsymbol{\delta}}^S$ as

$$\hat{\boldsymbol{\delta}}^S = [(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'][(\ln \mathbf{y} - \ln \phi(\mathbf{X})) + (\ln \phi(\mathbf{X}) - \ln \hat{\phi}^S(\mathbf{X}))],$$

we find that

$$
\begin{aligned}
E(\hat{\boldsymbol{\delta}}^S) &= E([(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'][(\ln \mathbf{y} - \ln \phi(\mathbf{X})) \\
&\quad + (\ln \phi(\mathbf{X}) - \ln \hat{\phi}^S(\mathbf{X}))]) \\
&= E([(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'][(\ln \mathbf{y} - \ln \phi(\mathbf{X}))) \\
&\quad + E([(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'](\ln \phi(\mathbf{X}) - \ln \hat{\phi}^S(\mathbf{X}))) \\
&= E((\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{Z}\boldsymbol{\delta}) + [(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}']E(\ln \phi(\mathbf{X}) \\
&\quad - \ln \hat{\phi}^S(\mathbf{X})) \\
&= \boldsymbol{\delta} - [(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}']E(\ln \hat{\phi}^S(\mathbf{X}) - \ln \phi(\mathbf{X})),
\end{aligned}
$$

For any given vector $\hat{\phi}^S(\mathbf{X})$, the potential bias of the estimator $\hat{\boldsymbol{\delta}}^S$ is due to the correlation of the z variables and the bias of the nonparametric frontier $E(\ln \hat{\phi}^S(\mathbf{X}) - \ln \phi(\mathbf{X}))$. Note that the unbiasedness of $\hat{\boldsymbol{\delta}}^S$ does not require that the estimator of the nonparametric frontier is unbiased, it suffices that the bias of $\hat{\phi}^S(\mathbf{X})$ does not correlate with the contextual variables.

Consider next the bias of the nonparametric part. The objective function of problem (3) can be stated using matrix algebra as

$$[(\ln \mathbf{y} - \mathbf{Z}\boldsymbol{\delta}) - \ln \hat{\phi}^S(\mathbf{X})]'[(\ln \mathbf{y} - \mathbf{Z}\boldsymbol{\delta}) - \ln \hat{\phi}^S(\mathbf{X})].$$

Minimization of this least-squares criterion requires that the conditional expectation satisfies:

$$E[\ln \hat{\phi}^S(\mathbf{X})|\mathbf{X}, \mathbf{Z}] = E[\ln \mathbf{y} - \mathbf{Z}\boldsymbol{\delta}|\mathbf{X}, \mathbf{Z}].$$

Substituting $\mathbf{y}$ by the right hand side of Eq. 1, we find that

$$
\begin{aligned}
E[\ln \mathbf{y} - \mathbf{Z}\boldsymbol{\delta}|\mathbf{X}, \mathbf{Z}] &= E[\ln \phi(\mathbf{X}) + \mathbf{Z}\boldsymbol{\delta} + \mathbf{v} - \mathbf{u}|\mathbf{X}, \mathbf{Z}] - \mathbf{Z}\boldsymbol{\delta} \\
&= \ln \phi(\mathbf{X}) - E(\mathbf{u}) = \ln \phi(\mathbf{X}) - \mu\mathbf{1}
\end{aligned}
$$

Therefore, the bias of $\ln \hat{\phi}^S(\mathbf{X})$ reduces to

$$E(\ln \hat{\phi}^S(\mathbf{X}) - \ln \phi(\mathbf{X})) = -\mu\mathbf{1}.$$

Since this is a vector of constants, we have $-(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'E(\mathbf{u}) = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mu\mathbf{1} = \mathbf{0}$. Therefore, the expected bias is zero and $E(\hat{\boldsymbol{\delta}}^S) = \boldsymbol{\delta}$. □

Note: if the assumption of homoskedasticity is relaxed, the nonparametric part $\ln \hat{\phi}^S(\mathbf{X})$ can be biased (i.e., the bias is not constant). This does not affect the unbiasedness of $\hat{\boldsymbol{\delta}}^S$ provided that the diagonal vector of the covariance matrix $\mathrm{Var}(\varepsilon)$ is uncorrelated with the columns of matrix $\mathbf{Z}$. For example, heteroskedasticity due to the firm size does not necessarily cause bias for the estimator $\hat{\boldsymbol{\delta}}^S$. However, if the contextual variables are the source of heteroskedasticity, or if heteroskedasticity is due to inputs $\mathbf{x}$ while $\mathbf{z}$ and $\mathbf{x}$ are correlated, then the bias of the nonparametric part correlates with the contextual variables and hence $\hat{\boldsymbol{\delta}}^S$ is biased. In that case, the bias is equal to.

$$[(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}']E(\ln \hat{\phi}^S(\mathbf{X}) - \ln \phi(\mathbf{X})).$$

Theorem 2 *If the following conditions are satisfied*

(i) *sequence,* $\{(y_i, \mathbf{x}_i, \mathbf{z}_i), \quad i = 1, \ldots, n\}$ *is a random sample of independent observations,*

(ii) $\lim_{n\to\infty} \mathbf{Z}'\mathbf{Z}/n$ *is a positive definite matrix,*

(iii) *the inefficiency terms* $\mathbf{u}$ *and the noise terms* $\mathbf{v}$ *are identically and independently distributed (i.i.d.) random variables with* $\mathrm{Var}(\mathbf{u}) = \sigma_u^2\mathbf{I}$ *and* $\mathrm{Var}(\mathbf{v}) = \sigma_v^2\mathbf{I}$,

then the StoNEZD-estimator for the coefficients of the contextual variables $(\hat{\boldsymbol{\delta}}^S)$ *is statistically consistent and asymptotically normally distributed according to:*

$$\hat{\boldsymbol{\delta}}^S \sim_a N(\boldsymbol{\delta}, (\sigma_v^2 + \sigma_u^2)(\mathbf{Z}'\mathbf{Z})^{-1}).$$

Proof The expected value follows directly from Theorem 1 (unbiasedness). Consider next the covariance matrix of the $\hat{\boldsymbol{\delta}}^S$:

$$\mathrm{Var}(\hat{\boldsymbol{\delta}}^S) = E\Big((\hat{\boldsymbol{\delta}}^S - \boldsymbol{\delta})(\hat{\boldsymbol{\delta}}^S - \boldsymbol{\delta})'\Big).$$

Noting that $\hat{\boldsymbol{\delta}}^S = \boldsymbol{\delta} + (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'(\ln \phi(\mathbf{X}) - \mathbf{u} + \mathbf{v} - \ln \hat{\phi}^S(\mathbf{X}))$, the covariance matrix can be stated as

$$
\begin{aligned}
\mathrm{Var}(\hat{\boldsymbol{\delta}}^S) = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'E((&\ln \phi(\mathbf{X}) - \mathbf{u} + \mathbf{v} \\
&- \ln \hat{\phi}^S(\mathbf{X}))(\ln \phi(\mathbf{X}) - \mathbf{u} + \mathbf{v} \\
&- \ln \hat{\phi}^S(\mathbf{X}))')\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}.
\end{aligned}
$$

Since the StoNEZD-estimator is consistent, $E(\ln \phi(\mathbf{X}) - \mathbf{u} - \ln \hat{\phi}^S(\mathbf{X})) \underset{a}{\to} \mathbf{0}$. Further, consistency of the StoNEZD-estimator implies $E((\ln \phi(\mathbf{X}) - \mathbf{u} - \ln \hat{\phi}^S(\mathbf{X}))(\ln \phi(\mathbf{X}) - \mathbf{u} - \ln \hat{\phi}^S(\mathbf{X}))') \underset{a}{\to} \sigma_u^2\mathbf{1}$. Thus,

$$
\begin{aligned}
\mathrm{Var}(\hat{\boldsymbol{\delta}}^S) &\underset{a}{\to} (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'(\sigma_u^2 + \sigma_v^2)\mathbf{1}\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1} \\
&= (\sigma_u^2 + \sigma_v^2)(\mathbf{Z}'\mathbf{Z})^{-1}
\end{aligned}
$$

Finally, consider the following sequence of the random variables $((v_1 - u_1)^2, \ldots, (v_n - u_n)^2)$, that are identically and independently distributed (by assumption), each with a finite expectation and variance, with the covariances $E((v_i - u_i)(v_h - u_h)) = 0\ \forall i \neq h$. Further, the sum $\sigma_u^2 + \sigma_v^2 = \sum_{i=1}^n (v_i - u_i)^2/n$ can be interpreted as a sample average of this sequence of i.i.d. random variables. Thus, the conditions of the central limit theorem are satisfied. This theorem implies that the limiting distribution of $\hat{\boldsymbol{\delta}}^S$ is normal even when v_i or u_i themselves are not normally distributed. □

Theorem 3 *If the conditions stated in Theorem 2 are satisfied and the skewness of the inefficiency terms* u_i *is*

finite such that $E(|u_i - \mu|^3) = \gamma < \infty$*, then the StoNEZD-estimator for the coefficients of the contextual variables* $(\hat{\boldsymbol{\delta}}^S)$ *converges to the true* $\boldsymbol{\delta}$ *at the standard parametric rate on the order of. Specifically, there exist a positive constant C such that for all n,*

$$|\hat{\boldsymbol{\delta}}^S - \boldsymbol{\delta}| \leq \frac{C\gamma}{\sqrt{n}(\sigma_v + \sigma_u)^3}\mathbf{1}.$$

Proof Recall that $\hat{\boldsymbol{\delta}}^S = \boldsymbol{\delta} + (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'(\ln \phi(\mathbf{X}) - \mathbf{u} + \mathbf{v} - \ln \hat{\phi}^S(\mathbf{X}))$, and thus,

$$\hat{\boldsymbol{\delta}}^S - \boldsymbol{\delta} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'(\ln \phi(\mathbf{X}) - \mathbf{u} + \mathbf{v} - \ln \hat{\phi}^S(\mathbf{X})).$$

We can express each $\hat{\delta}_k^S - \delta_k$ as a weighted average of random variables $d_i = \ln \phi(\mathbf{x}_i) - u_i + v_i - \ln \hat{\phi}_i^S$, $i = 1, \ldots, n$. Under the maintained assumptions, $E(d_i) = 0$ and $\mathrm{Var}(d_i) = \sigma_v^2 + \sigma_u^2$ for all $i = 1,\ldots,n$. Further, since are v_i symmetric and the StoNEZD-estimator is unbiased, the only source of skewness is the inefficiency term $\mathbf{u}$. The assumption $E(|u_i - \mu|^3) = \gamma < \infty$ ensures that $E(|d_i|^3) = C\gamma < \infty$, where C is a finite positive constant. Thus, all conditions of the Berry–Esseen theorem are satisfied. The inequality $\left|\hat{\boldsymbol{\delta}}^S - \boldsymbol{\delta}\right| \leq \frac{C\gamma}{\sqrt{n}(\sigma_v+\sigma_u)^3}\mathbf{1}$ follows from the application of the Berry-Esseen theorem. □

Theorem 4 *If the conditions stated in Theorem 2 are satisfied, then the StoNEZD-estimator for the coefficients of the contextual variables* $(\hat{\boldsymbol{\delta}}^S)$ *is asymptotically efficient. That is,*

$$\mathrm{AsyVar}(\hat{\boldsymbol{\delta}}^S) \leq \mathrm{AsyVar}(\hat{\boldsymbol{\delta}})$$

for any other consistent, asymptotically normally distributed estimator $\hat{\boldsymbol{\delta}}$.

Proof Firstly, recall from Theorem 3 that the asymptotic covariance matrix of the StoNEZD-estimator is

$$\mathrm{AsyVar}(\hat{\boldsymbol{\delta}}^S) = \left(\sigma_v^2 + \sigma_u^2\right)(\mathbf{Z}'\mathbf{Z})^{-1}.$$

Further, the consistency of the StoNEZD-estimator estimator implies that

$$\mathrm{Var}\left[\ln \mathbf{y} - \ln \hat{\phi}^S(\mathbf{X})\right] \underset{a}{\to} \left(\sigma_v^2 + \sigma_u^2\right)\mathbf{I}.$$

By construction, the StoNEZD-estimator estimator minimizes the sample variance within the class of monotonic increasing and concave functions ϕ. Otherwise, $\ln \hat{\phi}^S(\mathbf{X})$ is not the optimal solution to the QP problem (3). Thus, $\sigma_v^2 + \sigma_u^2$ is the smallest possible asymptotic variance for the estimator of the nonparametric function ϕ.

Second, consider an arbitrary linear estimator $\hat{\boldsymbol{\delta}} = [(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}' + \mathbf{D}](\ln \mathbf{y} - \ln \hat{\phi}(\mathbf{X}))$, where $\mathbf{D}$ is some arbitrary $k \times n$ non-zero matrix, $k < n$ that represents the deviations from the StoNEZD-based estimator. Any linear estimator of $\boldsymbol{\delta}$ can be expressed in this form. Denote by $\hat{\phi}(\mathbf{X})$ the vector of the estimated frontier outputs that satisfy monotonicity and concavity of the production function.

The expected value of $\hat{\boldsymbol{\delta}}$ is

$$\begin{aligned} E(\hat{\boldsymbol{\delta}}) &= E[[(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}' + \mathbf{D}](\ln \mathbf{y} - \ln \hat{\phi}(\mathbf{X}))] \\ &= E[[(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}' + \mathbf{D}][\ln \phi(\mathbf{X}) - \mathbf{Z}\boldsymbol{\delta} - \mathbf{u} \\ &\quad + \mathbf{v} - (\ln \hat{\phi}(\mathbf{X}) - \hat{\mathbf{u}})]] \end{aligned}$$

Any unbiased estimators for the frontier ϕ will satisfy $E\left[\ln \phi(\mathbf{X}) - \ln \hat{\phi}(\mathbf{X})\right] = \mathbf{0}$, and unbiased estimators of the inefficiency terms $\mathbf{u}$ will satisfy $E[\mathbf{u} - \hat{\mathbf{u}}] = \mathbf{0}$. Note that the $\boldsymbol{\delta}$ parameters can also be estimated without explicit estimation of the inefficiency term (as we do in the case of the StoNEZD-estimator). In the case both the function and the inefficiency are estimated unbiasedly, the unbiased estimator of the average function $\chi(\mathbf{x}) \equiv \phi(\mathbf{x}) - \mu$ must satisfy

$$\begin{aligned} E[\ln \chi(\mathbf{X})] - \ln \hat{\chi}(\mathbf{X})] &= E[(\ln \phi(\mathbf{X}) - \mathbf{u}) - (\ln \hat{\phi}(\mathbf{X}) - \hat{\mathbf{u}})] \\ &= 0. \end{aligned}$$

Distinction of the inefficiency term $\mathbf{u}$ does not influence unbiasness of the estimator for parameters $\boldsymbol{\delta}$. Finally, recall that $E[\mathbf{v}] = \mathbf{0}$. Combining the previous results, we find that

$$E(\hat{\boldsymbol{\delta}}) = \boldsymbol{\delta} + \boldsymbol{\delta}'\mathbf{D}\mathbf{Z}.$$

Therefore, any unbiased estimator $\hat{\boldsymbol{\delta}}$ must satisfy $\mathbf{D}\mathbf{Z} = \mathbf{0}$.

Next, denoting $\hat{\sigma}^2 = \mathrm{Var}(\ln y_i - \ln \hat{\phi}(\mathbf{x}_i))$, we have

$$\begin{aligned} \mathrm{Var}(\hat{\boldsymbol{\delta}}) &= \mathrm{Var}[[(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}' + \mathbf{D}](\ln \mathbf{y} - \ln \hat{\phi}(\mathbf{X}))] \\ &= \hat{\sigma}^2[(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}' + \mathbf{D}][(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}' + \mathbf{D}]' \\ &= \hat{\sigma}^2[(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1} + \mathbf{D}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}' \\ &\quad + (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{D} + \mathbf{D}\mathbf{D}'] \end{aligned}$$

As $\mathbf{Z}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1} = \mathbf{I}$ and $\mathbf{D}\mathbf{Z} = \mathbf{0}$, we have

$$\mathrm{Var}(\hat{\boldsymbol{\delta}}) = \hat{\sigma}^2[(\mathbf{Z}'\mathbf{Z})^{-1} + \mathbf{D}\mathbf{D}']$$

We have already noted that $s^2 \leq \hat{\sigma}^2$ for any consistent estimator. Second, matrix $\mathbf{D}\mathbf{D}'$ is always nonnegative definite for any non-zero $\mathbf{D}$ (e.g., Greene 2007, pp. 977–978). Therefore, $\mathrm{AsyVar}(\hat{\boldsymbol{\delta}}^S) \leq \mathrm{AsyVar}(\hat{\boldsymbol{\delta}})$ for any other consistent estimator $\hat{\boldsymbol{\delta}}$. □

References

Afriat SN (1967) The construction of a utility function from expenditure data. Int Econ Rev 8:67–77

Afriat SN (1972) Efficiency estimation of production functions. Int Econ Rev 13(3):568–598

Aigner D, Lovell CAK, Schmidt P (1977) Formulation and estimation of stochastic frontier production function models. J Econ 6:21–37

Banker RD, Natarajan R (2008) Evaluating contextual variables affecting productivity using data envelopment analysis. Oper Res 56(1):48–58

Chen X (2007) Large sample sieve estimation of semi-nonparametric models. In: Heckman JJ, Lleamer EE (eds) Chapter 76 in Handbook of econometrics, vol 6B, North-Holland

Estelle SM, Johnson AL, Ruggiero J (2010) Three-stage DEA models for incorporating exogenous inputs. Comput Oper Res 37(6):1087–1090

Greene WH (2007) Econometric analysis, 6th edn edn. Prentice Hall, Englewood Cliffs

Groeneboom P, Jongbloed G, Wellner JA (2001) A canonical process for estimation of convex functions: the "invelope" of integrated Brownian motion plus t(4). Ann Stat 29(6):1620–1652

Gstach D (1998) Another approach to data envelopment analysis in noisy environments: DEA+. J Prod Anal 9(2):161–176

Hall M, Winsten C (1959) The ambiguous notion of efficiency. Econ J 69(273):71–86

Hanson DL, Pledger G (1976) Consistency in concave regression. Ann Stat 4(6):1038–1050

Hildreth C (1954) Pont estimates of ordinates of concave functions. J Am Stat Assoc 49(267):598–619

Johnson AL, Kuosmanen T (2010) On one-stage and two-stage DEA estimation of the effects of contextual variables. Working paper

Kuosmanen T (2006) Stochastic nonparametric envelopment of data: combining virtues of SFA and DEA in a unified framework. MTT discussion papers, Helsinki

Kuosmanen T (2008) Representation theorem for convex nonparametric least squares. Econ J 11:308–325

Kuosmanen T, Johnson AL (2010) Data envelopment analysis as nonparametric least squares regression. Oper Res 58(1):149–160

Kuosmanen T, Johnson AL (2011) Stochastic axiomatic estimation of joint production: does competition affect the performance of the railroad firm? Working paper

Kuosmanen T, Kortelainen M (2011) Stochastic non-smooth envelopment of data: semi-parametric frontier estimation subject to shape constraints. J Product Anal (in press)

Kuosmanen T, Cherchye L, Sipilainen T (2006) The law of one price in data envelopment analysis: restricting weight flexibility across firms. Eur J Oper Res 170(3):735–757

Li S-K, Ng YC (1995) Measuring the productive efficiency of a group of firms. Int Adv Econ Res 1(4):377–390

MacKinnon JG, White H (1985) Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties. J Econ 29(3):305–325

Meeusen W, van den Broek J (1977) Efficiency estimation from Cobb–Douglas production function with composed error. Int Econ Rev 8:435–444

Ray SC (1988) Data envelopment analysis, nondiscretionary inputs and efficiency: an alternative interpretation. Socioecon Plann Sci 22(4):167–176

Ray SC (1991) Resource-use efficiency in public schools: a study of Connecticut data. Manage Sci 37(12):1620–1628

Simar L, Wilson PW (2007) Estimation and inference in two-stage, semi-parametric models of production processes. J Econ 136(1):31–64

Simar L, Wilson PW (2010) Two-stage DEA: caveat emptor. Technical report #10045, Institut de Statistique, Universite Catholique de Louvain, Louvain-la-Neuve, Belgium. http://www.stat.ucl.ac.be/ISpub/tr/2010/TR10045.pdf. Accessed 11 Dec 2010

Timmer CP (1971) Using a probabilistic frontier production function to measure technical efficiency. J Polit Econ 79:767–794

Wang HJ, Schmidt P (2002) One-step and two-step estimation of the effects of exogenous variables on technical efficiency levels. J Prod Anal 18(2):129–144

White H (1980) A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. Econometrica 48(4):817–838