J Prod Anal (2011) 36:205–218
DOI 10.1007/s11123-011-0230-6

Two-stage DEA: *caveat emptor*

Léopold Simar · Paul W. Wilson

Published online: 9 July 2011

Abstract This paper examines the wide-spread practice where data envelopment analysis (DEA) efficiency estimates are regressed on some environmental variables in a second-stage analysis. In the literature, only two statistical models have been proposed in which second-stage regressions are well-defined and meaningful. In the model considered by Simar and Wilson (J Prod Anal 13:49–78, 2007), truncated regression provides consistent estimation in the second stage, where as in the model proposed by Banker and Natarajan (Oper Res 56: 48–58, 2008a), ordinary least squares (OLS) provides consistent estimation. This paper examines, compares, and contrasts the very different assumptions underlying these two models, and makes clear that second-stage OLS estimation is consistent only under very peculiar and unusual assumptions on the data-generating process that limit its applicability. In addition, we show that in either case, bootstrap methods provide the only feasible means for inference in the second stage. We also comment on ad hoc specifications of second-stage regression equations that ignore the part of the data-generating process that yields data used to obtain the initial DEA estimates.

Keywords Technical efficiency · Two-stage estimation · Bootstrap · Data envelopment analysis (DEA)

L. Simar
Institut de Statistique, Biostatistique et Sciences Actuarielles, Université Catholique de Louvain, Voie du Roman Pays 20, 1348 Louvain-la-Neuve, Belgium
e-mail: leopold.simar@uclouvain.be

P. W. Wilson (✉)
The John E. Walker Department of Economics, Clemson University, 222 Sirrine Hall, Clemson, SC 29634-1309, USA
e-mail: pww@clemson.edu

JEL Classification C12 · C14 · C61 · D24

1 Introduction

Two-stage estimation procedures wherein technical efficiency is estimated by data envelopment analysis (DEA) or free disposal hull (FDH) estimators in the first stage, and the resulting efficiency estimates are regressed on some environmental variables in a second stage (hereafter referred to simply as "second-stage regressions"), remain popular in the literature. The Google Scholar search engine returned about 1,590 articles for the period 2007–2010 after a search on "efficiency," "two-stage," and "dea" for the period 2007–2010 on 16 August 2010. Replacing "dea" with "fdh" returned 194 hits. A large number of these papers use either ordinary least squares (OLS) or tobit regression in the second stage and rely on conventional methods for inference.

Simar and Wilson (2007, hereafter referred to as SW) considered a well-defined, coherent statistical model in which a second-stage regression is meaningful in the sense that the form of the second-stage regression equation is determined by the structure of the model in the first stage where the initial DEA estimates are obtained. In an attempt to rationalize studies where second-stage regressions have been estimated but no statistical model has been specified, SW introduced assumptions that lead to a truncated regression in the second stage which can be estimated consistently using the maximum likelihood (ML) method. As discussed below in Sect. 2, the assumption leading to a truncated regression in the second stage can be easily replaced to obtain a logistic or other parametric regression equation, or even a fully non-parametric regression equation. In any case, however, conventional inference methods fail to give valid

inference due to the fact that in the second-stage, *true* efficiency remains unobserved and must be replaced with DEA *estimates* of efficiency, and these are correlated by construction. SW showed how bootstrap methods can be used for inference in the case of a truncated regression, and these methods are easy to extend to cases where different assumptions, leading to different forms of the second-stage regression equation in their model, are made.

Banker and Natarajan (2008a, hereafter referred to as BN) proposed an alternative well-defined, coherent statistical model in which a second-stage regression is meaningful. In the BN model, the second-stage regression equation is log-linear, and OLS provides consistent estimation. BN did not mention in their paper how inference might be made in the second stage, but the on-line appendix (Banker and Natarajan 2008b; hereafter referred to as BN2) cited in their paper contains a proof of one of the propositions in their paper, and statements in the proof indicate that conventional OLS standard error estimates can be used for making inference in the usual way. However, as discussed below in Sect. 3, some statements in the proof are demonstrably false. Moreover, consistency of OLS in the second-stage regression depends crucially on the assumptions of the BN model. As also discussed below in Sect. 3, some of these assumptions are quite strong (i.e., restrictive), and should not be expected to hold in general. As demonstrated below, OLS is inconsistent if *any* of several restrictive assumptions in the BN model fail to hold.

Unfortunately, the BN paper makes a number of overreaching statements, leaving the impression that the usefulness of OLS in second-stage regressions is a general result, when in fact the result is specific to the BN model and its restrictive assumptions as discussed below in Sect. 3 Others have added to the confusion. For example, Sufian and Habibullah (2009, p. 341) write,

> In an influential development, Banker and Natarajan (2008a) provide proof that the use of a two-stage procedure involving DEA followed by an ordinary least square [sic] regression yields consistent estimators of the regression coefficients.

Cummins et al. (2010, p. 1526, third full paragraph) make similar statements. While BN leave this impression (e.g., see the quote from BN, p. 56, below in Sect. 3.1), the claim that OLS yields consistent estimation in the second stage is not true in general as discussed below in Sect. 3.2.

To our knowledge, SW and BN are the only papers to propose well-defined, coherent statistical models that lead to meaningful second-stage regressions in the sense defined above. Unfortunately, several topical papers, including Hoff (2007), McDonald (2009), and Ramalho et al. (2010) have recently argued that log-linear specifications (estimated by OLS), censored (i.e., tobit) specifications (estimated by ML), or other particular parametric specifications should be used in the second stage, but these papers do so without specifying a well-defined statistical model in which such structures would follow from the first stage where the initial DEA estimates are obtained. As such, these approaches are not structural, but instead are ad hoc; given the lack of a statistical model, it is unknown what might be estimated by such approaches. These problems are discussed in further detail in Sect. 4.

Unfortunately, BN, Hoff (2007), and McDonald (2009) have been cited by a number of empirical researchers as justification for using OLS in second-stage regressions. In particular, BN is often cited uncritically, without mentioning, considering, or testing the assumptions of the BN model when OLS estimation is used in second-stage regressions. Worse, studies that do this often provide OLS standard error estimates, which are inconsistent due to the correlation of DEA efficiency estimates and hence fail to be useful for valid inference. Examples include Chang et al. (2004), who cite a working paper version of BN and use OLS to estimate a *linear* second-stage regression, despite the fact that DEA efficiency estimates are bounded at unity. Examples also include Chang et al. (2008), Sufian and Habibullah (2009), Barkhi and Kao (2010), Cummins et al. (2010), Davutyan et al. (2010), Erhemjamts and Leverty (2010), and others.

In the following sections, we attempt to clear up some of the confusion that has developed. In the next section, we revisit SW in an attempt to state clearly, without too much technicality, what the main points of SW were, and to dispel some myths that have arisen. In Sect. 3, we critically examine the BN model by providing a detailed discussion of the assumptions and claims in the BN paper. Section 4 provides some brief comments on ad hoc specification of second-stage regression equations outside the context of a well-defined statistical model. The final section gives a summary, where we compare the assumptions required by the model considered by SW and those required by the BN model, letting the reader decide which might be less restrictive or more useful in typical empirical situations.

2 Simar and Wilson (2007) revisited

SW cited 48 published papers that regressed DEA efficiency estimates on some environmental variables in a second stage, and commented that "as far as we have been able to determine, none of the studies that employ this two-stage approach have described the underlying data-generating process." SW went on to (1) define a statistical model where truncated (but not censored, i.e., tobit, nor OLS) regression yields consistent estimation of model features; (2) demonstrated that conventional, likelihood-based

approaches to inference are invalid; (3) and developed a bootstrap approach that yields valid inference in the second-stage regression when such regressions are appropriate. It is important to note that SW did not advocate two-stage procedures; rather, the point of the paper was (1) to rationalize what has been done in the literature by providing a coherent, well-defined statistical model where a second-stage regression would be appropriate; and (2) to show how valid inference could be made in the second-stage regression. With regard to the first point, as far as we know the model provided by SW was the first complete description of a data-generating process (DGP) where second-stage regression would be appropriate. SW did not claim that this was the only such model; in fact, BN have introduced an alternative model as discussed below in Sect. 3.

The statistical model in SW is defined by Assumptions A1–A8 listed in their paper. These assumptions augment the standard non-parametric production model where DEA efficiency estimators are consistent (e.g., see Kneip et al. 1998; Simar and Wilson, or Kneip et al. 2008) to incorporate environmental variables. Specifically, the Farrell (1957) output efficiency measure δ_i is assumed to be a function $\psi(\boldsymbol{Z}_i, \boldsymbol{\beta})$ of environmental covariates $\boldsymbol{Z}_i$ and parameters $\boldsymbol{\beta}$ plus an independently distributed random variable ϵ_i representing the part of inefficiency not explained by $\boldsymbol{Z}_i$ (see SW, Assumptions A2). In addition, since $\delta_i \geq 1$ by definition, ϵ_i is assumed (in Assumption A3 of SW) to be distributed $N(0, \sigma_\epsilon^2)$ with left-truncation at $1 - \psi(\boldsymbol{Z}_i, \boldsymbol{\beta})$. Assumption A2 of SW implies

$$\delta_i = \psi(\boldsymbol{Z}_i, \boldsymbol{\beta}) + \epsilon_i \geq 1; \tag{1}$$

after rearranging terms, $\epsilon_i \geq 1 - \psi(\boldsymbol{Z}_i, \boldsymbol{\beta})$, which explains why ϵ_i must be truncated on the left at $1 - \psi(\boldsymbol{Z}_i, \boldsymbol{\beta})$.

SW note (pp. 35–36) that their Assumptions A1–A2 imply a "separability" condition, and that this condition may or may not be supported by the data, and hence that the condition should be tested. Here, we use the word "separability" as it was used in SW, and differently than it is sometimes used. Specifically, by "separability," we mean that the support of the output variables does not depend on the environmental variables in $\boldsymbol{Z}$. To illustrate this condition, consider the two DGPs given by

$$Y^* = g(X)e^{-(Z-2)^2 U} \tag{2}$$

and

$$Y^{**} = g(X)e^{-(Z-2)^2} e^{-U} \tag{3}$$

where $g(X) = \left(1 - (X-1)^2\right)^{1/2}, X \in [0,1], Z \in [0,4]$, and $U \geq 0$ is a one-sided inefficiency process. Setting $U = 0$ in (2), (3) gives the frontiers for the two DGPs, as illustrated in Fig. 1, where the frontier corresponding to (2) is shown in the left panel, and the frontier corresponding to (3) is shown in the right panel. To help visualize the frontiers, Fig. 2 shows contours of the two surfaces depicted in Fig. 1. Clearly, the frontiers are very different; it is clear that for a given level of the input variable X, the maximal output level Y^* in (2) does not vary with Z, as indicated by the vertical, linear contours in the left panel of Fig. 2. However, the maximal output level Y^{**} in (3) does vary with Z, and the corresponding contours in the right panel of Fig. 2 are non-linear. The "separability" condition discussed by SW is satisfied by the DGP in (2), but not by the DGP in (3).

To further illustrate the implications of the "separability" condition, consider the observations (0.5, 0.2, 1.75), (0.5, 0.2, 2.0), and (0.5, 0.2, 2.5) for (X, Y, Z). If the true DGP is as given by (2), the some algebra reveals that the *true* Farrell output efficiencies for each of the three observations are (approximately) 0.8660/0.2 = 4.33. On the other hand, if the true DGP is given by (3), then the Farrell output efficiencies corresponding to the three observations listed above are (approximately) 0.8136/0.2 = 4.068, 0.8660/0.2 = 4.33, and 0.6745/0.2 = 3.375 (respectively). As noted in the previous paragraph, the frontier corresponding to (2) is invariant with respect to Z, while the frontier corresponding to (3) is not. It is clear that whether the "separability" condition holds has an impact on the underlying, true efficiency levels, and this impact may be large. In the example considered here, the

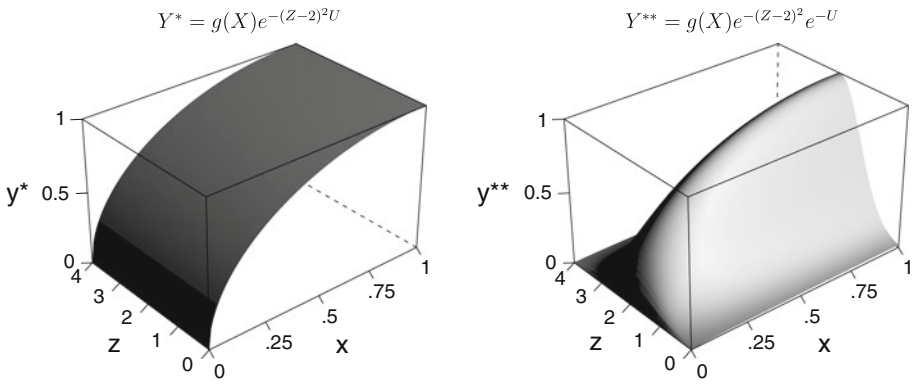

Fig. 1 Illustration of "Separability" condition described by Simar and Wilson (2007)

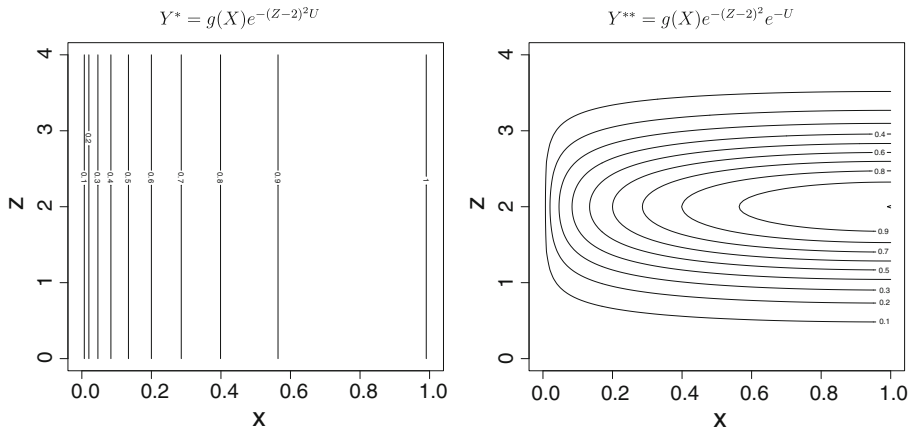

Fig. 2 Contours of frontiers corresponding to Eqs. (2), (3)

output efficiency level for the third observation is about 28.3% larger if the DGP is given by (2), where the "separability" condition is satisfied, as opposed to the case where the DGP is given by (3), where the "separability" condition is not satisfied.

Daraio et al. (2010) provide a fully non-parametric test of this condition. If it is rejected, then the conditional efficiency measures described by Daraio and Simar (2005, 2006) are appropriate. The non-parametric estimators of these measures described by Daraio and Simar (2005, 2006) use smoothing techniques, and statistical properties of the estimators have been established by Jeong et al. (2010). In addition, Bădin et al. (2010) provide a data-driven method for selecting bandwidths for use with these estimators.

To understand the importance of the "separability" condition, let $\boldsymbol{X} \in \mathbb{R}_+^p$ denote a vector of p input quantities, and let $\boldsymbol{Y} \in \mathbb{R}_+^q$ denote a vector of q output quantities. In addition, let $\boldsymbol{Z} \in \mathcal{Z} \subseteq \mathbb{R}^r$ denote a vector of r environmental variables with domain $\mathcal{Z}$. Let $\mathcal{S}_n = \{(\boldsymbol{X}_i, \boldsymbol{Y}_i, \boldsymbol{Z}_i)\}_{i=1}^n$ denote a set of observations. Assumptions A1–A2 in SW imply that the sample observations $(\boldsymbol{X}_i, \boldsymbol{Y}_i, \boldsymbol{Z}_i)$ in $\mathcal{S}_n$ are realizations of identically, independently distributed random variables $(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z})$ with probability density function $f(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z})$ which has support over a compact set $\mathcal{P} \subset \mathbb{R}_+^{p+q} \times \mathbb{R}^r$ with level sets $\mathcal{P}(\boldsymbol{z})$ defined by

$$\mathcal{P}(\boldsymbol{z}) = \{(\boldsymbol{X}, \boldsymbol{Y}) \mid \boldsymbol{Z} = \boldsymbol{z},\ \boldsymbol{X} \text{ can produce } \boldsymbol{Y}\}. \tag{4}$$

Now let

$$\Psi = \bigcup_{\boldsymbol{z} \in \mathcal{Z}} \mathcal{P}(\boldsymbol{z}) \subset \mathbb{R}_+^{p+q}. \tag{5}$$

Under the "separability" condition, $\mathcal{P}(\boldsymbol{z}) = \Psi\ \forall\ \boldsymbol{z} \in \mathcal{Z}$ and hence $\mathcal{P} = \Psi \times \mathcal{Z}$. If this condition is violated, then $\mathcal{P}(\boldsymbol{z}) \neq \Psi$ for some $\boldsymbol{z} \in \mathcal{Z}$; i.e., $\mathcal{P}(\boldsymbol{z}) \neq \mathcal{P}(\tilde{\boldsymbol{z}})$ for some $\boldsymbol{z} \neq \tilde{\boldsymbol{z}}, \boldsymbol{z}, \tilde{\boldsymbol{z}} \in \mathcal{Z}$. Whether this is the case or not is ultimately an empirical question; again, Daraio et al. (2010) provide a method for testing $H_0: \mathcal{P}(\boldsymbol{z}) = \Psi\ \forall\ \boldsymbol{z} \in \mathcal{Z}$ versus $H_1: \mathcal{P}(\boldsymbol{z}) \neq \Psi$ for some $\boldsymbol{z} \in \mathcal{Z}$. The null hypothesis constitutes a strong assumption, and we expect that in many samples, the null will be rejected. As an example, Daraio et al. (2010) revisit the empirical example based on Aly et al. (1990) that was presented in SW, and easily reject separability. The model introduced by BN does not impose separability, but as discussed below in Sect. 3, it imposes other restrictive conditions that are not likely to be satisfied by real data.

Returning to the illustration in Fig. 1, given a sample $\{(\boldsymbol{X}_i, \boldsymbol{Y}_i, \boldsymbol{Z}_i)\}_{i=1}^n$, what would it mean to estimate efficiency with DEA using the observations $\{(\boldsymbol{X}_i, \boldsymbol{Y}_i)\}_{i=1}^n$ if the underlying technology is the one in the right-hand panel? The preceding argument makes the answer clear: for a particular observation $(\boldsymbol{X}_i, \boldsymbol{Y}_i)$, DEA would estimate the distance not to the frontier $\mathcal{P}(\boldsymbol{Z}_i)$, but to the boundary of the set Ψ described in (5). In terms of the right-hand panel in Fig. 1, the frontier of the corresponding set Ψ is identical to the frontier shown in the left-hand panel of Fig. 1. Hence the DEA estimator, for a point $(\boldsymbol{X}_i, \boldsymbol{Y}_i)$, measures distance not to the technology, but to a frontier that is *very* different from the frontier shown in the right-hand panel of Fig. 1.

In terms of the specific example considered above, note that in both (2) and (3), output levels range from 0 to 1. Figure 3 shows, for $X = 0.5$, the frontiers corresponding to the two DGPs in (2), (3), with the left panel of Fig. 3 corresponding to (2) and the right panel corresponding to (3). The maximum output level shown in the right-hand panel of Fig. 3 is the same as the maximum output level for any value of Z in the left-hand panel of Fig. 3, since $g(X)$ is the same in (2), (3). If the separability condition is not satisfied, as in the right-hand panel of Fig. 3, measuring efficiency while ignoring this fact leads to meaningless results in the first stage of any two-stage estimation procedure. In terms of Fig. 3, if the true DGP is given by (3), the Farrell output efficiency measure projects the three hypothetical observations listed above onto a horizontal

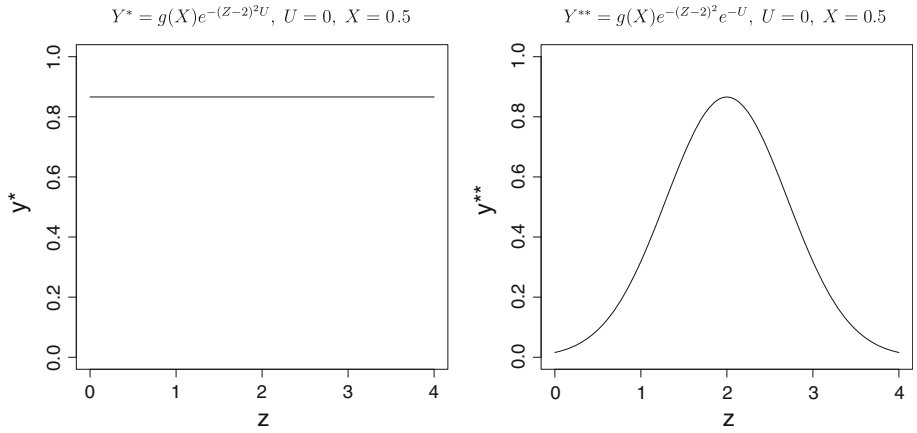

Fig. 3 Slices of production sets corresponding to Eqs. (2–3) with $X = 0.5$

line tangent to the frontier in the right-hand panel of Fig. 3 instead of projecting the observations onto the actual frontier.

In situations where the "separability" condition is satisfied, *if* the δ_i were observed, it would be straightforward to estimate (1). One might assume $\psi(\boldsymbol{Z}, \boldsymbol{\beta}) = \boldsymbol{Z\beta}$ and estimate the model by the ML method using standard software; note, however, that in the model given by (2), the relation between Farrell output efficiency and Z is given by

$$\delta = \frac{g(X)}{Y^*} = e^{-(Z-2)^2 U} \tag{6}$$

and hence is non-linear in Z. One could model this explicitly if the true DGP in (2) were known, or alternatively, one could allow $\psi(\boldsymbol{Z}, \boldsymbol{\beta})$ and the distribution of ϵ to be nonparametric and estimate $\psi(\cdot)$ using the local likelihood method discussed by Park et al. (2008).

Unfortunately, however, the δ_i are not observed. SW present two approaches for dealing with this problem. In the first approach, DEA estimates $\widehat{\delta}_i$ from the first stage estimation are used to replace the unobserved δ_i in (1) with $\psi(\boldsymbol{Z}_i, \boldsymbol{\beta}) = \boldsymbol{Z}_i\boldsymbol{\beta}$. Since the DEA estimates are consistent under the assumptions of the model in SW, ML estimation of the truncated regression

$$\widehat{\delta}_i = \boldsymbol{Z}_i\boldsymbol{\beta} + \xi_i \geq 1 \tag{7}$$

appearing in equation (13) of SW yields consistent estimates of $\boldsymbol{\beta}$. However, as SW note, inference is problematic due to the fact that $\widehat{\delta}_i$ has replaced the unobserved δ_i, and while the $\widehat{\delta}_i$ consistently estimate the δ_i, the DEA estimators converge slowly, at rate $n^{-2/(p+q+1)}$, and are biased. Consequently, the inverse of the negative Hessian of the log-likelihood corresponding to (7) does not consistently estimate the variance of the ML estimator of $\boldsymbol{\beta}$. The bootstrap procedure given in Algorithm #1 in SW provides a bootstrap procedure for making inference about $\boldsymbol{\beta}$ when (7) is estimated by ML.

SW also show how bootstrap methods can be used to construct bias-corrected estimates $\widehat{\widehat{\delta}}_i$ of the unobserved δ_i. Replacing the δ_i in (7) with the bias-corrected estimator $\widehat{\widehat{\delta}}_i$ and setting $\psi(\boldsymbol{Z}_i, \boldsymbol{\beta}) = \boldsymbol{Z}_i\boldsymbol{\beta}$ yields another truncated regression model in which ML estimation produces consistent estimates of $\boldsymbol{\beta}$. However, the issues for inference remain as before. SW provide a second bootstrap procedure based on bias-corrected estimates of δ_i in their Algorithm #2.

In either of the algorithms given by SW, it would be straightforward to allow $\psi(\boldsymbol{Z}_i, \boldsymbol{\beta})$ as well as the distribution of ϵ to be nonparametric, and to replace ML estimation of the truncated regression in Algorithms #1 and #2 with nonparametric estimation methods as mentioned above. The assumption of linearity of $\psi(\boldsymbol{Z}_i, \boldsymbol{\beta})$ in SW was made to correspond to what is typically done in the literature, and to what was done in the numerous articles cited in the introduction of SW. One could also assume different parametric forms, such as a logistic regression.

It is important to note that SW did not advocate using two-stage methods. As noted earlier, the goal was to provide a well-defined statistical model that could rationalize what has been done in the literature. In the end, the model in SW requires truncated regression in the second stage. Within the assumptions of the model in SW, tobit regression constitutes a mis-specification. The simulation results presented by SW confirm that under the assumptions of the model in SW, tobit estimation in the second stage yields biased and inconsistent estimates.

As far as we are aware, no statistical model in which second-stage tobit regression of DEA efficiency estimates on some environmental variables would produce consistent estimates has been presented in the literature. Similarly, BN (Sect. 4.3) also remark, "we cannot theoretically justify the use of a tobit regression in the second stage in terms of an underlying DGP…" A number of papers (e.g., Hoff 2007) argue that tobit regression is appropriate since in a

given sample, a number of DEA estimates will equal unity. This is by construction. However, under standard assumptions where properties of DEA estimators have been derived (e.g., Kneip et al. 2008), it is clear that the mass of estimates equal to one are due to the bias of the DEA frontier estimator. In other words, the estimates equal to one are spurious. If one were able to observe a sample of *true* efficiencies, one would not see a group of values equal to one.[1]

In the next section, we examine the alternative model proposed by BN.

3 The model of Banker and Natarajan (2008)

3.1 Model structure

BN present a model containing a one-sided inefficiency process, a two-sided, but bounded, noise process, and with "contextual variables affecting productivity," referring to environmental variables as contextual variables. In their abstract, BN state that

> Conditions are identified under which a two-stage procedure consisting of DEA followed by ordinary least squares (OLS) regression analysis yields consistent estimators of the impact of contextual variables. Conditions are also identified under which DEA in the first stage followed by ML estimation (MLE) in the second stage yields consistent estimators of the impact of contextual variables. This requires the contextual variables to be independent of the input variables.

As will be demonstrated below, these claims are true, but only under a set of assumptions that are rather restrictive. Unfortunately for the inattentive reader, BN give the impression at various points in their paper that their claims hold in general; e.g., in discussing their contributions in Sect. 5 of their paper, on p. 56, they state:

> Specifically, we prove that when data are generated by a monotone increasing and concave production function separable from a parametric function of the contextual variables, a two-stage approach comprising a DEA model followed by an ordinary least squares (or ML estimation) model yields consistent estimators of the impact of the contextual variables.

This is not true in general. In fact, as will be shown below, considerably more is assumed than what is revealed in this statement. The claims are specific to the BN model, and hold only under a number of restrictive conditions as will be explained below.

BN present their DGP in Sect. 2 of their paper in terms of a univariate output (i.e., $q = 1$); the DGP can be represented by

$$Y = \phi(X)e^{-\mathbf{Z}\boldsymbol{\beta}+V-U} \tag{8}$$

where Y is an output quantity, X is an input quantity, $\mathbf{Z}$ is a vector of r environmental variables, $\boldsymbol{\beta}$ is a vector of r parameters, U is a one-sided inefficiency process, and V is a two-sided noise process. On p. 50 of their paper, BN list their assumptions, including (1) $X \geq 0$; (2) $U \geq 0$; (3) $\mathbf{Z} \geq 0$; (4) $\boldsymbol{\beta} \geq 0$; (5) $-V^M \leq V \leq V^M$, where $V^M \geq 0$ is a constant; (6) X, Z, U, and V are mutually independent; (7) each random variable has finite variance; and (8) $E(V) = 0$. In addition, though not stated explicitly, both U and V are assumed to be distributed *identically* across all observations; i.e., both U and V are assumed to have constant variance, constant mean, and the same distribution for all observations.

In terms of the notation used in Sect. 2, the production set defined by (4) corresponding to (8) is given by

$$\mathcal{P}(\mathbf{z}) = \left\{(X, Y) \mid \mathbf{Z} = \mathbf{z},\ X \geq 0,\ Y \leq \phi(X)e^{V^M-\mathbf{Z}\boldsymbol{\beta}}\right\}, \tag{9}$$

and the set Ψ defined by (5) is given by

$$\Psi = \left\{(X, Y) \mid X \geq 0,\ Y \leq \phi(X)e^{V^M}\right\}. \tag{10}$$

Hence the DGP in (8) does not satisfy the "separability" condition described by SW since the support of Y depends on $\mathbf{Z}$. Moreover, since $\mathbf{Z}$ is assumed to be independent of U, the environmental variables cannot be interpreted as affecting inefficiency; instead, they affect the shape of the frontier in the BN framework. In addition, if $V^M > 0$, then standard efficiency measures (e.g., the Shephard 1970 output distance function) cannot be interpreted as measures of inefficiency in the context of (8) since they confound information about inefficiency, determined by U, with the boundary on the noise process, V^M.

On the surface, the assumptions required by the BN model seem innocuous. In fact, some of them are very restrictive, and at least one is almost certain to be violated by most data empirical researchers are likely to encounter. First, the assumption that all of the coefficients on the environmental variables are non-negative means that the researcher must know a priori the direction of the effects of the environmental variables. In some cases, this might be reasonable, but in many cases it is not. For example, one might use for elements of $\mathbf{Z}$ variables describing various regulatory regimes faced by firms in an industry.

[1] One could perhaps assume that the joint density of input-output vectors includes a probability mass along the frontier, but given the bias of the DEA frontier estimator and the resulting mass of observations for which the corresponding DEA efficiency estimate will equal unity, it is difficult to imagine how such a model could be identified from the model in Kneip et al. (2008). In addition, the properties of DEA estimators in such a model are unknown.

Depending on the nature of the regulation, and the actions of the regulating authority, regulations faced by businesses might hinder production, or to the extent that they limit competition, they might stimulate production by firms that are allowed to operate. Presumably, one of the reasons for engaging in empirical research is to check what the data have to say about whether there is an effect from a variable, what its direction might be, and only finally, what the magnitude of the effect might be. Assuming a priori the direction of the effect of environmental variables will surely be problematic in some, perhaps many, applications.

Second, the assumption that X and Z are independent is not likely to hold in economic data. For example, in agricultural applications, one might use rainfall as an environmental variable, but farmers surely do not choose input levels independently of rainfall—farmers in Belgium do not irrigate their crops, but farmers in west Texas must do so. One might consider replacing the elements of $\boldsymbol{Z}$ with instrumental variables in order to satisfy independence with X, but this is problematic for several reasons. Instruments are often not available, and introduce measurement error. Furthermore, using instruments in nonlinear models is problematic, and there is no theory for what the implications might be for doing so in a frontier model. The implications are perhaps even more uncertain in the context of a non-parametric or semi-parametric model.

Third, the implicit assumption of homoskedasticity and homogeneity for V and U are not likely to hold. Larger firms are likely to have better access to good managers than small firms, and hence may be more efficient than small firms. Similarly, output is likely to be more variable for large firms than for small firms, implying the assumption of constant variance for the noise term V is dubious. In cross-sectional regressions involving production or cost functions, one typically finds heteroskedasticity, further calling into question the assumptions required by this model.[2]

Fourth, the assumption that noise is bounded is problematic. Apart from the question of why noise should be bounded, and what its economic meaning might be, one might ask *if it is bounded*, why should it be bounded symmetrically? Moreover, why should the bounds be the same for all firms? As noted above, output is likely more variable for large firms than for small firms, which are constrained by smaller capacity. Yet, the assumptions of constant bounds, constant variance, and identical distribution for V are essential for estimation of efficiency with the Gstach (1998) method used in the BN framework. Moreover, it should be noted that *none* of the 48 papers cited in the Introduction of SW assumed a noise term or used the Gstach method.

Fifth, BN assume the DGP is as given in (8). In particular, this implies several important restrictions in their model. Perhaps most important, the support of Y, i.e., the frontier, decreases in $\boldsymbol{Z}$ monotonically and at a partially parametric rate for a given input level since

$$\frac{\partial Y}{\partial \boldsymbol{Z}'} = -\boldsymbol{\beta}\phi(X)e^{-\boldsymbol{Z}\boldsymbol{\beta}+V-U} = -\boldsymbol{\beta}Y < 0. \tag{11}$$

Hence $\boldsymbol{Z}$ is assumed to have a specific, monotonic influence on the frontier. This is not plausible for some environmental variables. For example, returning to the agricultural example given above, more rainfall would likely benefit farmers in west Texas, but farmers on the Meghalaya plateau in northeastern India have far too much rain; the optimal amount of rain is somewhere between these two extremes, implying a non-monotonic relationship between crop production and rainfall.[3]

In standard linear regression problems, researchers typically allow for non-monotonicity by including quadratic terms in the response function. Here, however, this approach does not help. Suppose that $\boldsymbol{Z} = \begin{bmatrix} Z_1 & Z_1^2 \end{bmatrix}$ and $\boldsymbol{\beta} = \begin{bmatrix} \beta_1 & \beta_2 \end{bmatrix}'$; then differentiating (8) yields

$$\frac{\partial Y}{\partial \boldsymbol{Z}'} = -(\beta_1 + 2\beta_2 Z_1)Y. \tag{12}$$

Given the assumptions $\boldsymbol{Z} \geq 0$ and $\boldsymbol{\beta} \geq 0$ made by BN, this is clearly negative (or zero if $\beta_1 - \beta_2 = 0$), although it is nonlinear. But, since $\frac{\partial Y}{\partial \boldsymbol{Z}'} \leq 0$, the effect of the environmental variable is still monotonic.

Sixth, $\boldsymbol{Z}$ affects the support of Y, not the inefficiency process, thereby violating the "separability" condition discussed by SW. This means that the environmental variables are assumed to only affect the production possibilities, but not the level of inefficiency.[4] While this might

[2] In the model considered by SW, inefficiency explicitly depends on the environmental variables which may account for heteroskedasticity in the inefficiency process. SW did not consider heteroskedasticity in the error term of the second stage regression, but this could be modeled using standard techniques; i.e., σ_ε^2 appearing in Assumption A3 of SW could be parameterized in terms of additional covariates. See also Park et al. (2008).

[3] The Meghalaya plateau in northeastern India is considered to be one of the rainiest places on earth (Murata et al. 2007).

[4] On p. 50, in the fourth through seventh lines after equation no. 2), it is stated that

> The contextual variables are measured such that the weights $\beta_s, s = 1, \ldots, S$, are all nonnegative—i.e., the higher the value of the contextual variables, the higher is the inefficiency of the DMU.

This is false due to the structure in (8) and the independence of $\boldsymbol{Z}$ and U.

be reasonable for some situations, it is a maintained assumption that should be tested. In addition, this is rather different from the numerous papers cited by SW, where environmental variables affect the level of efficiency, but not the production possibilities themselves. Moreover, the fact that $\boldsymbol{Z}$ is assumed to have a *monotonic* effect on the frontier means that the environmental variables could be transformed (e.g., by taking their reciprocals) so that their effects are positive (instead of negative) and then treated as inputs. This was the approach of Banker and Morey (1986), which is not mentioned in BN. The Banker and Morey approach allows dependence between Z and X, and is more flexible in the sense that it allows Z to affect efficiency as well as the frontier. In addition, bootstrap methods could be used to test whether Z has an effect on the production process.

Seventh, BN state in their footnote number 3 (p. 50),

> Our extension to the multiple-output case involves an additional vector of random variables specifying the proportion of each output. The DGP then determines the output vector Y_j as in the single-output case on the ray defined by the vector of random variables specifying the output mix.

Due to the imprecision, it is difficult to know with certainty what is meant by this. Apparently, this means one should use the right-hand side of (8) to generate a quantity Y^*, then draw $q-1$ (for $q>1$) random variables α_j on [0, 1], and finally compute $Y_1=\alpha_1 Y^*,\ldots,Y_{q-1}=\alpha_{q-1}Y^*$ and $Y_q=(1-\sum_{j=1}^{q-1}\alpha_j)Y^*$. However, although Y^* is by construction a convex combination of $Y_1,\ldots,Y_q$, the resulting technology is not necessarily convex if inputs are also multivariate (i.e., $p>1$). Hence, extending the model in (8) to allow for multiple outputs (i.e., $q>1$) while preserving convexity of the production set is problematic.

3.2 OLS in the second stage

BN (Sect. 3.1) propose using OLS in a second-stage regression of Gstach (1998) efficiency estimates on the environmental variables $\boldsymbol{Z}$. Specifically, they define

$$\widetilde{\phi}(X)=\phi(X)e^{V^M} \tag{13}$$

and

$$\widetilde{\theta}=e^{(V-V^M)-U-\boldsymbol{Z\beta}}\leq 1, \tag{14}$$

which is the quantity estimated by the Gstach (1998) estimator. Then

$$Y=\widetilde{\phi}(X)\widetilde{\theta} \tag{15}$$

after substituting (13), (14) into (8) and where $\theta=\widetilde{\theta}e^{V^M}$.

From (8) and (14) it follows that

$$\log\widetilde{\theta}=\beta_0-\boldsymbol{Z\beta}+\delta \tag{16}$$

where $\beta_0=E(V-U)-V^M$ and $\delta=V-U-E(V-U)$ so that $E(\delta)=0$; this corresponds to BN's equation (10) after correcting typos in their paper (note that we have re-defined δ here as a residual, in order to follow the notation appearing in BN). BN observe correctly that $\widetilde{\theta}$ is unobserved, and propose replacing $\log\widetilde{\theta}$ on the left-hand side of (16) with

$$\log\widehat{\widetilde{\theta}}=\log\widetilde{\theta}+\eta, \tag{17}$$

i.e., the log of the estimate $\widehat{\widetilde{\theta}}$ obtained using the usual DEA estimator and sample observations on X and Y. Doing so yields

$$\log\widehat{\widetilde{\theta}}=\beta_0-\boldsymbol{Z\beta}+\widetilde{\delta} \tag{18}$$

where $\widetilde{\delta}=\delta+\eta$. Proposition 1 in BN states that the OLS estimator $\widehat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ in (18) is consistent. Indeed, (18) is asymptotically equivalent to (16) under the assumptions discussed above.[5]

While this is true under the assumptions of their model, at this point it should be clear that several of BN's assumptions are crucial to their results. In particular, if Z and U are not independent, then OLS estimation of $\boldsymbol{\beta}$ in (18) will not be consistent. In addition, if X and Z are dependent, and if X and U are dependent (as would be the case, for example, if larger firms are more efficient than smaller firms), then the OLS estimator will again be inconsistent. If V^M is not constant, then it is not clear what would be estimated by OLS (or any other estimator) in (18); apart from this, if V^M varies systematically with X (and hence the size of the firm, which may be likely as argued above), OLS is once again inconsistent.[6]

An additional problem arises in the proof of Proposition 1 appearing in BN2 and referenced on p. 52 of the BN paper, where it is claimed that

$$n^{1/2}\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)\xrightarrow{d}N\left(\mathbf{0},\sigma^2\boldsymbol{Q}^{-1}\right) \tag{19}$$

where $\boldsymbol{Q}=\mathrm{Plim}\left(\frac{\boldsymbol{Z'Z}}{n}\right)$ is a positive definite matrix. While the claim is true for $p+q<3$, as demonstrated in the Appendix the claim is false for cases where $p+q\geq 3$. This is because in the proof, the roles of the bias of DEA

[5] BN write (18) as $\log\widehat{\widetilde{\theta}}=\widetilde{\beta}_0-\boldsymbol{Z}\widetilde{\boldsymbol{\beta}}+\widetilde{\delta}$ in their equation (11), but substitution of the right-hand side of (17) for $\widetilde{\theta}$ on the left-hand side of (16) does not change the parameters on the right-hand side of (16). Equation (17) appears as equation (A3) in BN2, where it is noted that $\eta\geq 0$.

[6] In addition, if V^M is not constant, it is equally unclear what is estimated in the first stage.

estimates and the correlation among DEA estimates is ignored. This correlation is bounded, and disappears asymptotically, but only at a slow rate; see Simar and Wilson (2011, 2011a) for details. Since the bias and variance are unknown, the asymptotic normality result cannot be used for inference about $\boldsymbol{\beta}$. Consequently, bootstrap methods along the lines of SW are needed for inference, after adapting the methods of SW to account for the particular features of the BN model.

Unfortunately, a number of unsuspecting empirical researchers have taken the BN results at face value. For example, Erhemjamts and Leverty (2010) obtains some DEA estimates in a first stage exercise, then regresses these using OLS in a second stage regression while citing BN to justify this, but without testing or even questioning the assumptions of the BN model. The results that are reported in Table 4 of Erhemjamts and Leverty (2010) for OLS estimates of $\boldsymbol{\beta}$ include standard error estimates, which are presumably the usual OLS standard error estimates suggested by the claim in the proof of BN's Proposition 1. As discussed in the previous paragraph, however, the stated result in the proof is incorrect, and the OLS standard error estimates do not give consistent estimates of the standard error of $\widehat{\widetilde{\boldsymbol{\beta}}}$.[7]

3.3 Maximum likelihood estimation in the second stage

BN discuss in Sect. 3.2 of their paper how $\boldsymbol{\beta}$ can be estimated by ML, and claim in their Proposition 2 on p. 52 that the maximum likelihood estimator of $\widetilde{\boldsymbol{\beta}}$ is consistent. However, since the estimation here involves replacing a true efficiency measure with a DEA estimate, the implications for the ML estimator and for inference are similar to the case where OLS is used in the second stage—the problem is similar to that described by SW. For reasons that will become clear, bootstrap methods appear to be the only available method for valid inference about $\boldsymbol{\beta}$.

In addition to these problems, the approach here requires that one assume specific forms for the inefficiency and noise terms. In this respect, their approach is no less restrictive than the assumption of truncated normality by SW. Following either approach, distributional assumptions are required; while various assumptions can be made, both approaches require a distributional assumption in the second stage.

In Sect. 3.3 of their paper, BN discuss estimation of individual efficiencies when maximum likelihood has been used in the second stage. Their approach is similar to that of Jondrow et al. (1982); the conditional density of U given $V - U$ is derived while accounting for the bounds on V, and use this to derive the conditional mean $E(U \mid V - U)$.

BN remark (in the fourth line after their equation 15 on p. 52) that

$E(U \mid \epsilon)$ is a consistent estimator of U *given* ϵ,

where $\epsilon = V - U$. This is not true. First, $E(U \mid \epsilon)$ contains unknown parameters which must be estimated. If these unknown parameters are replaced with consistent estimators, then the resulting expression is a random variable depending on ϵ, which is unobserved. Of course, ϵ can be replaced with an estimated residual, but the result cannot be an estimator of U because U is a random variable. Random variables can be predicted, but not estimated; in addition, any meaningful and interesting prediction of a *continuous* random variable necessarily involves an interval, as opposed to a point. Moreover, U is unidentified; only the difference $V - U$ can be identified in the model. Within the model, it is impossible to distinguish, for example, $U = 0.5$, $V = 1.5$ from $U = 1$, $V = 2$, or an infinite number of other possibilities. It is also important to remember that consistency, while a fundamental property of an estimator, is also a weak property—nothing can be learned from a consistent estimate unless valid inference can be made. Simar and Wilson (2010) discuss in their Sects. 3.2 and 3.3 what can be estimated consistently from composite error models such as the one considered by BN as well as how valid inference can be made.

3.4 Simulations

In Sect. 4.1 of their paper, BN describe the design of their Monte Carlo experiments. Their simulated technology (they only consider $p = q = 1$) is a cubic polynomial in inputs. In all of their experiments, their *true* model is

$$Y = (-37 + 48X - 12X^2 + X^3)e^{-0.2Z+V-U}, \tag{20}$$

where X is uniform on [1,4], Z is uniform on [0,1], $U \sim N^+(0, 0.0225)$, and V is truncated $N(0,0.0016)$ with truncation at -0.24 and 0.24. In addition, the random variables X, Z, U, ad V are drawn independently, consistent with the assumptions of the BN model. The frontier corresponding to (20) is plotted in Fig. 4. From the illustration, the effect of the assumption that the environmental variable has a monotonic effect on the frontier is clear. It is equally clear from Fig. 4 that if $1/Z$ were included as an input, efficiency could be estimated in one stage, avoiding the problems of two-stage estimation. One could use either ordinary DEA

[7] Erhemjamts and Leverty (2010) is not alone in taking statements in BN uncritically and without question. Both McDonald, (2009, p. 797) and Ramalho et al. (2010, Sect. 2, eighth paragraph) state that the DGP proposed by BN is less restrictive than that considered by SW, without mentioning the various restrictions required by the BN model. This issue is revisited below in Sect. 5

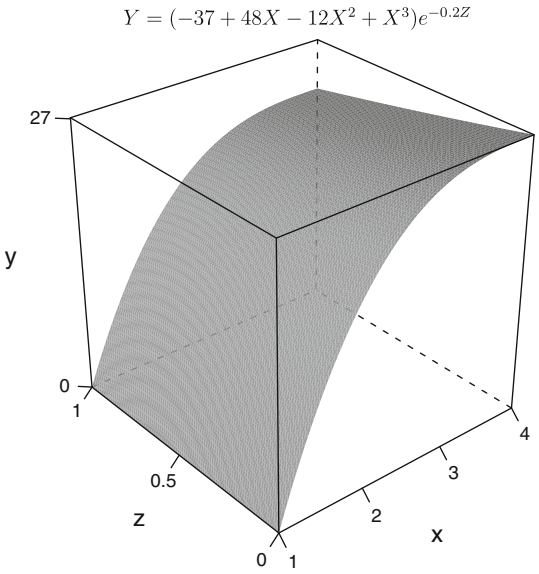

Fig. 4 Frontier in simulations of Banker and Natarajan (2008)

estimators or perhaps directional distance functions in order to estimate efficiency for given levels of *Z*.

In the experiments described in Sect. 4 of their paper, BN consider 12 different estimation procedures. The different procedures are evaluated in terms of the point estimates of β produced by each method, using root mean square error and mean absolute deviations as criteria by which to evaluate performance. BN do not attempt to make inferences about β in their experiments, nor do they consider inference about efficiency or anything else in their model. As noted above, conventional inference in the second stage is invalid, and bootstrap methods provide the only known (so far) route to valid inference about $\boldsymbol{\beta}$ in the BN model.

Since the simulated model represented by (20) is fully parametric, BN are able to consider various parametric estimation approaches as well as DEA. Their methods #5–12 involve either maximum likelihood, OLS, or corrected OLS estimation of translog or Cobb-Douglas functions, which mis-specify the simulated model in (20). It should be no surprise that these approaches do not provide good estimates of β. However, in Sect. 5 of their paper, in the sentence following the second quote given above in the first paragraph of Sect. 3.1, BN state that

> Results from extensive Monte Carlo simulations indicate that two-stage DEA-based procedures with OLS, ML, or even tobit estimation in the second stage significantly outperform the parametric methods.

This is similar to what is written in the abstract:

> Simulation results indicate that DEA-based procedures with OLS, maximum likelihood, or even tobit estimation in the second stage perform as well as the best [emphasis added] of the parametric methods in the estimation of the impact of contextual variables on productivity.

A similar statement is made at the end of the seventh paragraph of Sect. 1 in BN. None of these statements are true in general. It is true that the DEA-based methods outperform estimation with the mis-specified parametric models, which is not surprising. But as the results in Table 1 of the BN paper clearly show, the DEA-based methods do not perform as well as the parametric methods when the model is correctly specified (method no. 3 in BN). Moreover, as discussed above, the DEA-based methods perform well *under the numerous assumptions of the BN model* discussed above, but cannot be expected to perform well *in general*. The results are specific to the model defined by BN, with all of its restrictions.

4 The "Instrumentalist" approach

With regard to second-stage regressions of DEA efficiency estimates on environmental variables, it is apparently not uncommon to adopt the view that in the second stage, the DEA "scores" are simply astatistical, atheoretical measures of distance to an observed "best-practice frontier" (e.g., Hoff 2007; McDonald 2009; Ramalho et al. 2010). McDonald (2009) calls this the "instrumentalist" approach. Ramalho et al. (2010) summarizes the view by noting that in this framework,

> …DEA scores are treated as descriptive measures of the relative technical efficiency of the samples DMUs. Given this interpretation, the frontier can be viewed as a (within-sample) observed best-practice construct and, therefore, in stage two, the DEA scores can be treated like any other dependent variable in regression analysis. Hence, parameter estimation and inference [emphasis added] in the second stage may be carried out using standard procedures.

One can certainly view DEA scores as simply measured distance to an observed best-practice frontier. However, one might reasonably ask whether if an entrepreneur starts a new firm, will it lie beyond this observed best-practice frontier, and if so, how far might it lie beyond? Or one might ask whether the observed firms can improve their performance, and if so, by how much? Can the firms on the observed "best-practice" frontier improve their performance? Again, if so, by how much? Such questions can only be answered by *inference*. And, to be meaningful, inference requires a coherent, well-defined statistical model describing the DGP and providing a probabilistic structure

for inputs, outputs, and environmental variables. Such a model is conspicuously absent in Hoff (2007), McDonald (2009) and Ramalho et al. (2010).

In addition, if one posits a second-stage regression model with DEA "scores" on the left-hand side, then these must be viewed as random variables if a stochastic error term is included on the right-hand side. If inference is to be made in the second-stage regression, then the error term must be stochastic, for inference is neither meaningful nor well-defined otherwise. If the error term, and hence the DEA scores are viewed as random, then one must consider from where the DEA scores have come. In two-stage approaches, the DEA scores come from a first stage; hence the first-stage model will determine what is appropriate in the second stage regression.

As the discussion in Sects. 2 and 3 have revealed, the structure assumed in the first stage is crucial for determining what type of model should be estimated in the second stage. In the model considered by SW, the "separability" condition discussed in Sect. 2 in order to interpret the first-stage efficiency estimates sensibly. In the BN model, it is equally important that the environmental variables be independent of inefficiency and have a monotonic, exponential-linear effect on the frontier for similar reasons. Simply positing an ad hoc second-stage regression equation without considering a statistical model for the first stage amounts to a type of reduced form model in which it is hard to know what is being estimated.

5 Summary and conclusions

Footnote 1, near the end of Sect. 1 in the BN paper, states that

> The DGP assumed by Simar and Wilson is more restrictive than the DGP considered in this study because it does not contain a two-sided noise term and also imposes a DMU-specific truncated normal density on the inefficiency term. Based on their restrictive setup, Simar and Wilson argue that ML estimation of a truncated regression rather than Tobit is the preferred approach in a second-stage analysis that relates the DEA productivity estimator to the contextual variables.

This refrain has been repeated almost verbatim by others, including McDonald (2009, p. 797) and Ramalho et al. (2010, Sect. 2, eighth paragraph). It is true that BN allow for noise, while SW do not. However, as discussed above in Sect. 3, the noise allowed by BN must be (1) bounded, and (2) the bounds must be constant. The second assumption—that the bounds must be constant—was shown in Sect. 3 to be critical to the success of the BN approach. However, this is a strong assumption, akin to assuming homoskedasticity, which is frequently violated with cross-sectional data, and especially with data used in production or cost functions.

It is also true that SW assume a truncated normal density in their Assumption A3. Necessarily, the numerous studies that have employed tobit estimation in second-stage regressions have assumed a censored normal density. Again, the goal of SW was to match as closely as possible what empirical researchers have been doing while providing a well-defined statistical model in which a second-stage regression would be meaningful. Other assumptions can be made, or the second stage regression can be estimated non-parametrically using the local ML method discussed by Park et al. (2008). Moreover, as discussed above in Sect. 3.3, BN also introduce distributional assumptions when ML estimation is used in the second stage.

In addition, as the discussions in Sects. 2 and 3 have made clear, the BN model requires several additional assumptions that are much more restrictive than those required by the model described by SW. The BN model assumes that the effects of the environmental variables are monotonic; the model described by SW does not. The BN model assumes that the environmental variables are independent with respect to the input variables; the model described by SW does not. The BN model assumes that the inefficiency process is independent of the input variables; the model described by SW does not.

The BN model assumes that the environmental variables only affect the frontier, but not the inefficiency process; the model described by SW assumes that the environmental variables only affect the inefficiency process, but not the frontier (this is the "separability" condition described above). Hence both models are restrictive in terms of what the environmental variables are assumed to affect. As noted above, SW warn that the "separability" condition should be tested, and a method for testing this has been provided by Daraio et al. (2010). The corresponding assumption in the BN model should also be tested. In situations where environmental variables affect the frontier as well as the inefficiency process, one can use estimators of conditional measures of efficiency described by Daraio and Simar (2005).

Footnote 1 in the BN paper continues with the following:

> Although the Simar and Wilson paper substantially differs from this study in theoretical development and research design, our main result, that OLS is appropriate to evaluate the impact of contextual variables on productivity, is more robust and more appropriate for productivity research than Simar and Wilson's result that is valid under only much more restrictive assumptions about the DGP.

The reader can decide, in view of the preceding discussion, whether the model described by SW is more or less restrictive than the BN model. However, as the discussion in Sect. 3 has made clear, the claims that OLS is (1) appropriate for second stage regressions, and (2) more robust and more appropriate than the approach described by SW are not true in general, but instead depend crucially on the numerous assumptions underlying the BN model. As we have noted above, several of the assumptions required by the BN model are likely to be unsupported by economic data, and should in any case be tested.

Even if one were to accept all of the assumptions required by the BN model, problems remain for inference. It is not enough to obtain point estimates of $\boldsymbol{\beta}$ in the BN model; one must make inference before anything can be learned. Since the asymptotic bias and variance of OLS and ML estimators of $\boldsymbol{\beta}$ in the BN model are unknown, bootstrap methods are needed for valid inference about $\boldsymbol{\beta}$.

We do not recommend the use of second-stage regressions involving DEA efficiency scores. However, if one chooses to do so, the issues that have been raised here should be considered carefully. Regardless of whether one adopts the model considered by SW, the BN model, or some other model yet to be presented, one should carefully consider what restrictions are necessary, and whether these are reasonable. Ideally, restrictions should be tested. In addition, one should carefully consider how valid inference can be made. To do these things, one must have a coherent, well-defined statistical model. Finally, let the buyer beware—*caveat emptor*.

Acknowledgments Financial support from the ''Inter-university Attraction Pole'', Phase VI (No. P6/03) from the Belgian Government (Belgian Science Policy) and from l'Institut National de la Recherche Agronomique (INRA) and Le Groupe de Recherche en Economie Mathématique et Quantitative (GREMAQ),Toulouse School of Economics, Toulouse, France are gratefully acknowledged. Part of this research was done while Wilson was a visiting professor at the Institut de Statistique Biostatistique et Sciences Actuarielles, Université Catholique de Louvain, Louvain-la-Neuve, Belgium. We have benefited from discussions with Valentin Zelenyuk; of course, any remaining errors are solely our responsibility.

Appendix: OLS estimation in BN's second stage

The first stage estimation in BN's approach provides an estimator $\widehat{\widetilde{\theta}}_i \leq 1$ of $\widetilde{\theta}_i$ for $i = 1, \ldots, n$ where i indexes observations. The properties of DEA estimators have been developed by Korostelev et al. (1995a, b), Kneip et al. (1998), Kneip et al. (2008, 2011b), Park et al. (2010) and Simar and Wilson (2011), and depend on assumptions about returns to scale. In particular, if variable returns to scale (VRS) are assumed, then the DEA estimator converges at rate $n^{2/(p+q+1)}$, which is slower than the usual parametric rate $n^{1/2}$ for $p + q > 3$. BN ignore this in the proof (appearing in BN2) of their Proposition 1, and this leads to important errors and false statements.

BN suggest re-writing (17) as

$$\log \widetilde{\theta} = \log \widehat{\widetilde{\theta}} - \eta \tag{21}$$

and using the right-hand side of this to replace $\log \widetilde{\theta}$ in (16) to obtain (18). Then the error term $\widetilde{\delta}$ appearing in (18) is equal to $\delta + \eta$. BN propose estimating (18) by OLS, and claim in their proof of their Proposition 1 that

$$\sqrt{n}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \xrightarrow{d} N\left(\mathbf{0}, \sigma^2 \boldsymbol{Q}^{-1}\right) \tag{22}$$

where $\boldsymbol{Q} = \text{Plim}(n^{-1}\boldsymbol{Z}'\boldsymbol{Z})$.[8] As shown below, these claims are false.

Recall that $\eta_i \geq 0$ for all $i = 1, \ldots, n$, with i indexing the sample observations. Simar and Wilson (2011) and Kneip et al. (2011a) prove, under mild regularity conditions,

$$n^{\gamma}\eta_i \xrightarrow{\mathcal{L}} G(\mu_0, \sigma_0^2), \tag{23}$$

where $G(\cdot)$ is an unknown, non-degenerate distribution with mean $\mu_0 > 0$ and variance $\sigma_0^2 > 0$ (both finite and unknown), and $\gamma = 2/(p + q + 1)$ for the VRS case (or $\gamma = 2/(p + q)$ for the constant returns to scale (CRS) case). In addition, as shown in Kneip et al. (2008, 2011a, b), the asymptotic covariances between η_i and η_j is asymptotically non-zero for a number of observations $j = 1, \ldots, n, j \neq i$, which is of order $O(n^{\gamma})$. To summarize, as $n \to \infty$,

$$E(\eta_i) \approx n^{-\gamma}\mu_0, \tag{24}$$

$$\text{VAR}(\eta_i) \approx n^{-2\gamma}\sigma_0^2, \tag{25}$$

and

$$\text{COV}(\eta_i, \eta_j) \approx \left\{ \begin{array}{ll} n^{-2\gamma}\alpha & \text{for } O(n^{\gamma}) \text{observations } j \neq i; \\ 0 & \text{for the remaining observations} \end{array} \right\} \tag{26}$$

for some bounded but unknown constant α.

Recall that the error term $\widetilde{\delta}$ in (18), i.e., the equation that BN estimate by OLS, equals $\delta + \eta$ as shown above. Consequently, the properties of η play an important role in determining the properties of the OLS estimator $\widehat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$. Let $\boldsymbol{Z}$ be an $n \times (r + 1)$ matrix with ith row given by $[1 \quad -\boldsymbol{Z}_i]$, and let $\boldsymbol{Y} = \left[\log \widehat{\widetilde{\theta}}_1 \quad \cdots \quad \log \widehat{\widetilde{\theta}}_n\right]'$. In

[8] In the statement of their Proposition 1, BN correctly define $\boldsymbol{Q}$ as $\text{Plim}(n^{-1}\boldsymbol{Z}'\boldsymbol{Z})$, but in equation (A4) of the proof appearing in BN2, $\boldsymbol{Q}$ is implicitly defined as $n^{-1}\boldsymbol{Z}'\boldsymbol{Z}$. We use the definition $\boldsymbol{Q} = \text{Plim}(n^{-1}\boldsymbol{Z}'\boldsymbol{Z})$ in all that follows.

addition, let $\boldsymbol{\beta}^* = [\beta_0 \quad \boldsymbol{\beta}']'$ and $\widehat{\boldsymbol{\beta}}^* = \left[\widehat{\beta}_0 \quad \widehat{\boldsymbol{\beta}}'\right]'$. Then OLS estimation on (18) yields

$$\begin{aligned}\widehat{\boldsymbol{\beta}}^* &= (\mathscr{Z}'\mathscr{Z})^{-1}\mathscr{Z}'\mathscr{y} \\ &= (\mathscr{Z}'\mathscr{Z})^{-1}\mathscr{Z}'\left(\mathscr{Z}\boldsymbol{\beta}^* + \widetilde{\boldsymbol{\delta}}\right)\end{aligned} \tag{27}$$

where $\widetilde{\boldsymbol{\delta}} = \left[\widetilde{\delta}_1 \quad \ldots \quad \widetilde{\delta}_n\right]'$. Taking expectations,

$$\begin{aligned}E\left(\widehat{\boldsymbol{\beta}}^* \mid \mathscr{Z}\right) &= \boldsymbol{\beta}^* + (\mathscr{Z}'\mathscr{Z})^{-1}\mathscr{Z}'E(\widetilde{\boldsymbol{\delta}} \mid \mathscr{Z}) \\ &= \boldsymbol{\beta}^* + (\mathscr{Z}'\mathscr{Z})^{-1}\mathscr{Z}'E(\boldsymbol{\delta} \mid \mathscr{Z}) \\ &\quad + (\mathscr{Z}'\mathscr{Z})^{-1}\mathscr{Z}'E(\boldsymbol{\eta} \mid \mathscr{Z}) \\ &= \boldsymbol{\beta}^* + (\mathscr{Z}'\mathscr{Z})^{-1}\mathscr{Z}'E(\boldsymbol{\eta} \mid \mathscr{Z}) \\ &\approx \boldsymbol{\beta}^* + n^{-\gamma}c_1\end{aligned} \tag{28}$$

as $n \to \infty$, where c_1 is a non-zero, bounded constant, due to the result in (24) and since (by BN's assumptions) $E(\boldsymbol{\eta} \mid \mathbf{Z}) = E(\boldsymbol{\eta}), E(\boldsymbol{\delta} \mid \mathbf{Z}) = 0$, and where $\boldsymbol{\delta} = [\delta_1 \quad \ldots \quad \delta_n]'$ and $\boldsymbol{\eta} = [\delta_1 \quad \ldots \quad \delta_n]'$.

From the last line in (28) it is clear that as $n \to \infty$,

$$\sqrt{n}\left(\widehat{\boldsymbol{\beta}}^* - \boldsymbol{\beta}^*\right) \approx O_p\left(n^{\frac{1}{2}-\gamma}\right), \tag{29}$$

which is rather different from what is claimed in the proof of Proposition 1 of BN (as noted earlier, BN claim that (19) holds). Recall that $\gamma = 2/(p+q+1)$ for the VRS case. As shown below, the left-hand side of (29) converges to a non-degenerate random variable with constant variance for $p+q \leq 3$, and to a random variable with variance approaching infinity as $n \to \infty$ for $p+q > 3$. In the CRS case, $\gamma = 2/(p+q)$, and hence $\sqrt{n}(\boldsymbol{\beta}^* - \boldsymbol{\beta}^*)$ converges to a non-degenerate random variable with constant variance for $p+q \leq 4$, and to a random variable with variance approaching infinity as $n \to \infty$ for $p+q > 4$. For $p+q = 3$ in the VRS case and for $p+q = 4$ in the CRS case, the left-hand side of (29) converges to a random variable with constant variance, but which is not normally distributed as shown below. In their Monte Carlo experiments, BN considered only the case where $p = q = 1$ with VRS, and consequently did not notice the errors in their proof of their Proposition 1.

Combining the results in (24–26), and using standard central-limit theorem arguments (see Kneip et al. 2011a for mathematical details), we have

$$\sqrt{n}\left(\widehat{\boldsymbol{\beta}}^* - \boldsymbol{\beta}^*\right) \xrightarrow{\mathcal{L}} N\left(\mathbf{0}, \sigma^2\mathbf{Q}^{-1}\right) + \sqrt{n}\zeta_n, \tag{30}$$

where ζ_n is a random variable such that $\sqrt{n}\zeta_n = o_p(1)$ if $\gamma > 1/2$ or $\sqrt{n}\zeta_n = O_p\left(n^{1/2-\gamma}\right)$ otherwise, and $\sigma^2 = \mathrm{VAR}(\delta) = \mathrm{VAR}(V) + \mathrm{VAR}(U)$ (as in BN).[9] The result in (30) is very different from (22), which is the result claimed at the end of the proof appearing in BN2 of BN's Proposition 1. Although the OLS estimator $\widehat{\boldsymbol{\beta}}^*$ of $\boldsymbol{\beta}^*$ is consistent, (22) cannot be used for valid (asymptotic) inference. Moreover, even if $\gamma = 1/2$, (30) contains unknown constants and an unknown, bounded random variable. The left-hand side of (30) does not converge to anything that is bounded if $\gamma < 1/2$. Bootstrap methods appear to provide the only feasible avenue toward valid inference or hypothesis testing in the second-stage regression.[10]

The preceding discussion also illustrates how the numerous restrictive assumptions imposed on the BN model are crucial for consistency of OLS estimation in the second-stage regression. For example, if $\mathbf{Z}$ and U—which determines inefficiency—are correlated, then the error terms δ and $\widetilde{\delta}$ must be correlated with $\mathbf{Z}$, in which case OLS estimation in (18) would yield inconsistent estimates. As another example, if V^M, the bound on the noise process, is not constant, then OLS estimation may be problematic. If $V^M = \overline{V^M} + \zeta$, where $\overline{V^M}$ is constant and ζ is random with $E(\zeta) = 0$, then β_0 can be written as $\beta_0 = E(V-U) - \overline{V^M}$, but δ would have to be written as $\delta = V - U - E(V-U) - \zeta$. If $E(\zeta) \neq 0$, then OLS estimation of β_0 will be biased and inconsistent. Worse, regardless of whether $E(\zeta) = 0$, if ζ is not independent of $\mathbf{Z}$, then OLS estimation in (18) would yield inconsistent estimates of both β_0 and $\boldsymbol{\beta}$. If the environmental variables are related to the size of firms, and if the error bounds vary with firm size, the $\mathbf{Z}$ and ζ would clearly be correlated; this is likely to be the case in some applications.

Even more troubling is the assumption that V^M is finite, which implies that the noise term V is symmetrically truncated at $-V^M$ and V^M. Suppose, for example, that $V \sim N(0, \sigma_V^2)$, and suppose the researcher has a sample of n iid draws $\{V_1, \ldots, V_n\}$ from the $N(0,\sigma_V^2)$ distribution. Of course, one can easily find the sample maximum, and the maximum value in a normal sample of finite size will certainly be less than infinity. But, it is necessarily difficult, and maybe impossible, to test whether the distribution is

[9] In their proof appearing in BN2, BN ignore the role of the intercept β_0. Consequently, their expression for the variance of their OLS estimator would be wrong even if the rest of their derivations were correct, which they are not. In addition, in their Monte Carlo experiments, BN considered only the case where $p = q = 1$ with VRS, and consequently did not notice the errors in their proof of their Proposition 1.

[10] Most, if not all, of the papers that have used OLS to regress DEA efficiency scores on environmental variables while citing BN for justification have numbers of dimensions greater than three in their first-stage estimation. To give just a few examples, Cummins et al. (2010) use $p+q = 8$ or 9; Banker et al. (2010a) use $p+q = 6$; Banker et al. (2010b) use $p+q = 5$. Each of these rely on the usual OLS standard error estimate to make inference in the second-stage regressions, and consequently the inference in these papers is invalid.

truncated at a finite value. In situations in econometrics where truncated regression is used, the truncation typically arises from features of the sampling mechanism (e.g., survey design) or model structure (e.g., in SW, truncation arises from the fact that inefficiency has a one-sided distribution; it would make little sense to assume otherwise). Imposing finite bounds on a two-sided noise process, however, is a far more uncertain prospect.

If V^M is infinite, then the first-stage estimation using DEA estimators is inconsistent. From (13), it is clear that if V^M is infinite, then $\widetilde{\phi}(X)$ must be infinite. Re-arranging terms in (15) indicates that $\widetilde{\theta} = Y/\widetilde{\phi}(X)$ for the case of a univariate output considered by BN; hence if V^M is infinite, then $\widetilde{\theta}$ is undefined, in which case BN's second-stage regression is an ill-posed problem without meaning.

References

Aly HY, Grabowski CPRG, Rangan N (1990) Technical, scale, and allocative efficiencies in US banking: an empirical investigation. Rev Econ Stat 72:211–218

Banker RD, Cao Z, Menon N, Natarajan R (2010a) Technological progress and productivity growth in the US mobile telecommunications industry. Ann Oper Res 173:77–87

Banker RD, Lee SY, Potter G, Srinivasan D (2010b) The impact of supervisory monitoring on high-end retail sales productivity. Ann Oper Res 173:25–37

Banker RD, Morey RC (1986) Efficiency analysis for exogenously fixed inputs and outputs. Oper Res 34:513–521

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

Banker RD, Natarajan R (2008b) Online companion for "evaluating contextual variables affecting productivity using data envelopment analysis"—appendix: proofs of consistency of the second stage estimation. Oper Res online supplement, 1–6. Available at http://or.journal.informs.org/cgi/data/opre.1070.0460/DC1/1

Barkhi R, Kao YC (2010) Evaluating decision making performance in the GDSS environment using data envelopment analysis. Decis Support Syst 49:162–174

Bădin L, Daraio C, Simar L (2010) Optimal bandwidth selection for conditional efficiency measures: a data-driven approach. Eur J Oper Res 201:633–664

Chang H, Chang WJ, Das S, Li SH (2004) Health care regulation and the operating efficiency of hospitals: evidence from taiwan. J Account Public Policy 23:483–510

Chang H, Choy JL, Cooper WW, Lin MH (2008) The sarbanes-oxley act and the production efficiency of public accounting firms in suppying accounting auditing and consulting services: an application of data envelopment analysis. Int J Serv Sci 1:3–20

Cummins JD, Weiss MA, Xie X, Zi H (2010) Economies of scope in financial services: a DEA efficiency analysis of the US insurance industry. J Banking Finance 34:1525–1539

Daraio C, Simar L (2005) Introducing environmental variables in nonparametric frontier models: a probabilistic approach. J Prod Anal 24:93–121

Daraio C, Simar L (2006) A robust nonparametric approach to evaluate and explain the performance of mutual funds. Eur J Oper Res 175:516–542

Daraio C, Simar L, Wilson PW (2010) Testing whether two-stage estimation is meaningful in non-parametric models of production. Discussion paper #1031. Institut de Statistique, Université Catholique de Louvain, Louvain-la-Neuve, Belgium

Davutyan N, Demir M, Polat S (2010) Assessing the efficiency of turkish secondary education: heterogeneity, centralization, and scale diseconomies. Socio-Econ Plan Sci 44:3–44

Erhemjamts O, Leverty JT (2010) The demise of the mutual organizational from: An investigation of the life insurance industry. J Money Credit Banking 42:1011–1036

Farrell MJ (1957) The measurement of productive efficiency. J Royal Stat Soc A 120:253–281

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

Hoff A (2007) Second stage dea: comparison of approaches for modelling the dea score. Eur J Oper Res 181:425–435

Jeong SO, Park BU, Simar L (2010) Nonparametric conditional efficiency measures: asymptotic properties. Ann Oper Res 173:105–122

Jondrow J, Lovell CAK, Materov IS, Schmidt P (1982) On the estimation of technical inefficiency in the stochastic frontier production model. J Econ 19:233–238

Kneip A, Park B, Simar L (1998) A note on the convergence of nonparametric DEA efficiency measures. Econ Theory 14:783–793

Kneip A, Simar L, Wilson PW (2008) Asymptotics and consistent bootstraps for DEA estimators in non-parametric frontier models. Econ Theory 24:1663–1697

Kneip A, Simar L, Wilson PW (2011a) Central limit theorems for DEA scores: when bias can kill the variance. Discussion paper, Institut de Statistique Biostatistique et Sciences Actuarielles, Université Catholique de Louvain, Louvain-la-Neuve, Belgium

Kneip A, Simar L, Wilson PW (2011b) A computationally efficient, consistent bootstrap for inference with non-parametric DEA estimators. Comput Econ. (Forthcoming)

Korostelev A, Simar L, Tsybakov AB (1995a) Efficient estimation of monotone boundaries. Ann Stat 23:476–489

Korostelev A, Simar L, Tsybakov AB (1995b) On estimation of monotone and convex boundaries. Publications de l'Institut de Statistique de l'Université de Paris XXXIX 1:3–18

McDonald J (2009) Using least squares and tobit in second stage dea efficiency analyses. Eur J Oper Res 197:792–798

Murata F, Hayashi T, Matsumoto J, Asada H (2007) Rainfall on the Meghalaya plateau in northeastern India—one of the rainiest places in the world. Nat Hazards 42:391–399

Park BU, Jeong S-O, Simar L (2010) Asymptotic distribution of conical-hull estimators of directional edges. Ann Stat 38:1320–1340

Park BU, Simar L, Zelenyuk V (2008) Local likelihood estimation of truncated regression and its partial derivative: Theory and application. J Econ 146:185–2008

Ramalho EA, Ramalho JJS, Henriques PD (2010) Fractional regression models for second stage DEA efficiency analyses. J Prod Anal 34:239–255

Shephard RW (1970) Theory of cost and production functions. Princeton, Princeton University Press

Simar L, Wilson PW (2000) Statistical inference in nonparametric frontier models: the state of the art. J Prod Anal 13:49–78

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

Simar L, Wilson PW (2010) Estimation and inference in cross-sectional, stochastic frontier models. Econ Rev 29:62–98

Simar L, Wilson PW (2011) Inference by the m out of n bootstrap in nonparametric frontier models. J Prod Anal. (Forthcoming)

Sufian F, Habibullah MS (2009) Asian financial crisis and the evolution of korean banks efficiency: a DEA approach. Glob Econ Rev 38:335–369