Estimating Production Functions using Costs when Outputs are Restricted*

EMIR MALIKOV[1] SHUNAN ZHAO[2]
SUBAL C. KUMBHAKAR[3] DAVID H. BERNSTEIN[4]

[1]University of Nevada, Las Vegas
[2]Oakland University
[3]Binghamton University
[4]University of Miami

October 10, 2024

Abstract

Conventional proxy-variable approaches to production-function identification assume that output quantities are flexibly adjustable, with firms endogenously choosing them in pursuit of profits. But in not so few industries outputs are exogenously restricted, constraining input allocations by producers. Model assumptions of most proxy estimators become untenable in such settings. We provide a new structural methodology based on a dual formulation of firm production, indirectly identifying technology and productivity from cost data while appropriately accommodating exogenous output restrictions. We showcase our methodology on a panel of U.S. natural gas-fired power-generation plants in 2005–2017 facing price-inelastic demand that restricts their outputs in the short run. We find that power plants were scale-efficient, with (short-run) marginal costs of net electricity generation declining over time. With little evidence of productivity growth in the industry during our sample period, we conclude that the reduction in marginal costs was fueled mainly by diminishing input prices and, potentially, convergence to the optimal long-run input allocation.

Keywords: productivity, proxy variable, structural identification, endogeneity, electric power

JEL Classification: C20, C50, D24, L10, Q40

*Correspondence to: Emir Malikov, Department of Economics, Lee Business School, University of Nevada, Las Vegas, Las Vegas, NV, USA; emir.malikov@unlv.edu.

1 Introduction

Standard "proxy-variable" techniques for estimating production functions and firm productivity in the manner of Olley and Pakes (1996) and Levinsohn and Petrin (2003) assume that the quantity of output produced is flexibly adjustable and, in fact, is a firm's endogenous (static) choice in pursuit of maximum profits. However, in a fair few industries, notably in utilities and services including electricity, healthcare, transportation and even banking, output quantities are exogenously—and typically externally—predetermined, with firms taking them as a given constraint. Structural assumptions that most proxy-variable production-function estimators explicitly or implicitly build on become untenable in such settings. In this paper, we provide a new structural proxy-based approach that relies on a dual cost-function formulation of firm production, facilitating an indirect identification of the production technology and productivity from cost data while appropriately accommodating exogenous output quantity restrictions that producers face.

Proxy-variable methods for structural identification of production functions have become very popular among applied researchers because of their simplicity and reliable performance relative to alternative approaches to handling endogeneity induced by unobservable firm heterogeneity—dubbed "productivity" in this literature—that rely on fixed effects or prices as instruments. The idea underlining these estimators is simple: given the assumed structure, unobserved productivity can be proxied by inverting the firm's conditional demand for a variable input such as intermediates. To derive this variable input demand (or equivalently, the static first-order conditions), proxy-variable methodologies typically model firms as profit-maximizers that can freely choose their outputs. Continuing with this usual structural formalization of production behavior when the target output quantity is not a control but, in effect, an exogenous state variable is problematic because a proxy function derived from the firm's optimality conditions to control for latent productivity would be misspecified.[1] Further, the econometric treatment of output as a stochastic "outcome variable" in the production-function estimation is also dubious conceptually, unless a non-unit Jacobian of transformation is accounted for.

To accommodate exogenous predeterminedness of the output quantity, we dispense with a primal formulation of firm production in favor of a dual paradigm. Whether profit-maximizers or not, constrained by exogenous output restrictions, firms must at the least minimize their costs. Therefore, we adopt a cost-function-based formulation of firm production that provides a means to indirectly identify key technological metrics and productivity of the firm while, by definition, treating target outputs as constraints. Of note, cost minimization need *not* be the ultimate objective of firms who cannot freely adjust their outputs: we only assume that it is a necessary condition for their optimization. Our framework makes no assumption about firms' price-setting power in the (plausibly, monopolistically competitive) output market in which they may still seek to maximize profits given the output restrictions.[2]

[1] The target output quantity would be erroneously omitted from a proxy function. A functional form of the latter could also be misspecified unless the proxy is modeled nonparametrically.

[2] For instance, this is a popular structural formalization of how commercial banks operate; see Humphrey and Pulley

Besides properly accommodating output constraints, the additional non-trivial benefit of using the indirect cost-based approach is its ability to admit *multiple* outputs. Multi-product firms have lately been garnering much interest in the literature, with a number of papers attempting to extend structural proxy-variable techniques for production-function estimation to multi-output settings (see De Loecker et al., 2016; Grieco and McDevitt, 2017; Malikov and Lien, 2021; Dhyne et al., 2022; Orr, 2022; Valmari, 2023). While such extensions have proven to be quite non-trivial in a primal paradigm, requiring additional structural assumptions and, often, information beyond that which is available in most datasets in order to operationalize them (such as product-specific input allocations), the cost function is seamlessly generalizable to take on multiple outputs. For its structural identification, we only require that the underlying multi-product technology be homothetically separable in outputs which, in effect, is tantamount to the assumption of factor neutrality of latent firm productivity.[3]

Our cost-based identification methodology is also robust to cross-firm input price dispersion that may be endogenously correlated with firm productivity. Given the well-documented significant variation in input prices across firms and over time (e.g., Ornaghi, 2006; Foster et al., 2008; Kugler and Verhoogen, 2011), ignoring it or erroneously assuming it away can result in inconsistency of proxy-variable estimators. This happens because, when omitted from a proxy control function, these prices bias production technology estimates as they are correlated with inputs/outputs due to allocation adjustments that they induce within the firm. In our case, explicitly allowing for cross-firm price dispersion is also imperative because input prices are among immediate arguments of a cost function. Following Grieco et al. (2016), we do so while still maintaining the assumption of perfectly competitive markets for freely varying factors. That is, we continue to assume that firms are price-takers in variable input markets but the input prices are allowed to be *in*homogeneous across firms and over time. Although firms take these heterogeneous prices as given when making variable input allocations to minimize costs (i.e., no "quantity discounts"), we do not preclude prices from depending on other firm characteristics, notably productivity: e.g., a more productive firm may negotiate a better price from its suppliers. This is not a minor relaxation: the possibility that input prices may correlate with latent firm productivity renders them "econometrically" endogenous when estimating cost functions with proxies. To address this, we make use of a structural link between the cost function (of a known functional form) and the optimality conditions for flexible inputs derived from the firm's static cost-minimization problem subject to its quasi-fixed inputs and the output constraints. We propose a two-step estimation algorithm and show how to implement it under log-linear and log-quadratic specifications.

We demonstrate the ability of our cost-based estimator to successfully identify key technological parameters (such as returns to scale) through a small set of Monte Carlo simulations. The results are encouraging and show that our approach recovers the true parameters well. We also use simulations

(1997) and myriad studies adopting their "alternative profit function" approach that followed.

[3] Hicksian conceptualization of firm productivity continues to be the dominant modeling choice among proxy-variable production-function estimators. Few recent exceptions include Doraszelski and Jaumandreu (2018, 2019), Zhang (2019), Demirer (2020), Malikov et al. (2024), Zhao et al. (2024).

to demonstrate how failing to recognize exogenous output restrictions in the conventional primal production-function estimators can lead to biased and misleading estimates of technology. We then apply our methodology to a panel of U.S. natural gas-fired power plants in 2005–2017. A utility industry such as electricity generation provides one of the most apt settings for showcasing our methodology because output quantities that power plants produce are exogenously restricted and not freely adjustable (in a pursuit of maximum profits) in the short run. This is so because the demand for electricity is, effectively, perfectly price-inelastic, with plants constrained to match it. Formalizing a two-output cost function, we find that power plants are scale-efficient, operating at unitary returns to scale. Estimated (short-run) marginal costs of net electricity generation exhibit substantial variation, with an additional kilowatt-hour costing between 4 and 13 cents within the 10th-to-90th percentile range, but there is evidence of their decline over time. At the same time, we find little support for any significant productivity growth in the industry during our sample period. As such, the reduction in marginal costs was fueled mainly by diminishing input prices (particularly, natural gas) and, potentially, convergence to the optimal long-run input allocation.

Our paper contributes to the growth of a nascent literature on how to structurally identify production technologies of firms facing output constraints. To our knowledge, ongoing work by Atkinson and Luo (2024) is the only other attempt at adapting proxy-variable methods to such settings. Unlike us, in developing their approach they greatly focus on environmental aspects of coal-fired power generation and pollution control, compelling them to impose more structure onto the model which renders it distinctly application-specific. Our methodology formalizes a more general production setup and, thus, is more amenable to a variety of industries.[4] While both our and their papers estimate cost functions making use of Markovian properties of latent productivity common to a structural production-function analysis, we achieve model identification differently, without needing to supplement structural approach with instruments.

The rest of the paper is organized as follows. Section 2 describes the model of multi-output production subject to output constraints. We describe our identification strategy in Section 3 and detail the corresponding estimation procedure in Section 4. Section 5 discusses how to easily adapt our proposed estimator to accommodate the setting when output quantities are not restricted but flexibly chosen by firms. We examine the finite-sample performance of our methodology using simulations in Section 6 and provide an empirical illustration in Section 7. Section 8 concludes.

2 Model Setup

Consider the production process of a multi-product firm $i=1,\dots,N$ at time $t=1,\dots,T$ that involves both the quasi-fixed (subject to adjustment frictions) and flexible factors of production. We denote

[4] More specifically, Atkinson and Luo (2024) develop a three-step estimation procedure to estimate a hedonic coal price function of coal plants in the first step, an SO_2 abatement cost function in the second step and a (production) cost function in the third step. Their focus is on the trade-off that coal plants face to reduce SO_2 emission: purchasing permits, switching to cleaner fuels or operating flue gas desulfurization. As a result, their model and estimation methodology are customized specifically to their data and application.

these inputs as $\boldsymbol{K}_{it} = (K_{1,it},\ldots,K_{M,it})'$ and $\boldsymbol{V}_{it} = (V_{1,it},\ldots,V_{J,it})'$, respectively. For instance, consistent with the empirical evidence on adjustment costs, it is common in the literature to treat physical capital as a quasi-fixed input (subject to time-to-install costs), whereas inputs such as labor, intermediates/materials and energy are typically modeled as being freely varying and without dynamic implications.

Before we introduce a multi-output formulation of the production process, let us first consider a more typical case when the firm is thought to produce a single output (Q_{it}) which is formalized using the standard production function $F(\cdot)$ with factor-neutral productivity:

$$Q_{it} = F(\boldsymbol{K}_{it},\boldsymbol{V}_{it})\exp\{\omega_{it}\} \quad \Leftrightarrow \quad \underbrace{Q_{it}\exp\{-\omega_{it}\}}_{Q^*_{it}} = F(\boldsymbol{K}_{it},\boldsymbol{V}_{it}), \tag{2.1}$$

where ω_{it} is the latent firm productivity, and Q^*_{it} is the net producible output (net of Hicksian productivity, that is) given the inputs.

Note that the firm's production process in (2.1) can be formulated more generally in an implicit-function form using the so-called "transformation function" $T(\cdot)$ (see Shephard, 1971; Chambers, 1988), viz., $T\left(Q^*_{it},\boldsymbol{K}_{it},\boldsymbol{V}_{it}\right) = 1$, the extension of which to multiple outputs then becomes immediately apparent:

$$T\Big(\underbrace{\boldsymbol{Q}_{it}\exp\{-\omega_{it}\}}_{\boldsymbol{Q}^*_{it}},\boldsymbol{K}_{it},\boldsymbol{V}_{it}\Big) = 1, \tag{2.2}$$

where $\boldsymbol{Q}_{it} = \left(Q_{1,it},Q_{2,it},\ldots,Q_{D,it}\right)'$ includes $D \geq 1$ observable output quantities produced by the firm.

In what follows, we characterize structural assumptions about the firm's technology, productivity, economic environment and the timing of its decision-making process which facilitate a structural identification of this multi-product firm's production technology and (relative) productivity.

Assumption 1 (i) *Technology relating inputs and multiple outputs is represented by the transformation function $T(\cdot)$ that is continuous, twice differentiable, nondecreasing in inputs, nonincreasing in outputs, concave in inputs and outputs, and homothetically separable in outputs: $T\big(Y\left(\boldsymbol{Q}_{it}\exp\{-\omega_{it}\}\right), \boldsymbol{K}_{it},\boldsymbol{V}_{it}\big) = 1$, where $Y(\cdot)$ is homogeneous of arbitrary degree ϕ^Q.* (ii) *Among the factors of production: $\boldsymbol{K}_{it}$ are dynamic inputs subject to adjustment frictions and $\boldsymbol{V}_{it}$ are flexible inputs with no dynamic implications.* (iii) *Firm productivity ω_{it} is scalar, defined at the firm level and scales all outputs equiproportionately.*

Except for the normalizing assumption that $T(\cdot)$ is homothetically separable, the regularity conditions on the transformation function follow from the standard neoclassical assumptions about the production technology (e.g., see Diewert, 1973). Although, at first, the additionally imposed separability:

$$T\big(\boldsymbol{Q}_{it}\exp\{-\omega_{it}\},\boldsymbol{K}_{it},\boldsymbol{V}_{it}\big) = T\left(Y\left(\boldsymbol{Q}_{it}\exp\{-\omega_{it}\}\right),\boldsymbol{K}_{it},\boldsymbol{V}_{it}\right), \tag{2.3}$$

where $Y(\cdot)$ is a homogeneous output index function, might appear somewhat restrictive, notice that it is also routinely imposed onto a standard production function with Hicksian productivity akin to (2.1) whereby the (single) output is assumed to be separable from an input index $F(\cdot)$ with the corresponding degree of homogeneity set to $\phi^Q = 1$. Also, see Malikov and Lien (2021), O'Donnell (2014) and Greene (2008) for some recent examples of separability of the production technology being used in a multi-product settings. In our case, output separability is, in effect, tantamount to the assumption of factor neutrality of firm productivity which significantly aids the identification. Thus, in line with the go-to formulation in the production function estimation literature, by Assumptions 1(i) and (iii) we continue to maintain that latent firm heterogeneity ω_{it} is not only scalar but also Hicks-neutral. Given the multiple products, the formulation in (2.2) also implies that ω_{it} expands all outputs radially. Extensions to allow for multidimensional and/or factor-augmenting productivities is a promising area for future research.

Further, since $\boldsymbol{K}_{it}$ are subject to adjustment costs, the firm optimizes them dynamically at time $t-1$ rendering them predetermined state variables at time t. On the other hand, freely varying inputs $\boldsymbol{V}_{it}$ are chosen statically and contemporaneously with the time of production. While the exact number of flexible inputs is to be determined by the empirical application, we require that there is at least one such input (i.e., $J \geq 1$) because our identification strategy uses variation in *variable* costs.

Let Ξ_{it} denote the information available to the firm i for making period t production decisions.

Assumption 2 *Firm productivity ω_{it} evolves according to an exogenous time-homogeneous first-order Markov process: $\mathscr{P}_\omega(\omega_{it}|\Xi_{it-1}) = \mathscr{P}_\omega(\omega_{it}|\omega_{it-1})$.*

Following the convention in proxy variable literature, latent firm productivity is said to be first-order Markovian. Here, for simplicity, we assume it evolves exogenously (e.g., see Olley and Pakes, 1996; Levinsohn and Petrin, 2003; Ackerberg et al., 2015; Gandhi et al., 2020) but it is straightforward to generalize to a "controlled" process à la Doraszelski and Jaumandreu (2013) or De Loecker (2013) by incorporating productivity modifiers such as lagged R&D, export status, etc., depending on the empirical context. Assumption 2 implies the following regression for ω_{it}:

$$\omega_{it} = g(\omega_{it-1}) + \zeta_{it}, \tag{2.4}$$

where $g(\cdot) \equiv \mathbb{E}[\omega_{it}|\omega_{it-1}]$ is a conditional-expectation function, and ζ_{it} is the mean-orthogonal productivity innovation normalized to have a zero mean: $\mathbb{E}[\zeta_{it}|\Xi_{it-1}] = \mathbb{E}[\zeta_{it}] = 0$.

Assumption 3 *The output quantities $\boldsymbol{Q}_{it}$ for the time t production are exogenously predetermined, and the firm takes them as given. Conditional on Ξ_{it-1}, log-outputs at time t are mean-orthogonal to the firm's contemporaneous productivity draw ω_{it}.*

This assumption echos our discussions in the introduction whereby, in many service industries like utilities, the outputs that firms supply are exogenously given, with the firms having little, if any, ability to deviate from the demanded quantities. In effect, here we implicitly assume that firms face perfectly inelastic (residual) demand in the short run, but we place no restrictions on their pricing behavior.

Because the restricted nature of outputs, which the firm must take as a given constraint in their optimization at time t, does not on default guarantee their "exogeneity" in an econometric sense, we formalize the latter in a form of $\mathbb{E}[\ln \boldsymbol{Q}_{it}|\omega_{it}, \Xi_{it-1}] = \mathbb{E}[\ln \boldsymbol{Q}_{it}|\Xi_{it-1}]$ in Assumption 3. In positing that exogenous log-outputs at time t have zero partial correlations with ω_{it}, conditional on past information Ξ_{it-1}, we effectively rule out contemporaneous correlation between them and productivity innovation ζ_{it}. This admits a number of alternative timing schedules of when and how firms might observe their output quantities. For example, it is consistent with the following scenario. The firm learns the (log of) output quantity to supply at time t contemporaneously and uses this information to make its optimal variable input allocations. However, since predetermined quasi-fixed inputs do not become productive until the next period t, firms may optimally choose them at time $t-1$ based on the MSE-optimal forecasts of future log-outputs: $\ln \overline{\boldsymbol{Q}}_{it+s} \equiv \mathbb{E}[\ln \boldsymbol{Q}_{it+s}|\Xi_{it-1}]$ for $s = 0, 1, 2, \dots$. Following the firm's learning of actual output quantities it needs to supply $\boldsymbol{Q}_{it}$ at time t, given the optimality of output forecasts the corresponding one-period-ahead forecast errors in logs $\boldsymbol{\epsilon}_{it} = \ln \boldsymbol{Q}_{it} - \ln \overline{\boldsymbol{Q}}_{it}$ should be unforecastable as are the random productivity innovation ζ_{it}. Assumption 3 further restricts these errors to be conditionally mean-independent.

Assumption 4 *Markets for flexible factors are perfectly competitive. Suppliers use linear pricing, but input prices may be heterogeneous across firms and over time and, moreover, may depend on firm characteristics, including productivity.*

Here we relax the assumption of homogeneous and exogenous factor prices commonly made in many productivity studies—implicitly or explicitly—which is usually the case when the information on input prices is unavailable (see Olley and Pakes, 1996; Levinsohn and Petrin, 2003; Gandhi et al., 2020; Kim et al., 2019; Flynn et al., 2019; Malikov and Lien, 2021; Malikov et al., 2022, 2024, and many references therein). Following Grieco et al. (2016), we continue to assume that firms are price takers in the variable input markets but the prices need no longer be the same across firms: all variable input prices $\boldsymbol{P}_{it}^{V} = (P_{1,it}^{V}, \dots, P_{J,it}^{V})'$ are therefore indexed by i. More importantly, per Assumption 4, although firms take $\boldsymbol{P}_{it}^{V}$ as given when making input allocation decisions concerning flexible $\boldsymbol{V}_{it}$, these prices are allowed to correlate with their other (observed and unobserved) characteristics. Thus, while we rule out the possibility of quantity discounts which would endogenize input prices with respect to purchasing decisions about $\boldsymbol{V}_{it}$, firms are however allowed to be offered heterogeneous prices based on their differential productivity levels or their size (e.g., capital stock). This is supported by empirical evidence. For example, drawing on the uncommonly rich data from the Colombian Manufacturing Census, Kugler and Verhoogen (2012) find a positive connection between input prices and plant size: a 10% greater plant size is associated with about 0.11–0.12% higher prices paid for material inputs. Typically, firm size and productivity are closely correlated. Using a large panel of publicly traded U.S. firms in the 1970–1989 period, Dhawan (2001) finds that small firms are significantly more productive but also riskier than their larger counterparts, implying a trade off between efficiency and riskiness. Additionally, Manova and Zhang (2012), Verhoogen (2008) and Tybout (2003) find that exporting plants tend to pay higher input prices than non-exporting plants, and one of the main empirical regularities characterizing exporters is that they tend to be more productive than non-exporters (De Loecker, 2007). Therefore, input prices and productivity can be correlated through channels such as firm size and exporting behavior.

Although at first this might appear as a minor relaxation, it has important implications for identification. Namely, by allowing cross-firm input price dispersion to correlate with latent firm productivity, we immediately invalidate the potential use of firm-level input prices as instruments such as the case in, e.g., Doraszelski and Jaumandreu (2013, 2018, 2019) or De Loecker and Warzynski (2012). This also invalidates the use of control functions to proxy for (exogenous) input price dispersion, a modeling approach oft adopted when such prices are unobserved in the data (e.g. De Loecker et al., 2016). As it is to become apparent later on, the possibility that flexible input prices may depend on firm productivity can render them "endogenous" in an econometric meaning of the term which needs be addressed.

Assumption 5 *After observing their productivity draw ω_{it}, outputs $\boldsymbol{Q}_{it}$ and input prices $\boldsymbol{P}_{it}^{V}$ firms minimize (short-run) variable costs in a static optimization at time t, given the already predetermined dynamic inputs.*

Because firms have no monopsony power in the markets for variable inputs, the static cost minimization,

$$\min_{\mathrm{V}_{1,it},\ldots,\mathrm{V}_{J,it}} \quad \sum_{j=1}^{J} P_{j,it}^{V}\mathrm{V}_{j,it} + \lambda_{it}\left[T\Big(Y\left(\boldsymbol{Q}_{it}^{*}\right),\boldsymbol{K}_{it},\mathrm{V}_{1,it},\ldots,\mathrm{V}_{J,it}\Big)-1\right], \tag{2.5}$$

where λ_{it} is the Lagrangian multiplier, yields the usual variable cost function under perfect competition (see McFadden, 1978; Chambers, 1988):

$$C_{it} = H\left(\boldsymbol{P}_{it}^{V},\boldsymbol{K}_{it},Y\left(\boldsymbol{Q}_{it}^{*}\right)\right) = H\left(\boldsymbol{P}_{it}^{V},\boldsymbol{K}_{it},Y\left(\boldsymbol{Q}_{it}\right)\exp\{-\phi^{Q}\omega_{it}\}\right), \tag{2.6}$$

where $C_{it}=\sum_{j=1}^{J}P_{j,it}^{V}V_{j,it}$ is the (observed) minimum variable costs, and we have made use of the ϕ^{Q}-degree homogeneity of the output index $Y(\cdot)$ in the last equality. Under the regularity conditions on $T(\cdot)$ in Assumption 1, it follows that the short-run variable cost function $H(\cdot)$ is continuous, differentiable, nondecreasing, concave and linearly homogeneous in factor prices $\boldsymbol{P}_{it}^{V}$, nondecreasing and separable in the net producible outputs $\boldsymbol{Q}_{it}^{*}$. Given linear homogeneity, using any one of input prices—say, $P_{1,it}^{V}$ —we can rewrite (2.6) equivalently as

$$\widetilde{C}_{it} = H\left(1,\widetilde{\boldsymbol{P}}_{it}^{V},\boldsymbol{K}_{it},Y\left(\boldsymbol{Q}_{it}\right)\exp\{-\phi^{Q}\omega_{it}\}\right), \tag{2.7}$$

where $\widetilde{C}_{it}=C_{it}/P_{1,it}^{V}$, $\widetilde{\boldsymbol{P}}_{it}^{V}=(P_{2,it}^{V},\ldots,P_{J,it}^{V})'/P_{1,it}^{V}$.

In the next section, we discuss the identification of a multi-output production technology as well as the (relative) firm productivity from the variable cost function (2.7) under different specifications thereof. But first, a few remarks about the functional form of the firm's cost structure.

Naturally, restrictions put on the production technology, i.e., on the transformation function in (2.1), imply restrictions on the variable cost function in (2.6)–(2.7) because theoretical properties of $H(\cdot)$ follow from those of $T(\cdot)$. Hence, if one assumes a specific functional form for $T(\cdot)$, it will also have concrete implications for the functional form of $H(\cdot)$ as well. Under duality (see, e.g., Shephard, 1953; Diewert, 1971; Chambers, 1988), this is reversible, and by identifying parameters of the cost function $H(\cdot)$ we can indirectly recover—if not all, then at least economically important—parameters in $T(\cdot)$ as well as productivity ω_{it}. As McFadden (1978) puts it, due to being a value function the cost function is a "sufficient statistic" for all the economically relevant characteristics of the underlying production technology. Therefore, the assumption of a log-linear $H(\cdot)$ that we consider in Section 3 can be viewed as just that or, alternatively, this functional form can be justified as being implied by the Cobb-Douglas form of technology $T(\cdot)$.[5] For more flexible specifications such as, e.g., log-quadratic (see Appendix C), this functional-form "self-duality" may not hold and nor is it necessary in practice. Whatever the assumed functional form of $H(\cdot)$, under duality we can always recover the most relevant technological metrics such as returns to scale and firm productivity which are, arguably, the most commonly reported statistics in studies estimating production functions.[6]

Let us also concretize the measure of returns to scale, or scale elasticity, in a multi-product case. When outputs are many, the (long-run) returns to scale are defined by comparing a proportional change in all inputs to the resultant proportional change in all outputs (e.g., Färe and Primont, 1995): $\text{RTS} = \frac{\partial \ln \kappa}{\partial \ln \lambda}\big|_{\kappa=\lambda=1}$ for $T\left(\kappa \boldsymbol{Q}^*, \lambda \boldsymbol{K}, \lambda \boldsymbol{V}\right) = 1$. It is easy to show that $\text{RTS} = -\sum_{X\in\{\boldsymbol{K},\boldsymbol{V}\}} \frac{\partial \ln T(\cdot)}{\partial \ln X} \Big/ \sum_{d=1}^{D} \frac{\partial \ln T(\cdot)}{\partial \ln Q_d}$. Under duality, we can then recover RTS from the variable cost function as a reciprocal of economies of scale adjusted to account for quasi-fixed inputs (Panzar and Willig, 1977; Caves et al., 1981):

$$\text{RTS} = \frac{1-\sum_{X\in\boldsymbol{K}} \frac{\partial \ln C(\cdot)}{\partial \ln X}}{\sum_{d=1}^{D} \frac{\partial \ln C(\cdot)}{\partial \ln Q_d}} = \frac{1-\sum_{X\in\boldsymbol{K}} \frac{\partial \ln \widetilde{C}(\cdot)}{\partial \ln X}}{\sum_{d=1}^{D} \frac{\partial \ln \widetilde{C}(\cdot)}{\partial \ln Q_d}}, \tag{2.8}$$

where it is easy to see that $\partial \ln C/\partial \ln X = \partial \ln \widetilde{C}/\partial \ln X$ and $\partial \ln C/\partial \ln Q_d = \partial \ln \widetilde{C}/\partial \ln Q_d$.[7]

When it comes to measuring firm productivity, we can accomplish this using costs seamlessly because the cost function inherits ω_{it} from the transformation function via net producible outputs.

3 Identification Strategy

Using a dual approach, we model the short-run cost function to accommodate exogenous output quantity restrictions. To identify (2.7) from observational data is not straightforward because firm productivity ω_{it}, which scales each output ($Q^*_{d,it} = Q_{d,it}/\exp\{\omega_{it}\}\ \forall d$) as well as correlates with quasi-

[5] We demonstrate this "self-duality" of the Cobb-Douglas function in Appendix A.

[6] Other economically relevant characteristics of the underlying production technology that can be measured from the dual cost function include the Allen and Morishima elasticities of substitution, technical change measuring (secular) temporal shifts in technology, etc.

[7] The presence of cost elasticities of quasi-fixed inputs in (2.8) "adjusts" for the short-run nature of the cost function. If all inputs are flexible, then RTS simplifies to a straight reciprocal of economies of scale: $\left[\sum_{d=1}^{D} \frac{\partial \ln \widetilde{C}(\cdot)}{\partial \ln Q_d}\right]^{-1}$.

fixed inputs $\boldsymbol{K}_{it}$ and, potentially, the relative input prices $\widetilde{\boldsymbol{P}}^V_{it}$, is unobserved by an econometrician. Ignoring ω_{it} in the variable-cost-function regression is ill-advised because it would result in an endogeneity-inducing omitted confounder. We handle this problem along the lines of the Olley and Pakes (1996) and Levinsohn and Petrin (2003)-style proxy approach (developed in a production-function context) which we adapt to our cost-function setup.

Since the Cobb-Douglas remains a workhorse specification among practitioners in the proxy productivity literature (Olley and Pakes, 1996; Levinsohn and Petrin, 2003; Doraszelski and Jaumandreu, 2013; Ackerberg et al., 2015; Collard-Wexler and De Loecker, 2015; Konings and Vanormelingen, 2015; Kim et al., 2019; Malikov and Zhao, 2023), here we focus on the case when the variable cost function $H(\cdot)$ also takes the log-linear form. From (2.7), we have

$$\widetilde{c}_{it} = \alpha^0 + \sum_{j=2}^{J} \alpha^V_j \widetilde{p}^V_{j,it} + \sum_{m=1}^{M} \alpha^K_m k_{m,it} + \sum_{d=1}^{D} \alpha^Q_d q_{d,it} - \phi^Q \omega_{it}, \tag{3.1}$$

where lower-case variables denote the logs of respective variables in level, and ω_{it} is normalized to have zero mean.[8] Of note, ϕ^Q is not a "free" parameter but, in fact, $\phi^Q = \sum_{d=1}^{D} \alpha^Q_d$ since it is the degree of homogeneity of the output index $\exp\left\{\sum_{d=1}^{D} \alpha^Q_d q_{d,it}\right\}$. Further notice that the log-linearity of the Cobb-Douglas form delivers output separability (from other cost-function arguments) by default even if the latter had not been explicitly assumed. However, the practical relevance of this assumption becomes apparent under more flexible functional forms.

Making use of the Markovian nature of ω_{it}, we can further rewrite (3.1) as

$$\widetilde{c}_{it} = \alpha^0 + \sum_{j=2}^{J} \alpha^V_j \widetilde{p}^V_{j,it} + \sum_{m=1}^{M} \alpha^K_m k_{m,it} + \sum_{d=1}^{D} \alpha^Q_d q_{d,it} - \phi^Q \times g(\omega_{it-1}) - \zeta^*_{it}, \tag{3.2}$$

where $\zeta^*_{it} \equiv \phi^Q \zeta_{it}$ is an error. The parameters of interest are cost elasticities $\boldsymbol{\theta} = \left(\alpha^V_2, \ldots, \alpha^V_J, \alpha^K_1, \ldots, \alpha^K_M, \alpha^Q_1, \ldots, \alpha^Q_D\right)'$.

To identify $\boldsymbol{\theta}$ in the variable cost function (3.2), we need to address (i) the latency of firm productivity ω_{it-1} and (ii) the potential endogeneity (in an econometric meaning of the term) of relative input prices $\{\widetilde{p}^V_{j,it}\}$ with respect to the innovation ζ^*_{it}. The latter occurs because, per Assumption 4, cross-sectional input price dispersion may be driven by productivity differences across firms. Hence, given the potential non-zero contemporaneous association between $\{\widetilde{p}^V_{j,it}\}$ and ω_{it}, $\mathbb{E}\left[\zeta^*_{it}\widetilde{p}^V_{j,it}\right] = \phi^Q \mathbb{E}\left[\zeta_{it}\widetilde{p}^V_{j,it}\right] \neq 0\ \forall j$ since ω_{it} includes an update ζ_{it}. Note, however, that the dynamic inputs $(k_{1,it}, \ldots, k_{M,it})' \in \Xi_{it-1}$ and are thus uncorrelated with productivity innovation ζ_{it} at time t, and so is the case with (linearly entering) log-outputs per Assumption 3.

We identify (3.2) in two steps. First, to address the endogeneity of $\{\widetilde{p}^V_{j,it}\}$, we estimate their corresponding parameters $\boldsymbol{\alpha}^V = \left(\alpha^V_2, \ldots, \alpha^V_J\right)'$ indirectly by utilizing the structural link between the variable cost function and the cost-minimizing first-order conditions. Namely, applying Shephard's

[8] Latent productivity and the intercept term are not separately identifiable, with the implication being that ω_{it} can be identified up to a constant only.

lemma (in logs) to (3.1):

$$\alpha_j^V = \frac{\partial \widetilde{c}_{it}}{\partial \widetilde{p}_{j,it}^V} = \frac{\partial c_{it}}{\partial p_{j,it}^V} = \frac{P_{j,it}^V V_{j,it}}{C_{it}} \equiv S_{j,it}^V \quad \forall j = 2, \ldots, J, \tag{3.3}$$

where $S_{j,it}^V$ is a cost share of the jth variable input. Taking the mean of (3.3) for each static input, we arrive as a moment restriction that identifies $\boldsymbol{\alpha}^V$:

$$\boldsymbol{\alpha}^V = \mathbb{E}\begin{bmatrix} S_{2,it}^V \\ \vdots \\ S_{J,it}^V \end{bmatrix}. \tag{3.4}$$

To proceed to the second step, with $\boldsymbol{\alpha}^V$ already identified from the data, let us go back to (3.2) and rewrite it as

$$\underbrace{\widetilde{c}_{it} - \sum_{j=2}^{J} \alpha_j^V \widetilde{p}_{j,it}^V}_{\widetilde{\varkappa}_{it}} = \alpha^0 + \sum_{m=1}^{M} \alpha_m^K k_{m,it} + \sum_{d=1}^{D} \alpha_d^Q q_{d,it} - \phi^Q \times g(\omega_{it-1}) - \zeta_{it}^*, \tag{3.5}$$

where, after replacing each α_j^V with its first-step estimator $\mathbb{E}[S_{j,it}^V]$, $\widetilde{\varkappa}_{it}$ on the left-hand side effectively becomes observed data. This eliminates $\{\widetilde{p}_{j,it}^V\}$ from among regressors, and the regression in (3.5) now contains no endogenous covariates that need instrumentation. However, to operationalize (3.5), we still need to deal with latency of ω_{it-1}. In effect, we do so by quasi-differencing the variable cost function. Namely, to identify the rest of the variable cost function (3.5), we proxy for firm productivity ω_{it-1} by inverting (3.1) for firm productivity, whereby $\omega_{it} = \frac{1}{\phi^Q}\left[\alpha^0 + \sum_{j=2}^{J} \alpha_j^V \widetilde{p}_{j,it}^V + \sum_{m=1}^{M} \alpha_m^K k_{m,it} + \sum_{d=1}^{D} \alpha_d^Q q_{d,it} - \widetilde{c}_{it}\right]$. Using the first-step estimator of $\{\alpha_j^V\}$ in the proxy function, we thus obtain the following nonlinear (in parameters) equation from (3.5) by substituting for ω_{it-1}:

$$\widetilde{\varkappa}_{it} = \alpha^0 + \sum_{m=1}^{M} \alpha_m^K k_{m,it} + \sum_{d=1}^{D} \alpha_d^Q q_{d,it} - \phi^Q \times g\left(\frac{1}{\phi^Q}\left[\alpha^0 + \sum_{m=1}^{M} \alpha_m^K k_{m,it-1} + \sum_{d=1}^{D} \alpha_d^Q q_{d,it-1} - \widetilde{\varkappa}_{it-1}\right]\right) - \zeta_{it}^*, \tag{3.6}$$

where all right-hand-side regressors are mean-orthogonal to ζ_{it}^* and, thus, are self-instrumenting. With $g(\cdot)$ approximated using sieves, (3.6) identifies all remaining key cost-function parameters of interest, namely $\left(\alpha_1^K, \ldots, \alpha_M^K, \alpha_1^Q, \ldots, \alpha_D^Q\right)'$. The above equation is just the Markov process equation for firm productivity with ω_{it} substituted away using the inverted variable cost function, viz.,

$$\frac{1}{\phi^Q}\left[\alpha^0 + \sum_{m=1}^{M} \alpha_m^K k_{m,it} + \sum_{d=1}^{D} \alpha_d^Q q_{d,it} - \widetilde{\varkappa}_{it}\right] = g\left(\frac{1}{\phi^Q}\left[\alpha^0 + \sum_{m=1}^{M} \alpha_m^K k_{m,it-1} + \sum_{d=1}^{D} \alpha_d^Q q_{d,it-1} - \widetilde{\varkappa}_{it-1}\right]\right) + \zeta_{it}, \tag{3.7}$$

which is a standard nonlinear regression, albeit with a non-unit Jacobian of transformation when written in this form. Of note, the presence of neither a scaling multiplier ϕ^Q nor an intercept α^0 impede the identification of $\left(\alpha_1^K,\ldots,\alpha_M^K,\alpha_1^Q,\ldots,\alpha_D^Q\right)'$ because ϕ^Q is not a "free" parameter but, recall, is subject to the restriction that $\phi^Q=\sum_{d=1}^D \alpha_d^Q$ and α^0 is the only "intercept" because $g(\cdot)$ contains no additive constant per the mean-zero normalization.

Let's examine the point identification of these parameters from the second-step equation (3.6) a bit more formally. For clarity, consider a special case when $M=D=1$ and let $g(\cdot)$ be linear with no intercept: $g(\cdot)=\rho_1\times(\cdot)$. Then, (3.6) simplifies to

$$\left[\widetilde{\varkappa}_{it}-\alpha^K k_{it}-\alpha^Q q_{it}\right]=\rho_0+\rho_1\left[\widetilde{\varkappa}_{it-1}-\alpha^K k_{it-1}-\alpha^Q q_{it-1}\right]-\zeta_{it}^*, \tag{3.8}$$

where now there is only one output-related and one dynamic-input-related coefficient (i.e., $\alpha^Q=\alpha_1^Q$ and $\alpha^K=\alpha_1^K$), and $\rho_0\equiv\alpha^0(1-\rho_1)$ is a new intercept. The parameters of interest are $\boldsymbol{\alpha}^{V-}=\left(\alpha^K,\alpha^Q\right)'$; and also define the auxiliary $\boldsymbol{\rho}=(\rho_0,\rho_1)'$. Let us write equation (3.8) as $\widetilde{\varkappa}_{it}=r_{it}\left(\boldsymbol{\rho},\boldsymbol{\alpha}^{V-}\right)-\zeta_{it}^*$, where $r_{it}(\cdot)$ is the nonlinear regression function and ζ_{it}^* is a mean-independent error. Consider now the identification of $\left(\boldsymbol{\rho}',\left(\boldsymbol{\alpha}^{V-}\right)'\right)'$ in the following just-identified GMM problem that is equivalent to the nonlinear least-squares problem:

$$\left(\boldsymbol{\rho}_\star',\left(\boldsymbol{\alpha}_\star^{V-}\right)'\right)'=\arg\min_{\boldsymbol{\rho},\boldsymbol{\alpha}^{V-}}\ \mathbb{E}\left[\nabla r_{it}\left(\boldsymbol{\rho},\boldsymbol{\alpha}^{V-}\right)\left(\widetilde{\varkappa}_{it}-r_{it}\left(\boldsymbol{\rho},\boldsymbol{\alpha}^{V-}\right)\right)\right]'\mathbf{W}\mathbb{E}\left[\nabla r_{it}\left(\boldsymbol{\rho},\boldsymbol{\alpha}^{V-}\right)\left(\widetilde{\varkappa}_{it}-r_{it}\left(\boldsymbol{\rho},\boldsymbol{\alpha}^{V-}\right)\right)\right], \tag{3.9}$$

where $\mathbf{W}$ is a symmetric positive-definite moment-weighting matrix, and $\nabla r_{it}\left(\boldsymbol{\rho},\boldsymbol{\alpha}^{V-}\right)$ is the vector of gradients with respect to parameters:

$$\nabla r_{it}\left(\boldsymbol{\rho},\boldsymbol{\alpha}^{V-}\right)=\begin{bmatrix}1\\ \widetilde{\varkappa}_{it-1}-\alpha^K k_{it-1}-\alpha^Q q_{it-1}\\ k_{it}-\rho_1 k_{it-1}\\ q_{it}-\rho_1 q_{it-1}\end{bmatrix}. \tag{3.10}$$

The corresponding 4×4 information matrix $\Psi\left(\boldsymbol{\rho},\boldsymbol{\alpha}^{V-}\right)=\mathbb{E}\left[\nabla r_{it}\left(\boldsymbol{\rho},\boldsymbol{\alpha}^{V-}\right)\nabla r_{it}\left(\boldsymbol{\rho},\boldsymbol{\alpha}^{V-}\right)'\right]$ has a full rank (we unpack this expression in Appendix B). Thus, the information matrix for the GMM problem in (3.9) when evaluated at the true parameter values $\Psi\left(\boldsymbol{\rho}_\star,\boldsymbol{\alpha}_\star^{V-}\right)$ will be full-rank, and the parameters of interest in $\boldsymbol{\alpha}^{V-}$ will therefore be identified.

Remark 1 Although very popular in applied productivity studies, the Cobb-Douglas form is restrictive in that it assumes constant (over time and across firms) elasticities which forces firms of all sizes—small and large—to have the same technology, including the same returns to scale. We consider a more flexible (restricted) translog functional form for the firm's variable cost structure $H(\cdot)$ capable of accommodating technological heterogeneity across firms in Appendix C.

4 Estimation

The estimation procedure involves two steps and is easy to implement. In the first step, for each variable input j, we estimate a $\widetilde{p}_j^V$-related parameter using a sample analogue of (3.4):

$$\widehat{\alpha}_j^V = (NT)^{-1} \sum_{i,t} S_{j,it}^V \quad \forall j = 2, \ldots, J, \tag{4.1}$$

using which we can then construct a feasible analogue of $\widetilde{\varkappa}_{it}$; let us denote it as $\widetilde{\varkappa}_{it}\big(\widehat{\boldsymbol{\alpha}}^V\big) = \widetilde{c}_{it} - \sum_{j=2}^J \widehat{\alpha}_j^V \widetilde{p}_{j,it}^V$.

The second step requires choosing how to model the $g(\cdot)$ function. A popular choice in applied research involving proxy-variable estimations of production functions is just to assume that firm productivity is a linear autoregressive process (e.g., see Zhang, 2017, 2019; Kim et al., 2019; Ackerberg et al., 2020; Grieco et al., 2020; Mo et al., 2021; Malikov et al., 2024), as we have also done when discussing identification of the second step in Section 3. As such, we continue assuming that $g(\cdot)$ is a linear (parametric) function,[9] viz., $g(\cdot) = \rho_1 \times (\cdot)$, with no intercept to normalize firm productivity to be zero-mean. We can then estimate a proxied productivity process in (3.6) via nonlinear least squares (NLLS). The second-step estimator is

$$\begin{bmatrix} \widehat{\boldsymbol{\rho}} \\ \widehat{\boldsymbol{\alpha}}^{V-} \end{bmatrix} = \arg\min_{\boldsymbol{\rho}, \boldsymbol{\alpha}^{V-}} \sum_{i,t} \Big[w_{it}\big(\boldsymbol{\alpha}^{V-} \big| \widehat{\boldsymbol{\alpha}}^V\big) - \rho_0 - \rho_1 w_{it-1}\big(\boldsymbol{\alpha}^{V-} \big| \widehat{\boldsymbol{\alpha}}^V\big)\Big]^2 \tag{4.2}$$

$$\text{s.t.} \quad w_{it}\big(\boldsymbol{\alpha}^{V-} \big| \widehat{\boldsymbol{\alpha}}^V\big) \equiv \widetilde{\varkappa}_{it}\big(\widehat{\boldsymbol{\alpha}}^V\big) - \sum_{m=1}^M \alpha_m^K k_{m,it} - \sum_{d=1}^D \alpha_d^Q q_{d,it},$$

where, recall that $\rho_0 \equiv \alpha^0(1-\rho_1)$, and $\boldsymbol{\alpha}^{V-} = \big(\alpha_1^K, \ldots, \alpha_M^K, \alpha_1^Q, \ldots, \alpha_D^Q\big)'$ is redefined to denote all remaining non-variable-input-price-related cost-function slope parameters.

Under our model assumptions, such a two-step estimator is consistent and asymptotically normal. This is straightforward to establish by recasting both steps in a multiple-equation/system generalized method of moments (GMM) framework which, incidentally, also permits the derivation of an asymptotic variance matrix that accounts for a sequential nature of the estimation (see Newey, 1984). That is, we can rewrite our estimator as a simultaneous multiple-equation system of moment restrictions where the instrument sets vary across equations. Specifically, referring to all parameters in (4.1)–(4.2) collectively as $\boldsymbol{\theta} = \Big(\boldsymbol{\alpha}^{V\prime}, \big(\boldsymbol{\alpha}^{V-}\big)', \boldsymbol{\rho}'\Big)'$ and writing the fully expanded residual function in second-step estimator (after the substitution for $\widetilde{\varkappa}_{it}$) as

$$e_{it}(\boldsymbol{\theta}) = \widetilde{c}_{it} - \sum_{j=2}^J \alpha_j^V \widetilde{p}_{j,it}^V - \sum_{m=1}^M \alpha_m^K k_{m,it} - \sum_{d=1}^D \alpha_d^Q q_{d,it} - \rho_0 - \rho_1 \Bigg(\widetilde{c}_{it-1} - \sum_{j=2}^J \alpha_j^V \widetilde{p}_{j,it-1}^V - \sum_{m=1}^M \alpha_m^K k_{m,it-1} - \sum_{d=1}^D \alpha_d^Q q_{d,it-1} \Bigg),$$

[9] Alternatively, one may treat $g(\cdot)$ as a nonparametric function, in which case it can be approximated using sieves.

we can equivalently formulate both estimation steps using their implied moment conditions:

$$\mathbb{E}\left[\boldsymbol{f}_{it}(\boldsymbol{\theta})\right] \equiv \mathbb{E}\begin{bmatrix} \begin{pmatrix} S^V_{2,it} \\ \vdots \\ S^V_{J,it} \end{pmatrix} - \boldsymbol{\alpha}^V \\ e_{it}(\boldsymbol{\theta}) \dfrac{\partial e_{it}(\boldsymbol{\theta})}{\partial\left(\left(\boldsymbol{\alpha}^{V-}\right)', \boldsymbol{\rho}'\right)'} \end{bmatrix} = \mathbf{0}. \tag{4.3}$$

The benefit of interpreting our sequential estimator as solving a multiple-equation GMM problem corresponding to a system of nonlinear moment equations in (4.3) simultaneously is that the standard large-N limit results for a class of such GMM estimators apply here. Furthermore, using the moment equivalents in (4.3) also facilitates asymptotic inference based on the optimal variance of GMM estimators $\mathbb{V}[\widehat{\boldsymbol{\theta}}] = [\mathbb{E}\frac{\partial \boldsymbol{f}_{it}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'}]^{-1}\mathbb{E}[\boldsymbol{f}_{it}(\boldsymbol{\theta})\boldsymbol{f}_{it}(\boldsymbol{\theta})'][\mathbb{E}\frac{\partial \boldsymbol{f}_{it}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}]^{-1}$ which, if desired, can also be robustified using usual off-the-shelf methods. Bootstrap provides an alternative avenue for testing.

5 What if output quantities are flexibly chosen?

The methodology considered in our paper is specifically designed for market environments when optimizing firms face exogenous output quantity restrictions. Under Assumption 3 which formalizes this, we have that $\mathbb{E}\left[\zeta_{it}|q_{d,it}\right] = 0\ \forall d$, implying that outputs $\{q_{d,it}\}$ can self-instrument in step two[10] of our identification scheme given in (3.7). This orthogonality allows us to estimate the second-step equation (4.2) via NLLS.

A natural question to contemplate is whether our methodology is robust—and if not, whether it can be robustified—to the violation of the above assumption in situations when firms can flexibly choose their output quantities in a pursuit of the maximum profits. The latter is a standard production model setup that would render $\{q_{d,it}\}$ no longer mean-orthogonal to the contemporaneous productivity, conditional on period $t-1$ information. That is, $\mathbb{E}\left[\zeta_{it}|q_{d,it}\right] \neq 0\ \forall d$ when $\{q_{d,it}\}$ are optimally chosen by the firm at time t because supply decisions are based on information Ξ_{it} which also includes ω_{it} and which, in turn, incorporates a period-t productivity update ζ_{it}. Note that, in such a case, the cost function formulation that our methodology is based on still remains valid for it represents the first-order optimality. But the NLLS projection, which we have relied on under Assumption 3, obviously cannot identify step-two equations because log-outputs $\{q_{d,it}\}$ are endogenous. The solution, however, is readily available: step two can still be identified from the (relevant) exogenous variation in the remaining covariates, whereby

$$\mathbb{E}\left[\zeta_{it}\left(1, k_{1,it}, \ldots, k_{M,it}, k_{1,it-1}, \ldots, k_{M,it-1}, q_{1,it-1}, \ldots, q_{D,it-1}, \widetilde{\varkappa}_{it-1}\right)'\right] = 0, \tag{5.1}$$

which now excludes contemporaneous $\{q_{1,it}, \ldots, q_{D,it}\}$ from the instrumental variables set. The identification of step two via the moment above is possible because period-t outputs $\{q_{d,it}\}$ in equation

[10] Recall that step one of our methodology is unaffected by outputs.

(3.7) are not "free" regressors but enter the model subject to a parameter restriction whereby the coefficients thereon are the same as those on weakly exogenous, self-instrumenting lagged $\{q_{d,it-1}\}$, so there is no need for external instruments. Step-two equations are identified exactly based on (5.1), with the number of valid instruments being equal to the number of unknown parameters. Obviously, the NLLS estimation procedure in step two would have to be switched for nonlinear (G)MM. Specifically, instead of (4.2) one would use the following second-step NLIV estimator of $(\boldsymbol{\rho}, \boldsymbol{\alpha}^{V-})$:

$$\min_{\boldsymbol{\rho},\boldsymbol{\alpha}^{V-}} \Big[\sum_{i,t}\mathbf{X}_{it}\Big(w_{it}(\boldsymbol{\alpha}^{V-}|\widehat{\boldsymbol{\alpha}}^{V}) - \rho_0 - \rho_1 w_{it-1}(\boldsymbol{\alpha}^{V-}|\widehat{\boldsymbol{\alpha}}^{V})\Big)\Big]'\sum_{i,t}\mathbf{X}_{it}\Big(w_{it}(\boldsymbol{\alpha}^{V-}|\widehat{\boldsymbol{\alpha}}^{V}) - \rho_0 - \rho_1 w_{it-1}(\boldsymbol{\alpha}^{V-}|\widehat{\boldsymbol{\alpha}}^{V})\Big)$$

$$\text{s.t.}\quad w_{it}(\boldsymbol{\alpha}^{V-}|\widehat{\boldsymbol{\alpha}}^{V}) \equiv \widetilde{\varkappa}_{it}(\widehat{\boldsymbol{\alpha}}^{V}) - \sum_{m=1}^{M}\alpha_m^K k_{m,it} - \sum_{d=1}^{D}\alpha_d^Q q_{d,it},$$

where $\mathbf{X}_{it}$ is the vector of instruments from (5.1).

The moment-based estimator above is also capable of addressing an "intermediate" case. For instance, certain firms may initially need to fulfill previously signed long-term contracts, producing and supplying contracted quantities of outputs, which can be regarded as econometrically exogenous. However, beyond these commitments, firms can then produce additional outputs based on market conditions to maximize their profits. In such a scenario of partially flexible outputs, firms still solve the same cost minimization problem as the one in (2.5), and therefore, the modified moment-based estimator remains valid.

A point worth emphasizing is that—because it does not impose orthogonality between the productivity innovation and contemporaneous output quantities—the above moment-based estimator is consistent regardless if the outputs are restricted *or* flexible. Whereas, the NLLS estimator is consistent *only* if outputs are restricted, albeit it is efficient in that case unlike the NLIV. As such, we can use these two alternative step-two estimators to formally discriminate between the cases of (endogenous) flexible outputs and (exogenous) restricted outputs via the standard Durbin-Wu-Hausman test. Specifically, denoting the NLLS and NLIV parameter estimators of the step-two equation as $\widehat{\boldsymbol{\theta}}$ and $\widetilde{\boldsymbol{\theta}}$, respectively, the Durbin-Wu-Hausman test statistic can be constructed as

$$\mathscr{H} = (\widehat{\boldsymbol{\theta}} - \widetilde{\boldsymbol{\theta}})' \left[\mathbb{V}[\widetilde{\boldsymbol{\theta}}] - \mathbb{V}[\widehat{\boldsymbol{\theta}}]\right]^{-1} (\widehat{\boldsymbol{\theta}} - \widetilde{\boldsymbol{\theta}}), \tag{5.2}$$

where the optimal variances are obtained as discussed in Section 4. The test statistic is asymptotically chi-squared distributed under the null of exogenously restricted output quantities, with the number of degrees of freedom equal to the rank of $\mathbb{V}[\widetilde{\boldsymbol{\theta}}] - \mathbb{V}[\widehat{\boldsymbol{\theta}}]$. We apply this test in our empirical application to provide formal statistical corroboration for our arguments that, in the U.S. electricity generation sector, power plants' outputs (power generation) is exogenously restricted in the short run.

6 Simulations

Before taking our methodology to data, we first demonstrate its ability to successfully identify the cost-function parameters in a small Monte Carlo study. In designing the data generating process we draw from Grieco et al. (2016), Gandhi et al. (2020) and Malikov and Zhao (2023). We consider a balanced panel of $N = \{200,400,600,800\}$ firms observed over $T = 10$ periods.[11] Each panel is simulated $R = 1{,}000$ times. To keep things simple, we consider the case of a self-dual Cobb-Douglas technology with a single output ($D = 1$), one quasi-fixed input ($M = 1$) and two variable inputs ($J = 2$). The output-separable transformation function is given by

$$[Q_{it}\exp\{-\omega_{it}\}]^{-\beta^Q} K_{it}^{\beta^K} V_{1,it}^{\beta_1^V} V_{2,it}^{\beta_2^V} = 1, \tag{6.1}$$

where we normalize $\beta^Q = 1$ and set $\beta^K = 0.25$, $\beta_1^V = 0.3$ and $\beta_2^V = 0.4$. It is homogeneous of degree $-\beta^Q$, and the production technology exhibits decreasing returns to scale: $\text{RTS} = -\left(\beta^K + \beta_1^V + \beta_2^V\right)/\left(-\beta^Q\right) = 0.95$.

Latent firm productivity is generated as an exogenous zero-mean linear AR(1) process $\omega_{it} = \rho_1\omega_{it-1} + \zeta_{it}$ with $\rho_1 = 0.7$ and $\zeta_{it} \sim i.i.d.\ \mathbb{N}(0,0.04)$. Variable input prices (in logs) are allowed to vary across firms/time and to be contemporaneously correlated with firm productivity via $p_{1,it}^V = \nu_{10}p_{1,it-1}^V + \nu_{11}\omega_{it} + \eta_{1,it}^V$ and $p_{2,it}^V = \nu_{20}p_{2,it-1}^V + \nu_{21}\omega_{it} + \eta_{2,it}^V$, where we set $\nu_{10} = \nu_{20} = 0.55$, $\nu_{11} = 0.3$, $\nu_{21} = 0.1$ and both $\eta_{1,it}^V$ and $\eta_{2,it}^V$ are drawn independently from $\mathbb{N}(0,0.07)$.

For exogenously predetermined/restricted output quantity that firms target, we consider two alternative log-linear AR(1) specifications,

$$q_{it} = \tau_0 + \tau_1 q_{it-1} + \tau_2\omega_{it-1} + \epsilon_{it}, \tag{6.2}$$

with $\tau_0 = 1$, $\tau_1 = 0.8$ and $\epsilon_{it} \sim i.i.d.\ \mathbb{N}(0,0.1)$: when [**a**] $\tau_2 = 0$ which implies a purely exogenous evolution and [**b**] $\tau_2 = 0.2$ wherewith output constraints positively correlate with productivity.

We generate the firm's input allocations as follows. A quasi-fixed dynamic input—think, physical capital—evolves according to $K_{it} = I_{it-1} + (1-\delta_i)K_{it-1}$, where firm-specific depreciation rates are drawn from $\delta_i \sim i.i.d.\ \mathbb{U}\{0.05, 0.075, 0.10, 0.125, 0.15\}$ and the investment function (in logs) is given by $\ln I_{it-1} = \theta_0\mathbb{E}[q_{it}|\Xi_{it-1}] + \theta_1 k_{it-1} + \theta_2\omega_{it-1}$, where $\theta_0 = 0.5$, $\theta_1 = 0.4$, $\theta_2 = 0.1$, and $\mathbb{E}[q_{it}|\Xi_{it-1}] = \tau_0 + \tau_1 q_{it-1}$ is the expected log-output. The allocations of flexible inputs are generated from the firm's static variable cost minimization problem (2.5) yielding the following (conditional) variable input demand functions:

$$V_{1,it} = \left[Q_{it}^*\right]^{\frac{\beta^Q}{\beta_1^V+\beta_2^V}} K_{it}^{\frac{-\beta^K}{\beta_1^V+\beta_2^V}} \left[\widetilde{P}_{2,it}^V \frac{\beta_1^V}{\beta_2^V}\right]^{\frac{\beta_2^V}{\beta_1^V+\beta_2^V}}, \tag{6.3}$$

[11] We have also experimented with shorter and longer panels, and the results remained qualitatively unchanged.

$$V_{2,it} = \left[Q^*_{it}\right]^{\frac{\beta^Q}{\beta_1^V+\beta_2^V}} K_{it}^{\frac{-\beta^K}{\beta_1^V+\beta_2^V}} \left[\frac{1}{\widetilde{P}^V_{2,it}} \frac{\beta_2^V}{\beta_1^V}\right]^{\frac{\beta_1^V}{\beta_1^V+\beta_2^V}}, \tag{6.4}$$

where, recall, that $Q^*_{it} = Q_{it}\exp\{-\omega_{it}\}$. The state variables are initialized as follows: $\omega_{0i} \sim i.i.d.\ \mathbb{U}(1,3)$, $q_{i0} \sim i.i.d.\ \mathbb{U}(2,5)$, $k_{i0} \sim i.i.d.\ \mathbb{U}(10,200)$, $p^V_{1,i0} \sim i.i.d.\ \mathbb{U}(0,1)$, $p^V_{2,i0} \sim i.i.d.\ \mathbb{U}(0,0.5)$.

Just like the underlying technology, the implied variable cost function is also Cobb-Douglas:

$$\widetilde{c}_{it} = \underbrace{\ln\left[\left(\frac{\beta_1^V}{\beta_2^V}\right)^{\frac{\beta_2^V}{\beta_1^V+\beta_2^V}}\left(1+\frac{\beta_2^V}{\beta_1^V}\right)\right]}_{\alpha^0} + \underbrace{\left(\frac{\beta_2^V}{\beta_1^V+\beta_2^V}\right)}_{\alpha_2^V}\widetilde{p}^V_{2,it} + \underbrace{\left(\frac{-\beta^K}{\beta_1^V+\beta_2^V}\right)}_{\alpha^K} k_{it} + \underbrace{\left(\frac{\beta^Q}{\beta_1^V+\beta_2^V}\right)}_{\alpha^Q} q_{it} - \underbrace{\left(\frac{\beta^Q}{\beta_1^V+\beta_2^V}\right)}_{\phi^Q}\omega_{it}, \tag{6.5}$$

from where we obtain that the "true" values of the cost-function parameters $(\alpha^0, \alpha_2^V, \alpha^K, \alpha^Q)'$, and note that $\phi^Q = \alpha^Q$. From here, we can also confirm that the cost-based dual measure of scale elasticity $\text{RTS} = \left(1-\alpha^K\right)/\phi^Q = \left(\beta^K+\beta_1^V+\beta_2^V\right)/\beta_Q$ equals the corresponding primal measure.

First, we evaluate the performance of our proposed cost-based NLLS estimator with the focus on its ability to successfully identify key technological parameters. We estimate our model via the two-step algorithm outlined in Section 4. For each simulation repetition, we obtain point estimates of the cost-function-function parameters and the corresponding RTS measure for which we report the root mean squared error (RMSE) computed over R simulations. Panel A of Table 1 summarizes the second-step results for both exogenous and ω-controlled DGPs for output.[12] The results are encouraging and show that our methodology recovers the true parameters well, lending strong support to the validity of our identification strategy. As expected of a consistent estimator, the estimation becomes more stable as N grows.

In Panel B of Table 1, we also assess the performance of our NLIV estimator of step two detailed in Section 5 when outputs are exogenously determined. Like the NLLS estimator, the moment-based estimator exhibits convergence in the MSE as the sample size increases, indicating its consistency. Expectedly, RMSE values of the NLIV estimator are larger than their counterparts for the NLLS estimator, because that the latter is more efficient in this case.

To demonstrate the usefulness of our methodology in settings when firms face output constraints, we inspect the performance of a standard production-function proxy estimator which ignores the output constraint and—to the contrary—like most such estimators, uses the structure implied by the firm's static optimization problem in which output quantity can be flexibly chosen to maximize profits. Given recent unidentification critiques of most off-the-shelf proxy-variable techniques including Olley and Pakes (1996) and Levinsohn and Petrin (2003), we employ the Gandhi et al. (2020) system-of-equations method which we implement parametrically along the lines of Malikov and Zhao (2023) and Zhang and Malikov (2023). Concretely, we are interested in applying the primal production-

[12] We omit the first-step results corresponding to α_2^V because, per (3.4) and given our DGP, it is estimated exactly.

TABLE 1. Simulation Results for the DGP with Restricted Q_t

	$N = 200$	$N = 400$	$N = 600$	$N = 800$
PANEL A: SECOND STEP VIA NLLS WITH Q_t SELF-INSTRUMENTING				
Exogenous Output Process				
α^K	0.0474	0.0316	0.0273	0.0237
α^Q	0.0207	0.0142	0.0113	0.0098
RTS	0.0340	0.0229	0.0198	0.0169
Productivity-Dependent Output Process				
α^K	0.0438	0.0291	0.0255	0.0216
α^Q	0.0207	0.0142	0.0112	0.0098
RTS	0.0314	0.0212	0.0185	0.0156
PANEL B: SECOND STEP VIA NLIV EXCLUDING Q_t FROM IVS				
Exogenous Output Process				
α^K	0.0679	0.0453	0.0372	0.0325
α^Q	0.1520	0.1044	0.0820	0.0694
RTS	0.0797	0.0525	0.0423	0.0346
Productivity-Dependent Output Process				
α^K	0.0637	0.0440	0.0382	0.0331
α^Q	0.1885	0.1431	0.1221	0.1003
RTS	0.0852	0.0646	0.0549	0.0450

Reported are simulation RMSE for the dual cost-function parameters from (6.5) estimated via two alternative step-two estimators. Step-one results corresponding to α_2^V are omitted because, given the DGP, it is estimated exactly. RTS$= (1-\alpha^K)/\alpha^Q$. $T = 10$ throughout.

function estimator on the Cobb-Douglas technology (with the $\beta^Q = 1$ normalization imposed):

$$Q_{it} = K_{it}^{\beta^K} V_{1,it}^{\beta_1^V} V_{2,it}^{\beta_2^V} \exp\{\omega_{it}\} = K_{it}^{\beta^K} V_{1,it}^{\beta_1^V} V_{2,it}^{\beta_2^V} \exp\{g(\omega_{it-1}) + \zeta_{it}\}. \tag{6.6}$$

Such an estimator addresses endogeneity of flexible inputs (with respect to the contemporaneous productivity innovation ζ_{it}) by relying on the structural relation implied by the firm's static profit maximization under perfect competition in both output and input markets, viz., $\max_{\{V_{1,it}, V_{2,it}, Q_{it}\}} P_{it}^Q Q_{it} - \sum_{j=1} P_{j,it}^V V_{j,it}$ subject to (6.6). Namely, from the corresponding first-order conditions—provided the objective is correctly formulated—we have that the jth input's elasticity equals the ratio of its expenditures to revenues, which suggests the following first-step estimator: $\beta_j^V = \mathbb{E}\left[P_{j,it}^V V_{j,it} / \left(P_{it}^Q Q_{it}\right)\right] \; \forall j = 1,2$, where P_{it}^Q is the (presumably observable) output price. Such a first step would recover parameters on endogenous variable inputs, with the second step amounting to estimation of the remaining production-function parameters by running nonlinear least squares on $\left(\widetilde{q}_{it} - \beta^K k_{it}\right) = g\left(\widetilde{q}_{it-1} - \beta^K k_{it-1}\right) + \zeta_{it}$, where $\widetilde{q}_{it} = q_{it} - \beta_1^V v_{1,it} - \beta_{2,it}^V v_{2,it}$ is constructed using first-step estimates of β_1^V and β_2^V.[13]

[13] For comparability with our proposed estimator, here we also assume a linear $g(\cdot)$ function. Further, to minimize distorting influences of output price variation, which plays no role in our cost-based approach but must be observed to implement the first step of production-function estimator, we set it to be time-invariant and constant across all firms as

TABLE 2. Simulation Results for Primal Production-Function Estimator

	$N = 50$	$N = 100$	$N = 200$	$N = 400$
Exogenous Output Process				
β^K	0.1752	0.1308	0.1139	0.1036
β_1^V	0.1926	0.2020	0.2094	0.2156
β_2^V	0.2568	0.2693	0.2792	0.2874
α^K	0.8259	0.6963	0.6292	0.5463
α^Q	2.6925	3.0649	3.4236	3.7634
RTS	0.5300	0.5462	0.5701	0.5930
Productivity-Dependent Output Process				
β^K	0.4280	0.4484	0.4565	0.4650
β_1^V	0.1773	0.1878	0.1963	0.2033
β_2^V	0.2363	0.2504	0.2617	0.2711
α^K	1.0556	1.1705	1.2577	1.3486
α^Q	2.1727	2.4963	2.8096	3.1065
RTS	0.8345	0.8836	0.9129	0.9386

Reported are simulation RMSE for the primal technology coefficients estimated via the production-function estimator and for the cost-function parameters they imply. Under the $\beta^Q = 1$ normalization: $\alpha^K = -\beta^K/(\beta_1^V + \beta_2^V)$, $\alpha^Q = 1/(\beta_1^V + \beta_2^V)$ and RTS$=(1-\alpha^K)/\alpha^Q = (\beta^K + \beta_1^V + \beta_2^V)$. $T = 10$ throughout.

Table 2 summarizes simulation results for this production-function estimator under the restricted-output DGPs. Besides the primal input elasticities of output (β), we also include the results for the elasticities of cost (α) that they imply for an easier comparison across the estimators. Unsurprisingly, the estimation of variable-input elasticities is inconsistent in both scenarios, since the first-step estimator is derived under a misspecified optimization objective for firms. When output evolves exogenously (top panel), we observe that a production-function estimator can still consistently recover the coefficient on fixed input. The latter occurs because the bias in quasi-differenced $\widetilde{q}_{it}$ and $\widetilde{q}_{it-1}$, which depends on q_{it}, is uncorrelated with capital given the exogenous evolution of the former, thereby not impeding the identification of β^K. This no longer occurs when q_{it} is productivity-dependent (bottom panel) which results in a correlation with capital that is also productivity-dependent. In this case, all input elasticity parameters and cost elasticities that they imply are unidentified, as is the RTS measure. This showcases the inappropriateness of using standard structural methods for production-function estimation when outputs are restricted.

We conclude by evaluating the finite-sample performance of our proposed NLLS and NLIV estimators in a situation that firms flexibly choose their output quantities. In this case, we expect our NLLS estimator to no longer be consistent due to the endogeneity of q_{it}, but our NLIV estimator should still perform well. We adjust the DGP by dispensing with the exogenous output process

$P^Q = \max(P_j^V)$ so that the only source of variation in revenues is the variation in output quantity.

TABLE 3. Simulation Results for the DGP with Flexible Q_t

	$N=200$	$N=400$	$N=600$	$N=800$
PANEL A: SECOND STEP VIA NLLS WITH Q_t SELF-INSTRUMENTING				
α^K	0.2261	0.2265	0.2268	0.2269
α^Q	0.3046	0.3044	0.3043	0.3044
RTS	0.0568	0.0560	0.0555	0.0556
PANEL B: SECOND STEP VIA NLIV EXCLUDING Q_t FROM IVS				
α^K	0.0745	0.0517	0.0398	0.0345
α^Q	0.0887	0.0620	0.0481	0.0422
RTS	0.0132	0.0092	0.0074	0.0065

Reported are simulation RMSE for the dual cost-function parameters from (6.5) estimated via two alternative step-two estimators. Step-one results corresponding to α_2^V are omitted because, given the DGP, it is estimated exactly. RTS= $(1-\alpha^K)/\alpha^Q$. $T=10$ throughout.

in (6.2) and, instead, generate Q_{it} from the production function in (6.1) using the optimal flexible outputs $V_{1,it}$ and $V_{2,it}$ that solve the first-order conditions of a profit-maximizing firm with flexible production:

$$V_{1,it} = \left(\beta_1^V \left(\frac{\beta_2^V}{\beta_1^V}\right)^{\beta_2^V} \frac{(P_{1,it}^V)^{\beta_2^V-1}}{(P_{2,it}^V)^{\beta_2^V}} P_{it}^Q K_{it}^{\beta^K} \exp\{\omega_{it}\}\right)^{\frac{1}{1-\beta_1^V-\beta_2^V}}$$

$$V_{2,it} = \left(\beta_2^V \left(\frac{\beta_1^V}{\beta_2^V}\right)^{\beta_1^V} \frac{(P_{2,it}^V)^{\beta_1^V-1}}{(P_{1,it}^V)^{\beta_1^V}} P_{it}^Q K_{it}^{\beta^K} \exp\{\omega_{it}\}\right)^{\frac{1}{1-\beta_1^V-\beta_2^V}}.$$

All other variables are generated using the same procedures as discussed previously. For simplicity, the output price is fixed to be time-invariant and common for all firms as $P^Q = \max_j(P_{j,it}^V)$.

The corresponding results are reported in Table 3. We observe that our NLLS estimator no long exhibits the characteristics of a consistent estimator; the RMSE values are large, and there is no convergence as the sample size increases. Conversely, the NLIV estimator continues to perform well, displaying a small and decreasing RMSE for all parameters.

7 Empirical Illustration

We apply our methodology to a panel of U.S. natural gas-fired electric power plants in 2005–2017. The electricity generation sector in the U.S. relies on natural gas for a plurality of its energy needs (40% in 2022, EIA 2023), which underscores our interest in gas-fired power. Peaker gas plants play a particularly crucial role in smoothing supply shocks stemming from non-dispatchable energy resources like wind and solar. More recently, liquefied natural gas have become strategically important too, in the wake of the Russian invasion of Ukraine in 2022 and international sanctions that ensued.

Given growing U.S. (and global) reliance on gas-based electricity, a production analysis of such power plants becomes ever more timely. We indirectly recover their production technology and productivity by estimating a short-run variable cost function.

A unique feature of electric power generation is that the demand for electricity is, effectively, perfectly price-inelastic in the short run (see Borenstein et al., 2002; Borenstein and Holland, 2005). When someone flips on the light switch, they do not first check the price of electricity. Power supply is thus constrained to match demand. Blackouts and brownouts are the result of under-supply and very costly,[14] while over-supply also leads to waste because of the lack of significant storage capacity. As such, a utility industry such as electricity generation provides an especially apt setting for illustrating practical usefulness of our methodology because output quantities that power plants produce are exogenously restricted and not freely adjustable (in a pursuit of maximum profits) in the short run. Power plants also do not enjoy much market power over key variable inputs aligning well with the perfectly competitive factor markets assumption of our methodology. The U.S. benefits from cheap and abundant gas supplied by many firms, and is currently the world's largest producer. Deregulation and market liberalization have also helped foster a natural gas market with relatively few barriers to entry, and a suite of financial products, such as hub index prices, make for highly competitive and liquid markets.

7.1 Data

The main data source for our analysis is the FERC Form 1: Electric Utility Annual Reports from the U.S. Federal Energy Regulatory Commission. This dataset provides plant-level observations for major steam-electric generating facilities in the U.S. that utilize natural gas as their primary fuel input. To model a short-run variable cost function of these electric power plants, we specify two outputs, three variable input prices and one quasi-fixed input: $D = 2$, $J = 3$ and $M = 1$.

The first output (Q_1) is the net power generation excluding the plant usage measured in kilowatt-hours (kWh), and the second output (Q_2) are plant-hours connected to the load. Net power generation is, naturally, the most common output of electricity plants considered the literature (e.g., Nerlove, 1963; Christensen and Greene, 1976; Considine, 2000; Bernstein and Parmeter, 2019). A closely related measure of power plant production are the plant-hours connected to the load, which refer to the number of hours per year that a generator produces electricity used by external consumers. Plant-hours connected to the load thus measures supply duration over the course of a year and is related to both reliability and the extent to which the plant is used for peaking or base load generation.

The three variable inputs are the heat oil in barrels (V_1), natural gas in thousands of cubic feet (V_2) and number of employees (V_3). The (average) price of labor (P_3^V) is plant-level and calculated as production expenses, including operating costs, maintenance supervision and engineering costs,

[14] Allcott et al. (2016) quantify the impacts of under-supply in India, finding that electricity shortages reduce manufacturing revenues by 5.6–7.7%, and drop plant producer surplus by 9.5%, highlighting the essential nature of reliability in electricity supply.

TABLE 4. Summary Statistics

	Mean	S.D.	1st Qu.	Median	3rd Qu.
Net generation (Q_1) $\times 10^{-3}$	1,020,174	2,115,881	32,435	186,662	1,123,659
Plant hours (Q_2)	4,235	6,347	416	2,101	6,668
Installed capacity (K)	507	518	154	340	644
Price of heat oil (P_1^V)	2.23	0.62	1.72	2.21	2.87
Price of natural gas (P_2^V)	9.76	2.49	7.87	9.10	11.6
Price of labor (P_3^V)	36,796	87,782	5,483	17,909	40,760
Total variable cost (C)	83,358,560	177,945,300	3,670,536	19,035,410	93,719,730

divided by the number of employees. Price information for oil and gas inputs are obtained from secondary sources. The price of heat oil (P_1^V) is for the New York Harbor No. 2 heating oil FOB spot price which, given the integration of oil markets, is a national-level series. We obtain this series from the U.S. Energy Information Agency (EIA). The price of natural gas (P_2^V) is the EIA average commercial price series which is reported at the state level. All nominal prices are deflated using a state-level implicit GDP deflator and expressed in the 2012 U.S. dollar, with the resultant variation in all real input prices being at least state-level and, in case of labor, plant-level. Lastly, the total installed capacity measured in megawatts per hour is treated as a quasi-fixed input (K). A plant's variable cost (C) is constructed as the real total cost of heat oil, natural gas and labor.

Table 4 provides summary statistics of the data. Our sample is an unbalanced panel of 351 plants operating between 2005 and 2017, with a total of 2,641 observations. For all variables, the mean is larger than the median, pointing to their distributions being skewed to the right. Both standard deviations and interquartile ranges of input prices indicate significant variation in our sample. This large variation, especially for the price of labor for which the third quartile is about 7.4 times its first quartile, suggests that conventional proxy-variable methods that assume homogeneous input prices across firms would be inappropriate for the analysis here.

7.2 Results

We begin our analysis by conducting the Durbin-Wu-Hausman test of the endogeneity of output quantities produced by power plants. Following the test procedure outlined in Section 5, we calculate the test statistic, yielding a value of 1.33 with the corresponding p-value of 0.93 (df=5). Consequently, we fail to reject the null hypothesis that the output quantities are exogenously restricted. The test result further confirms our previous arguments that power supply is constrained to match demand in the short run. Our treatment of power generation as exogenous also aligns with previous studies, such as Borenstein et al. (2002) and Borenstein and Holland (2005). Given the test result and insights from previous studies, for the remainder of this section, we implement step two of our cost-function estimator using NLLS.

Panel A of Table 5 reports two-step parameter estimates of the Cobb-Douglas short-run vari-

TABLE 5. Technology Estimates

PANEL A: *Dual Cost Approach*	$\widetilde{c}$	PANEL B: *Primal Approach*	q_1	q_2
$\widetilde{p}_1^V$		v_1	0.027	–0.013
			(0.027)	(0.061)
$\widetilde{p}_2^V$	0.943⋆⋆⋆	v_2	0.652⋆⋆⋆	0.181
	(0.003)		(0.233)	(0.345)
$\widetilde{p}_3^V$	0.052⋆⋆⋆	v_3	0.307⋆⋆⋆	0.362⋆⋆⋆
	(0.003)		(0.041)	(0.068)
k	0.010	k	0.665	–0.365
	(0.586)		(0.443)	(0.658)
q_1	0.738⋆⋆⋆	q_1		0.399⋆⋆⋆
	(0.262)			(0.102)
q_2	0.405⋆	q_2	0.173⋆⋆⋆	
	(0.229)		(0.033)	
RTS	0.866⋆⋆⋆	RTS	1.997⋆⋆⋆	0.276
	(0.157)		(0.321)	(0.673)

Reported are the elasticity parameter estimates from both steps of our cost-function proxy estimator in (4.1)–(4.2) [Panel A] and one-step parameter estimates for the DPSW-style "unsymmetric" transformation functions in (7.1)–(7.2) [Panel B]. Asymptotic standard errors in parentheses. ⋆p<0.1; ⋆⋆p<0.05; ⋆⋆⋆p <0.01.

able cost function for electric power plants (normalized using the price of heating oil input $P^V_{1,it}$).[15] Among flexible inputs, we find that the electric power generation is most cost-sensitive with respect to natural gas prices. This is unsurprising given that our analysis focuses specifically on natural gas fired power plants for which gas is the primary intermediate input. At the same time, given the statistical insignificance of generation capacity,[16] it appears that plants' short-run costs are near their optimal "long-run" levels. The cost elasticities of outputs are significant and, together, they imply RTS of $0.866 = (1-0.010)/(0.738+0.405)$. However, this estimate is not statistically different from 1, suggesting constant returns to scale in the industry. In "stochastic frontier" studies that utilize similar data but different methods, Bernstein (2020) also reports evidence of unitary returns to scale in gas-fired power generation between 1994 and 2016 as does Knittel (2002) between 1981 and 1996.

To underscore the importance of incorporating output constraints that render $\boldsymbol{Q}_{it}$ a firm's state—not control—variable in the production-technology estimation, aside from our *dual* cost-based methodology that explicitly incorporates this key structural detail, we also estimate a more traditional *primal* production-function model. Since there is more than one output in our case ($D = 2$), we implement the latter via the Dhyne et al. (2022) method (hereafter, DPSW) that extends single-product production-function proxy estimators to multi-output settings using "unsymmetric" transformation functions

[15] Asymptotic standard errors are reported in parentheses. To account for two-step estimation, the standard errors are calculated by evaluating the asymptotic variance of the GMM estimator corresponding to (4.3) at our two-step estimates as detailed in Section 4.

[16] A positive sign of the cost elasticity with respect to k is of no concern since it is estimated to be statistically no different from zero, which satisfies a theoretical property that the variable cost is non-increasing in fixed inputs in population.

from Diewert (1973). To maximize comparability with our approach, we also implement it using a Cobb-Douglas functional form. Concretely, depending on the choice of a left-hand-side output, the two such unsymmetric transformation functions are $Q_{1,it} = K_{it}^{\beta^K}\left[\Pi_{j=1}^{3} V_{j,it}^{\beta_j^V}\right] Q_{2,it}^{\beta^Q}\exp\{\beta^0+\omega_{1,it}\}$ and $Q_{2,it} = K_{it}^{\gamma^K}\left[\Pi_{j=1}^{3} V_{j,it}^{\gamma_j^V}\right] Q_{1,it}^{\gamma^Q}\exp\{\gamma^0+\omega_{2,it}\}$, from where it is also evident that, unlike our single-equation approach, the DPSW methodology is multiple-equation (one per output) and assumes that the first-order Markov productivity is product-specific: $\omega_{1,it}$ and $\omega_{2,it}$, respectively.

Per DPSW, these productivities evolve following $\omega_{1,it} = g_1(\omega_{it-1})+\zeta_{1,it}$ and $\omega_{2,it} = g_2(\omega_{it-1})+\zeta_{2,it}$, whereby each one of them autoregressively depend on the vector of firm's all output-specific productivities $\omega_{it-1} = (\omega_{1,it-1}, \omega_{2,it-1})'$. To overcome the apparent violation of a "scalar unobservability" condition underlying the proxy approach, DPSW require multiple proxies: at least one per unobservable. Using the two energy inputs (v_1 and v_2) as such proxy variables to control for both product-specific productivities, we have

$$q_{1,it} = \beta^0 + \beta^K k_{it} + \sum_{j=1}^{3} \beta_j^V v_{j,it} + \beta^Q q_{2,it} + \lambda_1\left(k_{it-1}, v_{1,it-1}, v_{2,it-1}\right) + \zeta_{1,it}, \tag{7.1}$$

$$q_{2,it} = \gamma^0 + \gamma^K k_{it} + \sum_{j=1}^{3} \gamma_j^V v_{j,it} + \gamma^Q q_{1,it} + \lambda_2\left(k_{it-1}, v_{1,it-1}, v_{2,it-1}\right) + \zeta_{2,it}, \tag{7.2}$$

where we have made use of Markovian structure of productivities and substitutes bivariate controls for $\omega_{1,it}$ and $\omega_{2,it}$. Following DPSW, (7.1)–(7.2) are estimated via GMM[17] using lagged labor $v_{3,it-1}$ and twice lagged energy inputs $(v_{1,it-2}, v_{2,it-2})'$ as instruments for $v_{3,it}$ and $(v_{1,it}, v_{2,it})'$, respectively, and lagged outputs ($q_{2,it-1}$ or $q_{1,it-1}$) as instruments for period-t outputs on the right-hand side ($q_{2,it}$ or $q_{1,it}$). Quasi-fixed capital k_{it} requires no instrumentation, as before.

We summarize technology parameter estimates for these DPSW-style transformation functions in Panel B of Table 5. Given the "unsymmetry" of such primal formulations of the multi-product technology, the coefficients corresponding to the same inputs can vary a lot across models, depending on the choice of a left-hand-side output. But of note, only the labor input (v_3) has a consistently significant elasticity of output. An input as critical to power generation as the natural gas (v_2) is significant only when the net generation output is used as a regressand. In both equations, the capital elasticity is also estimated very imprecisely. The cross-output coefficient estimates are positive suggesting—in line with a priori intuition—that net generation and the number of hours that a generator produces externally consumed electricity are complements.

Although the primal-approach and our cost-based parameter estimates are not directly comparable, we can meaningfully contrast the two methodologies by looking at the returns to scale that they imply. Depending on the normalizing output, the primal estimate of RTS ranges between 0.276 and 1.997, suggesting massive, either diseconomies or economies of scale. The inherent non-uniqueness of "unsymmetric" primal formulations, which the DPSW approach builds on, alone is unlikely to fully explain this implausible variation in the implied scale elasticity in power generating plants. Instead,

[17] After approximating $\lambda_m(\cdot)$ functions using quadratic polynomials.

we attribute it to the primal approach's underlying misconception of the production process that fails to recognize the constrained nature of outputs in power generation.

In what follows, we focus on the results from our cost-based approach. Based on the estimated variable cost function, we can recover marginal costs of both outputs. Specifically, we have

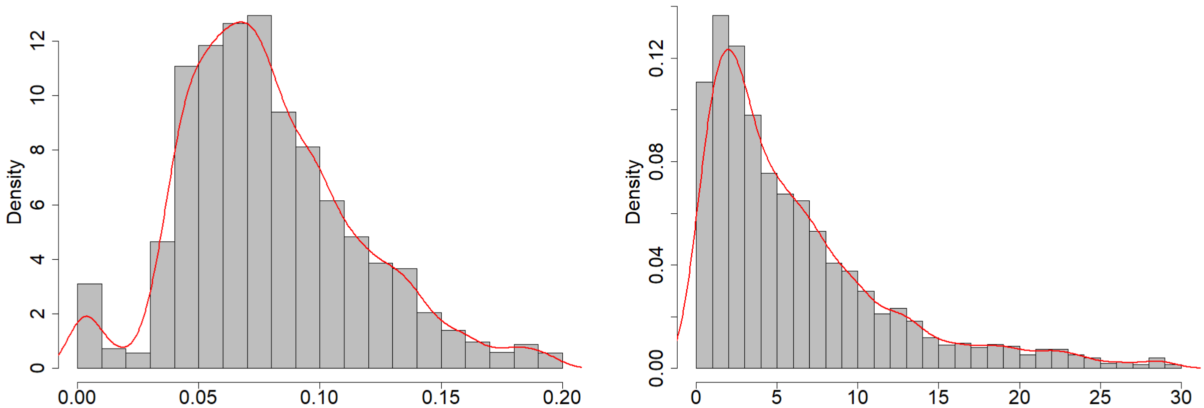
$$MC_{d,it} \equiv \frac{\partial C_{it}}{\partial Q_{d,it}} = \frac{\partial \widetilde{c}_{it}}{\partial q_{d,it}} \frac{C_{it}}{Q_{d,it}} \quad \forall\ d = 1,2, \tag{7.3}$$

where $\frac{\partial \widetilde{c}_{it}}{\partial q_{d,it}}$ is the cost elasticity of output d, which can be obtained from the estimated (3.1), and $\frac{C_{it}}{Q_{d,it}}$ is directly observable in the data. Given that these marginal costs are observation-specific, Figure 1 summarizes them via histograms along with kernel densities. For the net electricity generation (Q_1), the middle 80% of the marginal cost distribution is concentrated between (statistically significant) 0.04 and 0.13, suggesting that generating an additional kilowatt-hour of electricity increases the short-run cost by 4 to 13 cents for most plants. The average MC_1 estimate is 8 cents/kWh. These estimates are in line with the roughly 1 to 15 cent average short-run marginal cost estimates in the 2017 PJM State of the Market report (Monitoring Analytics, LLC, 2018) for combustion turbines and combined cycle plants in 2009–2017. The ratio of 90th to 10th percentiles of MC_1 estimates is 3.6. For plant hours connected to the load (Q_2), we find a much larger variation in marginal cost estimates. The mode of its distribution is about \$2,000–3,000 with the median estimate of about \$4,400, and the distribution has a long right tail, consistent with significant cross-plant heterogeneity (the 90th-to-10th-percentile ratio is 22.4).

Figure 2 shows geographic differences in marginal costs at the state level. The maps are drawn using average estimates of marginal costs for each state, with the states not appearing in our sample left blank. For net electricity generation, states with relatively higher marginal costs tend to concentrate in the Midwest: Kansas, Missouri, Iowa and Wisconsin. Besides them, power generating plants in Kentucky, Arkansas, Arizona and Montana also exhibit relatively higher average marginal costs. For plant hours [see Figure 2(b)], states in the Southeast are the ones that tend to have relatively higher marginal costs, but it is the highest in West Virginia. The latter is fairly unsurprising since West Virginia derives the vast majority of its electric power from coal, and natural gas generation in the state is predominantly small-scale simple-cycle plants which are generally used as peaker plants.

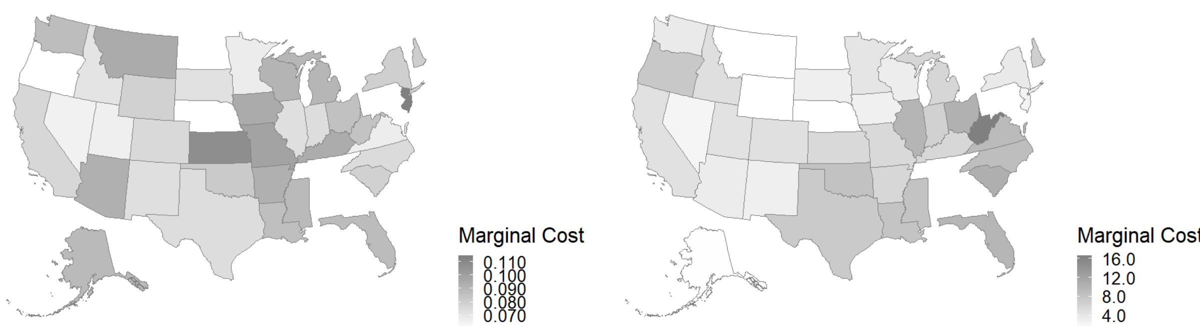
To investigate changes in marginal costs over time, Figure 3 reports boxplots of marginal cost estimates by the year. There is a clear downward trend in the marginal cost of net generation. The evidence suggests a shift of the entire distribution: the median and interquartile ranges are trending down, with whiskers also getting shorter over time; see Figure 3(a). For the marginal cost of plant hours, the downward trend is not as obvious but it in general got smaller in more recent years. Overall, these results suggest that gas-fired power generation got cheaper during our sample period.

Besides marginal costs, productivity of power generating plants is also of interest. In Figure 4, we report a histogram and boxplots of the estimated (log-)productivity (ω). While nominal magnitudes of productivity are not interpretable since it is identified up to a constant only, a wide dispersion thereof is informative of non-trivial (relative) heterogeneity across individual plants' productive efficiency.



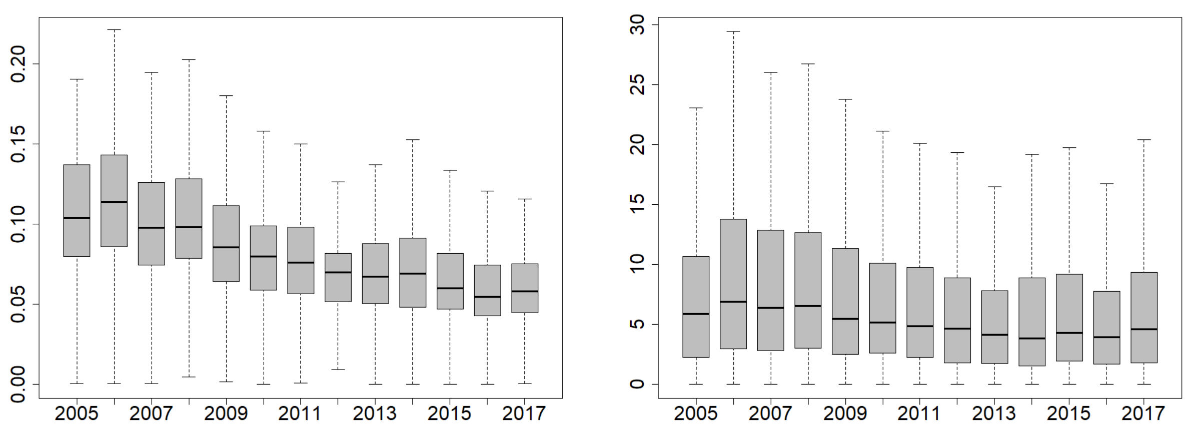
(A) Net Power Generation ($/KwH) (B) Plant Hours Connected to the Load ($1,000/hr)

FIGURE 1. Distributions of Marginal Costs



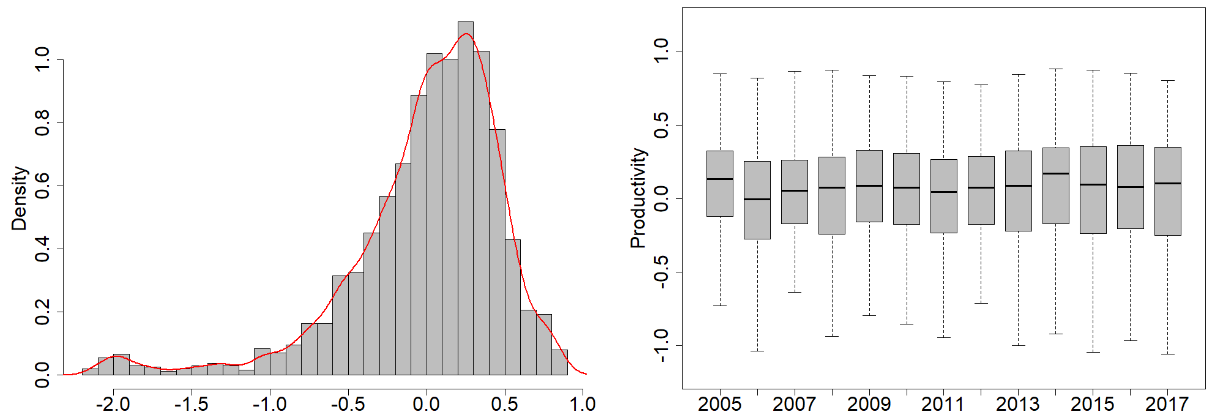
(A) Net Power Generation ($/KwH) (B) Plant-Hours Connected to the Load ($1,000/hr)

FIGURE 2. Average Marginal Costs by State (states without information are left out)



(A) Net Power Generation ($/KwH) (B) Plant Hours Connected to the Load ($1,000/hr)

FIGURE 3. Temporal Variation in Marginal Costs



(A) Pooled Distribution

(B) By-year Distributions

FIGURE 4. Plant Productivity Estimates

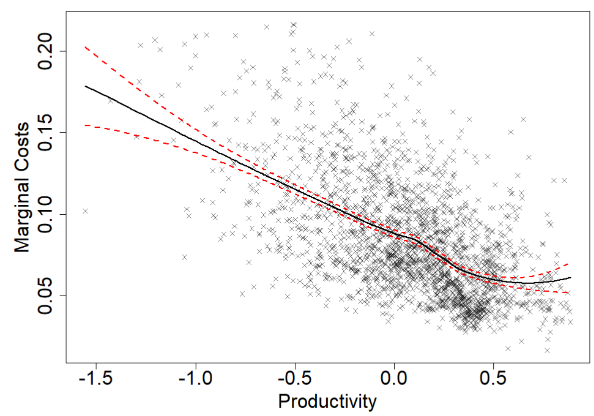


FIGURE 5. Marginal Costs of Net Power Generation vs. Plant Productivity (the fitted line is a local-linear regression with 95% confidence intervals)

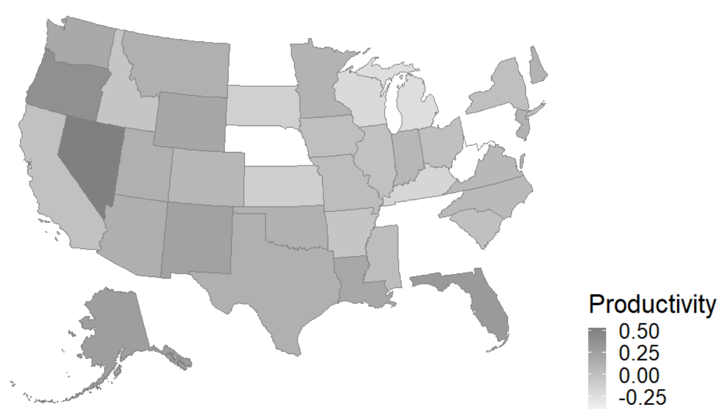


FIGURE 6. Average Plant Productivity by State (states without information are left out)

The mid-90% range of the estimated ω is about 3, implying that plants in the top 5th percentile are at least three times more productive that plants in the bottom 5th percentile of the distribution. At the same time, by-year distributions [see Figure 4(b)] show no evidence of productivity growth in the industry during the sample period. As such, the decline in (short-run) marginal costs that we have documented earlier must have been fueled mainly by input price declines and the convergence to the optimal long-run input allocation rather than improvements in factor-neutral productivity.

Temporal changes aside, we also correlate the *levels* of plant productivity and marginal costs of net power generation in Figure 5. Shown are plant-level marginal cost estimates plotted against productivity, where we also graph a fitted relationship between these two variables using local-linear kernel methods. Consistent with textbook microeconomics, we finds a strong negative association between the marginal cost and plant productivity, whereby more productive plants tend to have lower marginal costs of the electricity generation.

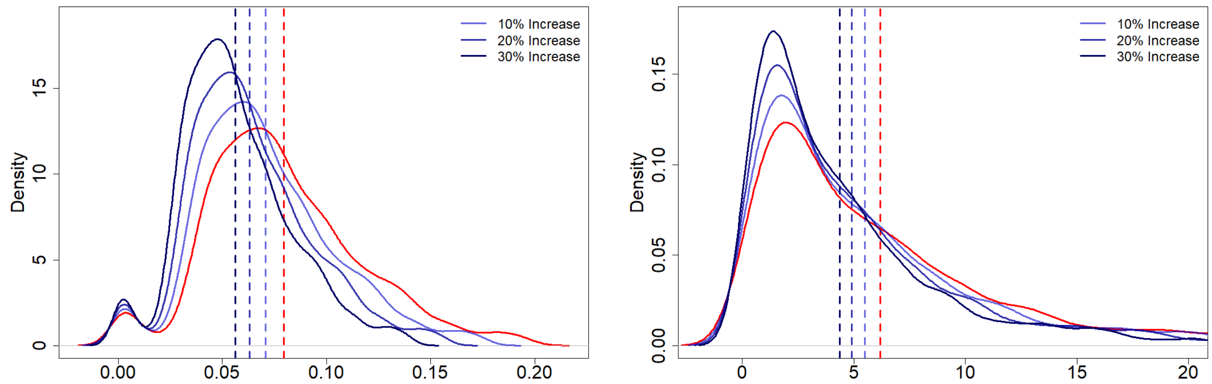
We conclude by examining a geographic distribution of plant productivity. Figure 6 maps average productivity by states. In most states, productivity of electric plants is roughly the same. The few outliers include Nevada, Oregon and Florida, which exhibit a higher than average productivity, whereas plants in Wisconsin, Michigan and Kentucky appear to be less productive.

7.3 Counterfactual analysis

Using the estimated cost function, we examine how the marginal costs of electricity generation and plant-hours connection would have changed if power plants had become more technologically efficient. Recall that in Figure 4(b) we have observed no evidence of productivity growth in the industry during our sample period. Consequently, improvements in ω did not contribute to observed declines in marginal costs. Here, we consider counterfactual scenarios of 10%, 20% and 30% permanent increases in productivity of all power plants in our sample —that is, increases in the long-run mean productivity in the electricity sector—and investigate how these hypothetical productivity boosts would have reduced marginal costs of power plants.

Specifically, according to the marginal cost formula (7.3), assuming there is no change in the production technology of power plants and that output quantities are fixed at observed values (i.e., keeping $\frac{\partial \widetilde{c}}{\partial q_d}$ and Q_d fixed for each $d = 1,2$), increases in ω will affect marginal costs by altering the power plants' total variable costs (C). For the Cobb-Douglas specification in (3.1), a permanent $p\%$ increase in productivity raises ω by $\frac{p}{100}$, consequently reducing $\widetilde{c}$ by $\frac{p}{100}\phi^{Q}$, from where we can recover the counterfactual value of C.

In Figure 7, we plot densities of actual (in red) and counterfactual marginal costs of power generation and plant-hours connection and also show their mean values using vertical dashed lines. We observe that the industry-wide improvement in the long-run productivity materially shifts the distributions of marginal costs to the left. The effects are non-trivial and economically significant. For example, if power plants were 10% (20%) more productive ceteris paribus, the mean marginal cost of power generation would have been 7.1 (6.3) cents over the historical 8 cents per kWh.



(A) Power Generation (B) Plant Hours

FIGURE 7. Counterfactual Effects of Productivity Improvements

8 Conclusion

Motivated by the observation that output quantities produced in many utility and service industries, such as electricity, healthcare or transportation, are exogenously restricted, we consider a dual cost-function approach to indirectly identify the firm's production technology within a structural framework of proxy-variable methods. A dual paradigm is used since a primal formulation of firm production treats the output quantity as a firm's choice, which is inconsistent with its constrained nature in these industries. Besides, relative to conventional proxy-variable estimators, our methodology also has additional advantages. It imposes no structure onto the output market, easily admits multiple outputs and allows input prices to be endogenously correlated with firm productivity. None of these are trivial issues for the identification of production technologies via widely-used primal proxy-variable methods. To operationalize our cost-based estimator, we employ the Markov assumption about firm productivity and, under static cost minimization, use a structural link between the short-run cost function and the corresponding first-order conditions for flexible inputs. The proposed estimator contributes to a growing literature aimed at extending proxy-variable identification methods to new production settings.

References

Ackerberg, D., Frazer, G., Kim, K., Luo, Y., and Yingjun, S. (2020). Under-identification of structural models based on timing and information set assumptions. *Working Paper*.

Ackerberg, D. A., Caves, K., and Frazer, G. (2015). Identification properties of recent production function estimators. *Econometrica*, 83:2411–2451.

Allcott, H., Collard-Wexler, A., and O'Connell, S. D. (2016). How do electricity shortages affect industry? evidence from india. *The American Economic Review*, 106(3):587–624.

Atkinson, S. E. and Luo, R. (2024). Estimation of production technologies with output and environmental constraints. *International Economic Review*, 65(2):755–780.

Bernstein, D. H. (2020). An updated assessment of technical efficiency and returns to scale for U.S. electric power plants. *Energy Policy*, 147.

Bernstein, D. H. and Parmeter, C. F. (2019). Returns to scale in electricity generation: Replicated and revisited. *Energy Economics*, 82:4–15. Replication in Energy Economics.

Borenstein, S., Bushnell, J. B., and Wolak, F. A. (2002). Measuring market inefficiencies in California's restructered wholesale electricity market. *American Economic Review*, 92(5):1376–1405.

Borenstein, S. and Holland, S. (2005). On the efficiency of competitive electricity markets with time-invariant retail prices. *RAND Journal of Economics*, 36(3):469–493.

Caves, D. W., Christensen, L. R., and Swanson, J. A. (1981). Productivity growth, scale economies, and capacity utilization in US railroads, 1955-74. *American Economic Review*, 71(5):994–1002.

Chambers, R. G. (1988). *Applied Production Analysis: A Dual Approach*. Cambridge University Press.

Christensen, L. R. and Greene, W. H. (1976). Economies of Scale in U.S. Electric Power Generation. *Journal of Political Economy*, 84(4):655–676.

Collard-Wexler, A. and De Loecker, J. (2015). Reallocation and technology: Evidence from the US steel industry. *American Economic Review*, 105:131–171.

Considine, T. J. (2000). Cost Structures for Fossil Fuel-Fired Electric Power Generation. *The Energy Journal*, 21(2):83–104.

De Loecker, J. (2007). Do exports generate higher productivity? evidence from slovenia. *Journal of International Economics*, 73(1):69–98.

De Loecker, J. (2013). Detecting learning by exporting. *American Economic Journal: Microeconomics*, 5:1–21.

De Loecker, J., Goldberg, P. K., Khandelwal, A. K., and Pavcnik, N. (2016). Prices, markups, and trade reform. *Econometrica*, 84:445–510.

De Loecker, J. and Warzynski, F. (2012). Markups and firm-level export status. *American Economic Review*, 102:2437–2471.

Demirer, M. (2020). Production function estimation with factor-augmenting technology: An application to markups. *Working Paper, MIT*.

Dhawan, R. (2001). Firm size and productivity differential: theory and evidence from a panel of us firms. *Journal of Economic Behavior & Organization*, 44(3):269–293.

Dhyne, E., Petrin, A., Smeets, V., and Warzynski, F. (2022). Theory for extending single-product production function estimation to multi-product settings. *NBER Working Paper 30784*.

Diewert, W. E. (1971). An application of the Shephard duality theorem: A generalized Leontief production function. *Journal of Political Economy*, 79:481–507.

Diewert, W. E. (1973). Functional forms for profit and transformation functions. *Journal of Economic Theory*, 6:284–316.

Doraszelski, U. and Jaumandreu, J. (2013). R&D and productivity: Estimating endogenous productivity. *Review of Economic Studies*, 80:1338–1383.

Doraszelski, U. and Jaumandreu, J. (2018). Measuring the bias of technological chance. *Journal of Political Economy*, 126:1027–1084.

Doraszelski, U. and Jaumandreu, J. (2019). Using cost minimization to estimate markups. *Working Paper, University of Pennsylvania*.

Färe, R. and Primont, D. (1995). *Multi-Output Production and Duality: Theory and Applications*. Kluwer Academic Publishers, Boston, MA.

Flynn, Z., Gandhi, A., and Traina, J. (2019). Measuring markups with production data. *Working Paper, University of Pennsylvania*.

Foster, L., Haltiwanger, J., and Syverson, C. (2008). Reallocation, firm turnover, and efficiency: Selection on productivity or profitability? *American Economic Review*, 98(1):394–425.

Gandhi, A., Navarro, S., and Rivers, D. (2020). On the identification of gross output production functions. *Journal of Political Economy*.

Greene, W. H. (2008). The econometric approach to efficiency analysis. In Fried, H. O., Lovell, C. A. K., and Schmidt, S. S., editors, *The Measurement of Productive Efficiency and Productivity Change*. Oxford University Press.

Grieco, P. L. E., Li, S., and Zhang, H. (2016). Production function estimation with unobserved input price dispersion. *International Economic Review*, 57:665–689.

Grieco, P. L. E., Li, S., and Zhang, H. (2020). Input prices, productivity and trade dynamics: Long-run effects of liberalization on Chinese paint manufacturers. *Working Paper*.

Grieco, P. L. E. and McDevitt, R. C. (2017). Productivity and quality in health care: Evidence from the dialysis industry. *Review of Economic Studies*, 84:1071–1105.

Humphrey, D. B. and Pulley, L. B. (1997). Banks' Responses to Deregulation: Profits, Technology, and Efficiency. *Journal of Money, Credit and Banking*, 29(1):73–93.

Kim, K., Luo, Y., and Su, Y. (2019). A robust approach to estimating production functions: Replication of the ACF procedure. *Journal of Applied Econometrics*, 34:612–619.

Knittel, C. (2002). Alternative regulatory methods and firm efficiency: Stochastic frontier evidence from the U.S. electricity industry. *The Review of Economics and Statistics*, 84(3):530–540.

Konings, J. and Vanormelingen, S. (2015). The impact of training on productivity and wages: Firm-level evidence. *Review of Economics and Statistics*, 97:485–497.

Kugler, M. and Verhoogen, E. (2011). Prices, Plant Size, and Product Quality. *Review of Economic Studies*, 79(1):307–339.

Kugler, M. and Verhoogen, E. (2012). Prices, plant size, and product quality. *The Review of Economic Studies*, 79(1):307–339.

Levinsohn, J. and Petrin, A. (2003). Estimating production functions using inputs to control for unobservables. *Review of Economic Studies*, 70:317–341.

Malikov, E. and Lien, G. (2021). Proxy variable estimation of multiproduct production functions. *American Journal of Agricultural Economics*, 103(5):1878–1902.

Malikov, E., Zhang, J., Zhao, S., and Kumbhakar, S. C. (2022). Accounting for cross-location technological heterogeneity in the measurement of operations efficiency and productivity. *Journal of Operations Management*, 68:153–184.

Malikov, E. and Zhao, S. (2023). On the estimation of cross-firm productivity spillovers with an application to FDI. *Review of Economics and Statistics*, 105:1207–1223.

Malikov, E., Zhao, S., and Zhang, J. (2024). A system approach to structural identification of production functions with multiple-dimensional productivity. *Advances in Econometrics*, 46:211–263.

Manova, K. and Zhang, Z. (2012). Export prices across firms and destinations. *The Quarterly Journal of Economics*, page qjr051.

McFadden, D. L. (1978). Cost, revenue, and profit functions. In Fuss, M. and McFadden, D. L., editors, *Production Economics: A Dual Approach to Theory and Applications*, volume I: The Theory of Production. Amsterdam: North-Holland.

Mo, J., Qiu, L. D., Zhang, H., and Dong, X. (2021). What you import matters for productivity growth: Experience from Chinese manufacturing firms. *Journal of Development Economics*, 152. Article 102677.

Monitoring Analytics, LLC (2018). State of the market report for pjm. Technical report, Independent Market

Monitor for PJM.

Nerlove, M. (1963). Returns to Scale in Electricity Supply. In *Measurement in Economics-Studies in Mathematical Economics and Econometrics in Memory of Yehuda Grunfeld*, chapter 7, pages 167–198. Stanford Univ. Press.

Newey, W. K. (1984). A method of moments interpretation of sequential estimators. *Economics Letters*, 14(2):201–206.

O'Donnell, C. J. (2014). Econometric estimation of distance functions and associated measures of productivity and efficiency change. *Journal of Productivity Analysis*, 41:187–200.

Olley, G. S. and Pakes, A. (1996). The dynamics of productivity in the telecommunications equipment industry. *Econometrica*, 64:1263–1297.

Ornaghi, C. (2006). Assessing the effects of measurement errors on the estimation of production functions. *Journal of Applied Econometrics*, 21(6):879–891.

Orr, S. (2022). Within-firm productivity dispersion: Estimates and implications. *Journal of Political Economy*, 130:2771–2828.

Panzar, J. C. and Willig, R. D. (1977). Economies of scale in multi-output production. *Quarterly Journal of Economics*, 91:481–493.

Shephard, R. W. (1953). *Cost and Production Functions*. Princeton University Press.

Shephard, R. W. (1971). *Theory of Cost and Production Functions*. Princeton University Press.

Tybout, J. R. (2003). Plant-and firm-level evidence on new trade theories. *Handbook of International Trade*, 1:388–415.

Valmari, N. (2023). Estimating production functions of multiproduct firms. *Review of Economic Studies*, 90(6):3315–3342.

Verhoogen, E. A. (2008). Trade, quality upgrading, and wage inequality in the mexican manufacturing sector. *The Quarterly Journal of Economics*, 123(2):489–530.

Zhang, H. (2017). Static and dynamic gains from costly importing of intermediate inputs: Evidence from Colombia. *European Economic Review*, 91:118–145.

Zhang, H. (2019). Non-neutral technology, firm heterogeneity, and labor demand. *Journal of Development Economics*, 140:145–168.

Zhang, J. and Malikov, E. (2023). Detecting learning by exporting and from exporters. *Journal of Productivity Analysis*. forthcoming.

Zhao, S., Malikov, E., and Kumbhakar, S. (2024). On the estimation of nonneutral production technologies without homothetic separability. Working Paper.

Acknowledgments

We greatly benefited from discussions with participants of the 2023 North American Productivity Workshop at the University of Miami. Bernstein, who is also affiliated with the Federal Energy Regulatory Commission, would like to note that the views expressed here are his own and do not necessarily represent the views of the Federal Energy Regulatory Commission, the U.S. Department of Energy or the U.S. Government.

Disclosure Statement

The authors report there are no competing interests to declare.

Data Availability Statement

Data used in this study are included in the replication package accompanying the paper.

APPENDIX

A Self-Duality of the Cobb-Douglas

Throughout this appendix, we drop the "it" subscript to minimize notational clutter. To further streamline the exposition, let $\boldsymbol{K} = (K, L)'$ and $\boldsymbol{V} = (M, E)'$. Now, consider the Cobb-Douglas transformation function:

$$\Pi_{d=1}^{D} [Q_d^*]^{-\beta_d^Q} K^{\beta^K} L^{\beta^L} M^{\beta^M} E^{\beta^E} = 1, \tag{A.1}$$

which is separable and homogeneous in "net producible" outputs $Q_d^* = Q_d \exp\{-\omega\}\ \forall d$ and is formulated such that all elasticity parameters are non-negative.

The firm's variable cost minimization problem is

$$\min_{\{\mathcal{M},\mathcal{E},\lambda\}} \quad \mathcal{L} = P^M \mathcal{M} + P^E \mathcal{E} + \lambda \left[\Pi_{d=1}^{D} [Q_d^*]^{\beta_d^Q} - K^{\beta^K} L^{\beta^L} \mathcal{M}^{\beta^M} \mathcal{E}^{\beta^E} \right], \tag{A.2}$$

with the corresponding first-order conditions for variable inputs given by

$$\begin{aligned} P^M &= \lambda \beta^M K^{\beta^K} L^{\beta^L} M^{\beta^M - 1} E^{\beta^E} \\ P^E &= \lambda \beta^E K^{\beta^K} L^{\beta^L} M^{\beta^M} E^{\beta^E - 1}, \end{aligned}$$

from where we have that, at the optimum, $\frac{E}{M} = \frac{P^M}{P^E} \times \frac{\beta^E}{\beta^M}$. Substituting for these two inputs into the transformation function in (A.1), we obtain conditional demand functions:

$$\begin{aligned} M &= \left[\Pi_{d=1}^{D} [Q_d^*]^{\frac{\beta_d^Q}{\beta^M + \beta^E}} \right] \left[K^{\beta^K} L^{\beta^L} \right]^{-\frac{1}{\beta^M + \beta^E}} \left[\frac{P^E \beta^M}{P^M \beta^E} \right]^{\frac{\beta^E}{\beta^M + \beta^E}} \\ E &= \left[\Pi_{d=1}^{D} [Q_d^*]^{\frac{\beta_d^Q}{\beta^M + \beta^E}} \right] \left[K^{\beta^K} L^{\beta^L} \right]^{-\frac{1}{\beta^M + \beta^E}} \left[\frac{P^M \beta^E}{P^E \beta^M} \right]^{\frac{\beta^M}{\beta^M + \beta^E}}. \end{aligned}$$

Thus, the variable cost function is

$$C = P^M M + P^E E = \left[\Pi_{d=1}^D \left[Q_d^*\right]^{\frac{\beta_d^Q}{\beta^M+\beta^E}}\right]\left[K^{\beta^K} L^{\beta^L}\right]^{-\frac{1}{\beta^M+\beta^E}} \left(P^M\right)^{\frac{\beta^M}{\beta^M+\beta^E}} \left(P^E\right)^{\frac{\beta^E}{\beta^M+\beta^E}} \left[\frac{\beta^M}{\beta^E}\right]^{\frac{\beta^E}{\beta^M+\beta^E}} \left(1+\frac{\beta^E}{\beta^M}\right),$$

normalizing which using P^E, we arrive at

$$\widetilde{C} = \left[\Pi_{d=1}^D \left[Q_d^*\right]^{\frac{\beta_d^Q}{\beta^M+\beta^E}}\right] K^{\frac{-\beta^K}{\beta^M+\beta^E}} L^{\frac{-\beta^L}{\beta^M+\beta^E}} \left[\frac{P^M}{P^E}\right]^{\frac{\beta^M}{\beta^M+\beta^E}} \left[\frac{\beta^M}{\beta^E}\right]^{\frac{\beta^E}{\beta^M+\beta^E}} \left(1+\frac{\beta^E}{\beta^M}\right).$$

Logging both sides yields the log-linearized Cobb-Douglas cost function:

$$\widetilde{c} = \underbrace{\ln\left[\left(\frac{\beta^M}{\beta^E}\right)^{\frac{\beta^E}{\beta^M+\beta^E}}\left(1+\frac{\beta^E}{\beta^M}\right)\right]}_{\alpha^0} + \underbrace{\left(\frac{\beta^M}{\beta^M+\beta^E}\right)}_{\alpha^M}\widetilde{p}^M + \underbrace{\left(\frac{-\beta^K}{\beta^M+\beta^E}\right)}_{\alpha^K} k + \underbrace{\left(\frac{-\beta^L}{\beta^M+\beta^E}\right)}_{\alpha^L} l + \sum_{d=1}^D \underbrace{\left(\frac{\beta_d^Q}{\beta^M+\beta^E}\right)}_{\alpha_d^Q} q_d - \underbrace{\left(\sum_{d=1}^D \frac{\beta_d^Q}{\beta^M+\beta^E}\right)}_{\phi^Q}\omega, \tag{A.3}$$

where we have substituted $q_d - \omega$ for q_d^*. From above, it is evident that (A.3), which corresponds to a Cobb-Douglas transformation function, also has a log-linear Cobb-Douglas form. In fact, it is of the same form as a Cobb-Douglas variable cost function assumed in (3.1).

B Expanded Expression for $\Psi(\boldsymbol{\rho}, \boldsymbol{\alpha})$

Throughout this appendix, we drop the "i" subscript to minimize notational clutter. The vector of self-instrumenting gradients of the regression function $r_t(\boldsymbol{\rho}, \boldsymbol{\alpha}^{V-})$ with respect to the parameter vector given in (3.10) is

$$\nabla r_t(\boldsymbol{\rho}, \boldsymbol{\alpha}^{V-}) = \begin{bmatrix} 1 \\ \widetilde{\varkappa}_{t-1} - \alpha^K k_{t-1} - \alpha^Q q_{t-1} \\ k_t - \rho_1 k_{t-1} \\ q_t - \rho_1 q_{t-1} \end{bmatrix}.$$

The corresponding $\Psi(\boldsymbol{\rho}, \boldsymbol{\alpha}^{V-}) = \mathbb{E}\left[\nabla r_{it}(\boldsymbol{\rho}, \boldsymbol{\alpha}^{V-}) \nabla r_{it}(\boldsymbol{\rho}, \boldsymbol{\alpha}^{V-})'\right]$ matrix equals

$$\mathbb{E}\begin{bmatrix} 1 & [\widetilde{\varkappa}_{t-1} - \alpha^K k_{t-1} - \alpha^Q q_{t-1}] & k_t - \rho_1 k_{t-1} & q_t - \rho_1 q_{t-1} \\ [\widetilde{\varkappa}_{t-1} - \alpha^K k_{t-1} - \alpha^Q q_{t-1}] & [\widetilde{\varkappa}_{t-1} - \alpha^K k_{t-1} - \alpha^Q q_{t-1}]^2 & (k_t - \rho_1 k_{t-1})[\widetilde{\varkappa}_{t-1} - \alpha^K k_{t-1} - \alpha^Q q_{t-1}] & (q_t - \rho_1 q_{t-1})[\widetilde{\varkappa}_{t-1} - \alpha^K k_{t-1} - \alpha^Q q_{t-1}] \\ k_t - \rho_1 k_{t-1} & (k_t - \rho_1 k_{t-1})[\widetilde{\varkappa}_{t-1} - \alpha^K k_{t-1} - \alpha^Q q_{t-1}] & (k_t - \rho_1 k_{t-1})^2 & (q_t - \rho_1 q_{t-1})(k_t - \rho_1 k_{t-1}) \\ q_t - \rho_1 q_{t-1} & (q_t - \rho_1 q_{t-1})[\widetilde{\varkappa}_{t-1} - \alpha^K k_{t-1} - \alpha^Q q_{t-1}] & (k_t - \rho_1 k_{t-1})(q_t - \rho_1 q_{t-1}) & (q_t - \rho_1 q_{t-1})^2 \end{bmatrix}.$$

C Restricted Translog

Although very popular in applied productivity studies, the Cobb-Douglas form is restrictive in that it assumes constant (over time and across firms) elasticities which forces firms of all sizes—small and large—to have the same technology, including the same returns to scale. In what follows, we therefore consider a more flexible functional form for the firm's variable cost structure $H(\cdot)$ in (2.7) capable of accommodating technological heterogeneity across firms, viz.,

$$\begin{aligned}\widetilde{c}_{it} = \alpha^0 + \sum_{j=2}^{J} \alpha_j^V \widetilde{p}_{j,it}^V + \sum_{m=1}^{M} \alpha_m^K k_{m,it} + \sum_{m=1}^{M}\sum_{m'=1}^{M} \alpha_{mm'}^{KK} k_{m,it} k_{m',it} + \\ \sum_{j=2}^{J}\sum_{m=1}^{M} \alpha_{jm}^{VK} \widetilde{p}_{j,it}^V k_{m,it} + \sum_{d=1}^{D} \alpha_d^Q q_{d,it} + \sum_{d=1}^{D}\sum_{d'=1}^{D} \alpha_{dd'}^{QQ} q_{d,it} q_{d',it} - \phi^Q \omega_{it},\end{aligned} \tag{C.1}$$

where, just like in the Cobb-Douglas case described in Section 3, parameters need be restricted to ensure homogeneity of degree ϕ^Q in outputs. Given the Young symmetry, we can achieve the latter by imposing

$$\phi^Q = \sum_{d=1}^{D} \alpha_d^Q \quad \text{and} \quad \sum_{d=1}^{D} \alpha_{dd'}^{QQ} = 0 \ \forall d' \tag{C.2}$$

on (C.1) during the estimation.

It is immediately evident that, per (C.1), cost elasticities of quasi-fixed inputs and variable input prices are no longer constant. Among other things, the corresponding RTS measure now also depends on cost-function covariates and, thus, is observation-specific.[18] In effect, (C.1) is a restricted translog where, while allowing for quadratics and cross-products in dynamic inputs, we restrict the cost function to be partially linear in input prices. By suppressing "own" quadratic dynamics in input prices we guarantee that variable input cost shares do not depend on prices, which enables us to extend our identification arguments from Section 3 to this more flexible specification. Also note that the lack of cross-products of outputs with other cost covariates follows from the output separability assumption about the technology.

As in the Cobb-Douglas case, to address the endogeneity of relative input prices, we identify the parameters associated with $\{\widetilde{p}_{j,it}^V\}$ from the variable input cost share equations in the first step. Applying Shephard's lemma to (C.1), we now get

$$S_{j,it}^V = \alpha_j^V + \sum_{m=1}^{M} \alpha_{jm}^{VK} k_{m,it}, \tag{C.3}$$

which is, expectedly, no longer a constant. Expression (C.3) suggests a linear least-squares regression of the jth variable input's cost share on quasi-fixed inputs which, given the predeterminedness of the

[18]Namely, we now have that RTS $= \left[1 - \sum_m \left(\alpha_m^K + \sum_{m'} \alpha_{mm'}^{KK} k_{m',it} + \sum_j \alpha_{jm}^{VK} \widetilde{p}_{j,it}^V\right)\right] / \phi^Q$, whereas in the Cobb-Douglas case RTS $= \left(1 - \sum_m \alpha_m^K\right) / \phi^Q$ is fixed.

latter, identifies $\left(\alpha_j^V, \alpha_{j1}^{VK}, \ldots, \alpha_{jM}^{VK}\right)'$:

$$\begin{bmatrix} \alpha_j^V \\ \alpha_{j1}^{VK} \\ \vdots \\ \alpha_{jM}^{VK} \end{bmatrix} = \mathbb{E}\Big[(1, k_{1,it}, \ldots, k_{M,it})'(1, k_{1,it}, \ldots, k_{M,it})\Big]^{-1} \mathbb{E}\Big[(1, k_{1,it}, \ldots, k_{M,it})' S_{j,it}^V\Big] \quad \forall j = 2, \ldots, J. \tag{C.4}$$

To arrive at the second step, we rewrite the cost function in (C.1) as

$$\widetilde{\varkappa}_{it} = \alpha^0 + \sum_{m=1}^{M} \alpha_m^K k_{m,it} + \sum_{m=1}^{M} \sum_{m'=1}^{M} \alpha_{mm'}^{KK} k_{m,it} k_{m',it} + \sum_{d=1}^{D} \alpha_d^Q q_{d,it} + \sum_{d=1}^{D} \sum_{d'=1}^{D} \alpha_{dd'}^{QQ} q_{d,it} q_{d',it} - \phi^Q \omega_{it}, \tag{C.5}$$

where the new outcome variable $\widetilde{\varkappa}_{it} \equiv \widetilde{c}_{it} - \sum_{j=2}^{J} \alpha_j^V \widetilde{p}_{j,it}^V - \sum_{j=2}^{J} \sum_{m=1}^{M} \alpha_{jm}^{VK} \widetilde{p}_{j,it}^V k_{m,it}$ contains the already identified parameters from the first step and can thus be treated as an observable. The model in (C.5) no longer contains endogenous relative prices among its regressors and, to estimate it, we handle latency of firm productivity analogously to our earlier strategy: by quasi-differencing. Specifically, inverting (C.5) for ω_{it} to obtain a proxy function and plugging it in the Markov process for productivity, we obtain

$$\begin{aligned} &\frac{1}{\phi^Q}\left[\alpha^0 + \sum_{m=1}^{M} \alpha_m^K k_{m,it} + \sum_{m=1}^{M} \sum_{m'=1}^{M} \alpha_{mm'}^{KK} k_{m,it} k_{m',it} + \sum_{d=1}^{D} \alpha_d^Q q_{d,it} + \sum_{d=1}^{D} \sum_{d'=1}^{D} \alpha_{dd'}^{QQ} q_{d,it} q_{d',it} - \widetilde{\varkappa}_{it}\right] = \\ &g\left(\frac{1}{\phi^Q}\left[\alpha^0 + \sum_{m=1}^{M} \alpha_m^K k_{m,it-1} + \sum_{m=1}^{M} \sum_{m'=1}^{M} \alpha_{mm'}^{KK} k_{m,it-1} k_{m',it-1} + \sum_{d=1}^{D} \alpha_d^Q q_{d,it-1} + \sum_{d=1}^{D} \sum_{d'=1}^{D} \alpha_{dd'}^{QQ} q_{d,it} q_{d',it} - \widetilde{\varkappa}_{it-1}\right]\right) + \zeta_{it}, \end{aligned} \tag{C.6}$$

which, just like equation (3.7) in the Cobb-Douglas case, identifies all the remaining key cost-function coefficients via nonlinear least squares subject to parameter restrictions in (C.2).

C.1 Estimation of a Restricted Translog Cost Function

A restricted translog cost-function estimator is operationalized following the same procedure as in Section 4. In the first step, we estimate $\widetilde{p}_j^V$-related parameters using a sample analogue of (C.4), which yields the following OLS estimator:

$$\begin{bmatrix} \widehat{\alpha}_j^V \\ \widehat{\alpha}_{j1}^{VK} \\ \vdots \\ \widehat{\alpha}_{jM}^{VK} \end{bmatrix} = \Big[\sum_{i,t} (1, k_{1,it}, \ldots, k_{M,it})'(1, k_{1,it}, \ldots, k_{M,it})\Big]^{-1} \sum_{i,t} (1, k_{1,it}, \ldots, k_{M,it})' S_{j,it}^V \quad \forall j = 2, \ldots, J. \tag{C.7}$$

Using these first-step parameter estimates, we can then construct a feasible analogue of $\widetilde{\varkappa}_{it}$; let us denote it as $\widetilde{\varkappa}_{it}(\widehat{\boldsymbol{\alpha}}^V) = \widetilde{c}_{it} - \sum_{j=2}^{J} \widehat{\alpha}_j^V \widetilde{p}_{j,it}^V - \sum_{j=2}^{J} \sum_{m=1}^{M} \widehat{\alpha}_{jm}^{VK} \widetilde{p}_{j,it}^V k_{m,it}$, where $\widehat{\boldsymbol{\alpha}}^V$ includes all estimated input-price-related coefficients.

Assuming that $g(\cdot)$ is a linear (parametric) function with no intercept, viz., $g(\cdot) = \rho_1 \times (\cdot)$, we can then estimate the (negative of) second-step equation in (C.6) via restricted NLLS:

$$\begin{bmatrix} \widehat{\boldsymbol{\rho}} \\ \widehat{\boldsymbol{\alpha}}^{V-} \end{bmatrix} = \arg\min_{\boldsymbol{\rho},\boldsymbol{\alpha}^{V-}} \sum_{i,t} \left[w_{it}\left(\boldsymbol{\alpha}^{V-} \middle| \widehat{\boldsymbol{\alpha}}^{V}\right) - \rho_0 - \rho_1 w_{it-1}\left(\boldsymbol{\alpha}^{V-} \middle| \widehat{\boldsymbol{\alpha}}^{V}\right)\right]^2 \quad \text{subject to (C.2)}, \tag{C.8}$$

where $\rho_0 \equiv \alpha^0(1-\rho_1)$ as before but we redefine $\boldsymbol{\alpha}^{V-} = \left(\alpha_1^K, \dots, \alpha_M^K, \alpha_{11}^{KK}, \dots, \alpha_{MM}^{KK}, \alpha_1^Q, \dots, \alpha_D^Q\right)'$ and

$$w_{it}\left(\boldsymbol{\alpha}^{V-} \middle| \widehat{\boldsymbol{\alpha}}^{V}\right) \equiv \widetilde{\varkappa}_{it}\left(\widehat{\boldsymbol{\alpha}}^{V}\right) - \sum_{m=1}^{M} \alpha_m^K k_{m,it} - \sum_{m=1}^{M}\sum_{m'=1}^{M} \alpha_{mm'}^{KK} k_{m,it} k_{m',it} - \sum_{d=1}^{D} \alpha_d^Q q_{d,it} - \sum_{d=1}^{D}\sum_{d'=1}^{D} \alpha_{dd'}^{QQ} q_{d,it} q_{d',it}.$$

To rewrite this two-step sequential estimator as a simultaneous multiple-equation GMM system of moment conditions, let all parameters in (C.7)–(C.8) be denoted by $\boldsymbol{\Theta} = \left(\left(\boldsymbol{\alpha}^V\right)', \left(\boldsymbol{\alpha}^{V-}\right)', \boldsymbol{\rho}'\right)'$ and the residual function in the second step, after substituting the first-step estimator of $\widetilde{\varkappa}_{it}$ in, be $e_{it}(\boldsymbol{\Theta})$. Then, the two estimation steps imply the following (optimal) moment restrictions:

$$\mathbb{E}\left[\boldsymbol{f}_{it}(\boldsymbol{\Theta})\right] \equiv \mathbb{E}\begin{bmatrix} \begin{pmatrix} S_{2,it}^V - \sum_{m=1}^M \alpha_{2m}^{VK} k_{m,it} \\ \vdots \\ S_{J,it}^V - \sum_{m=1}^M \alpha_{Jm}^{VK} k_{m,it} \end{pmatrix} \otimes \left(1, k_{1,it}, \dots, k_{M,it}\right)' \\ e_{it}(\boldsymbol{\Theta}) \dfrac{\partial e_{it}(\boldsymbol{\Theta})}{\partial\left(\left(\boldsymbol{\alpha}^{V-}\right)', \boldsymbol{\rho}'\right)'} \end{bmatrix} = \mathbf{0}. \tag{C.9}$$