Measuring Markups with Revenue Data

Ivan Kirov
Paolo Mengano
James Traina*

August 2025

Abstract

When output prices are unobserved, standard production-based markup estimators are biased and inconsistent because they are unable to distinguish whether firms have higher revenues due to higher prices or higher quantities. Building on work designed for competitive environments, we propose a novel method that solves this problem using only revenue data. We flexibly model markups as a specified function of observables and fixed effects, supporting a broad class of variable-markup frameworks. We explicitly adopt a Markovian revenue productivity process, a commonly implicit assumption in the literature. Our suggested two-step approach is simple in concept and implementation, requiring only common regression techniques.

JEL Codes: C14, C33, D24, D43, L11
Keywords: Markups, Revenue, Omitted Price Bias, Production Functions, Productivity

*Kirov: Analysis Group, ikirov@uchicago.edu. Mengano: Esade Business School, paolo.mengano@esade.edu. Traina: NYU Stern Abu Dhabi, james.traina@nyu.edu. We would like to express our sincere gratitude to Erik Hurst, Chad Syverson, Ali Hortaçsu, Brent Neiman, and Luigi Zingales for their invaluable feedback and insightful discussions on earlier drafts of this paper. We also appreciate the constructive comments and suggestions received from participants at UChicago IO Lunch, as well as Anhua Chen, Zach Flynn, Mike Gibbs, Arshia Hashemi, and Uyen Tran for helpful comments and discussions. Corresponding email address: james.traina@nyu.edu.

Firms in modern economies are experiencing long-term structural shifts: average firm size is expanding, industries are concentrating, and business dynamism is declining (Grullon et al., 2019; Decker et al., 2016). Two competing narratives have emerged to explain these trends. One attributes them to technological change and intangible capital, with positive welfare implications (Autor et al., 2020; Bessen, 2020). The other attributes them to market power and entry barriers, with negative welfare implications (De Loecker et al., 2020; Gutiérrez and Philippon, 2019). Policy responses hinge on which narrative is correct: should we embrace these changes as inevitable products of technological progress, or intervene to preserve competitive markets? At the heart of this question is measuring firm-level markups—the ratio of price to marginal cost—at the macroeconomic scale. While measuring markups poses deep empirical challenges, the importance of the issue demands we confront them head-on.

Recent studies respond to this demand by adopting the production approach to markup estimation (De Loecker et al., 2020; Traina, 2018; Diez et al., 2018; Calligaris et al., 2018). This approach sidesteps the need to specify demand or competition, making it scalable across firms and time. However, the production-function estimators underpinning this approach were designed for perfectly competitive environments. In such settings, variation in prices only reflects variation in quality, so revenue is a good stand-in for "quality-adjusted" output. But imperfectly competitive environments muddle the link between revenue and output because we don't know whether revenues are high because of high quality-adjusted output or market power. Indeed, recent work argues for a fundamental role of detailed price information to recover unbiased and consistent markup estimates (Klette and Griliches, 1996; Doraszelski and Jaumandreu, 2019; Bond et al., 2021).[1] The empirical challenge is that most production datasets offer only revenue information, not price.[2] Researchers interested in studying competition at macroeconomic scales face a frustrating gap between their econometric frameworks and the data available to estimate them.

To bridge this gap, we present an approach that delivers unbiased and consistent markup estimates using only revenue data. Consider the mechanics of production-based markup estimation. Cost minimization implies markups equal the ratio of any flexible input's

[1]Using an industry price index to recover quantities does not solve this problem (Bond et al., 2021) except in two special cases: perfect competition, or no firm heterogeneity in prices (Hashemi et al., 2022).

[2]Compustat, Worldscope, and Orbis financial-statement datasets, as well as the US, Colombian, Chilean, India, Indonesia, and Slovenian manufacturing surveys, do not include output price information.

output elasticity to its revenue share. In practice, researchers observe revenue shares and estimate output elasticities, inferring production-based markups as the residual. The key challenge is to disentangle what part of revenue share variation reflects technology (output elasticity), and what part reflects market power (markup). To do so, we need separating variation that affects one but not the other. Economic theory tells us such variation exists—anything that affects markups through pricing, such as proxies for demand or competitive behavior, should shift markups without affecting technology.

Starting from the firm's cost-minimization problem, we draw out the link between the flexible input's revenue share and its production function and unobserved firm revenue productivity. We model markups as a flexible function of observables and fixed effects—demand conditions, market structure, and other determinants of market power—allowing us to recover revenue elasticities and productivity shocks realized only after firms make production decisions. To be precise, this markup function allows the share regression to identify revenue elasticities, which combine output elasticities with markups. We then leverage the revenue productivity process to separate these components. We show this approach solves the omitted price bias generated with revenue data, and our framework accommodates a broad class of variable-markup frameworks.

We depart from modeling the physical productivity process and instead model the revenue productivity process. Though unconventional at first blush, this approach is consistent with data limitations and grounded either implicitly or explicitly in the vast majority of earlier work (Klette and Griliches, 1996; De Loecker, 2011; De Loecker and Warzynski, 2012; De Loecker et al., 2020). Foster et al. (2008) is one of the few papers with independent data to study differences between physical and revenue productivity, and finds similar estimates for persistence and role in determining exit. Our choice can be viewed as an alternative to the traditional choice of modeling physical productivity as the single Markovian state variable that's a reasonable approximation of the changing conditions that inform firm decisions. In the Appendix, we discuss the microfoundations for this commonly-used Markovian revenue productivity assumption, showing how and when it emerges from equilibrium behavior in dynamic oligopoly models. With better data and extensions to our method to handle multiple unobservables, one can separate the two. But given current data limitations, in this paper we're trying to take a solid step in the right direction. And given this minimal setup, our two-step estimator is straightforward, consistent, and easily applied to readily accessible datasets.

Our key contribution is a novel method for inferring markups without requiring output price data, thus addressing a long-standing problem in estimating production functions: the omitted price bias. The first stage of our estimator is a share regression in the spirit of Gandhi et al. (2020), which directly estimates firm revenue elasticities using the cost minimization first-order condition. The second stage then uses information and timing assumptions to separate the effects of inputs on prices and output, recovering output elasticities. We can then measure unbiased and consistent firm-level markups with these estimates. Our framework also suggests modifications to traditional production function estimators (Blundell and Bond, 2000) that impose more restrictive (physical or revenue) productivity structures. Our approach broadly applies across data sources and market settings, including potential use in macroeconomic, trade, labor, and finance research. It might also outperform standard estimators even when price data are available. For example, even if we had information on output prices, we would need to quality-adjust physical quantities to make sure outputs are comparable across firms, yet quality data are even rarer than price data in practice. Many firms also produce multiple products, which complicates how researchers should interpret quantity data without added assumptions on the production process De Loecker (2011). Our approach eases both concerns and is thus attractive even with price data.

The literature on production function estimation has long noted the omitted price bias in measuring markups with revenue data. Klette and Griliches (1996) shows that markup estimators are biased and inconsistent without information on output prices. This bias arises because unobserved prices enter the residual of the estimating equation but also affect firm input choices (e.g., higher prices spur firms to increase output by using more inputs). Consequently, omitted prices cause a classic simultaneity problem. Doraszelski and Jaumandreu (2019) extend these results to modern production function estimators, while Bond et al. (2021) show the bias leads to measured markups of one for profit-maximizing firms. Our approach generalizes the solution in Klette and Griliches (1996), which relies on the restrictive assumptions of constant elasticity demand and monopolistic competition, thus implying no variation in markups across firms. Beyond typical parametric critiques, these assumptions narrow the very scope of competition. More recently, De Loecker et al. (2016) uses observed output prices to control for unobserved input price biases, but the authors work in a rare setting where output price data are available. Foster et al. (2008) suggests the omitted price bias is so significant that it changes the observed correlation between physical productivity and revenue productivity. De Ridder et al. (0225) shows although we cannot recover precise markup levels from revenue data, we can imperfectly learn about their dispersion across firms and over time.

In their simulations—which focus on a repeated, static, Cournot model of competition that we adopt for our Monte Carlo experiments—they find a correlation of 0.9 between revenue-based and true log markups. However, when applying their method to actual data, the correlation drops to 0.3. In sum, the literature has shown that the omitted price bias is important, but has not yet offered a general solution for resolving it. We contribute by offering a promising avenue to study market power in a wide range of settings by tailoring the proxy variable and productivity process assumptions.

A related literature examines transmission bias: firms optimally choose their inputs as a function of their productivity, so productivity simultaneously determines both output and inputs (Marschak and Andrews, 1944). In an empirical analysis, regressing output on inputs fails to identify a firm's production function because the unobservable productivity term transmits into input decisions. The literature overcomes the resulting bias with assumptions about a firm's production or productivity process. Broadly, proxy variable estimators (Olley and Pakes, 1996; Levinsohn and Petrin, 2003; Ackerberg et al., 2015; Gandhi et al., 2020) assume that observable inputs can control for unobserved productivity, while dynamic panel estimators (Blundell and Bond, 2000) parametrically impose linearity in the productivity process. Our suggested method is in the tradition of proxy variable estimators meaning that it can allow for flexible modeling of the productivity process, and it allows researchers to relax the critical assumption of competitive output markets. Our work also relates to earlier attempts at forming control functions for revenue productivity directly, as in Flynn et al. (2019). We also derive results showing how to extend methods in the tradition of dynamic panel estimators.

Our main estimator builds on this line of work by addressing omitted price and transmission bias simultaneously. We do so by specifying a control function for markups, as opposed to the typical control function for physical productivity. We then directly estimate the firm's revenue elasticity by regressing its flexible input's log cost share of revenue on log inputs and fixed effects. Though similar in spirit to the first stage of Gandhi et al. (2020), we do not identify the flexible input's output elasticity from this share regression. Instead, we use it to partially identify the elasticity by identifying a combination of the elasticity and markup. We then substitute this composite into the revenue production equation to control for unobserved revenue productivity. Together, this method allows us to identify physical output elasticities relying only on revenue data.

Section 1 introduces a general production model with imperfect competition in product markets. Section 2 describes how to adapt the production model for imperfectly competitive environments, and how to use our proposed method to recover markups. Section 3 compares our estimator to existing production function estimation methods and models of imperfect competition, and Section 4 presents our simulation study. Section 5 concludes.

1 A Production Model with Markups

In this section, we present a structural model consistent with widely used approaches to firm-level production function and markup estimation. We consider the contemporary case of imperfect competition in the output market and cost minimization for a flexible input. This setup links the firm's markup to the flexible input's output elasticity and cost share of revenue (Hall, 1988; Basu and Fernald, 2002; Petrin and Sivadasan, 2013; De Loecker and Warzynski, 2012). After deriving the link in our model, we discuss identification and estimation problems when researchers only have data on revenues (not quantities). Throughout, we use uppercase to refer to levels of variables and lowercase to refer to logs of variables.

1.1 Production Technology, Productivity, and Pricing

The economy is populated by profit-maximizing firms with idiosyncratic productivity. Each firm generates final output combining a competitively supplied flexible input X_{it}, and a vector of nonflexible inputs, K_{it}, according to the following function:

$$Q_{it} = \Omega_{it}\mathcal{E}_{it}F(X_{it}, K_{it}) \tag{1}$$

with Ω_{it} representing firm i's physical productivity (TFPQ in the parlance of Foster et al. 2008) and $\mathcal{E}_{it}$ is an unexpected change in physical productivity that occurs after firms have made decisions (Olley and Pakes, 1996; Gandhi et al., 2020).[3] $F(\cdot)$ is the firm's production function.[4] X_{it} is flexible in the sense that it is both variable and static: firms may adjust it in each period after observing the realization of state variables such as productivity, and its choice has no dynamic implications. K_{it} is fixed, dynamic, or both.

[3]The only restrictions on the ex-post shock ε_{it} are that it is independent of the firm's choice of flexible inputs and that its expectation is finite and constant across firms and time, allowing us to treat $\mathbb{E}[\mathcal{E}_{it}]$ as a parameter to be estimated or normalized. In principle, even full independence could be relaxed; see Gandhi et al. (2020) for further discussion.

[4]The production function is concave and differentiable at every point. We omit the time subscript for sake of exposition, but in principle the production function can vary over time.

For exposition, we refer to X_{it} as intermediates, K_{it} as capital, and assume capital is fixed so that it may not respond to current period state variables.

Taking logs, where $\omega_{it} = \log(\Omega_{it})$ and $\varepsilon_{it} = \log(\mathcal{E}_{it})$, firm production is thus:

$$q_{it} = f(x_{it}, k_{it}) + \omega_{it} + \varepsilon_{it} \tag{2}$$

Firms may have market power in the output market, so that a single firm's residual demand curve may not be perfectly elastic.[5] Firms choose a price and quantity pair $\{P_{it}, Q_{it}\}$ along their residual demand curves, subject to technology constraints. Firms may choose to price above marginal cost, thus generating a markup. Markups are equilibrium objects that depend on demand, costs, market structure, and possibly other factors.

1.2 Optimization, Information, and Timing

Firms choose planned output through a broader problem, such as (but not necessarily) profit maximization. We focus here on the cost minimization problem that is nested within it. Specifically, firms form a plan for output based on their expectations of productivity and demand conditions. They then choose inputs to minimize the cost of producing this planned level of output. This nested approach is standard in the literature and provides the basis for deriving a tractable markup expression from the firm's first-order condition that also allows for stochastic errors when mapping the model to data.

The timing of the problem is as follows. At the beginning of the period, firms know their capital stock, K_{it}, and their physical productivity, Ω_{it}. First, each firm plans an optimal markup, $\mathcal{M}_{it}$, relying on the expectation about productivity, Ω_{it}, along with other information. Second, they choose the corresponding intermediate inputs, X_{it}, to implement these plans given their residual demand curve, technology, capital stock, and expected productivity. Finally, the transitory shock $\mathcal{E}_{it}$ is realized, production occurs, and markups are realized. Markups are generally not orthogonal to productivity, since the firm uses its expectation about productivity to plan a markup.

[5]The residual demand curve accounts for competitor responses and thus is a mix of demand and conduct (Baker and Bresnahan, 1988). For example, a monopolist's residual demand curve is the market demand curve, while in a perfectly competitive market, a firm's residual demand curve is flat. It includes information on own and cross-price elasticities as well as on other firms' decisions, and it defines an optimal markup level that profit-maximizing firms endogenously set.

Given this setup, the firm's cost minimization problem with time t information, $\mathcal{I}_{it}$, is

$$\begin{aligned} \min_{X_{it}} \quad & C_{it}X_{it} \\ \text{s.t.} \quad & \mathbb{E}[Q_{it}|\mathcal{I}_{it}] = \mathbb{E}[\Omega_{it}\mathcal{E}_{it}|\mathcal{I}_{it}]F(X_{it}, K_{it}) \end{aligned}$$

with C_{it} being the cost of the flexible inputs. The Lagrangian of the cost minimization is

$$\mathcal{L}_{it} = C_{it}X_{it} + \Lambda_{it}(\mathbb{E}[Q_{it}|\mathcal{I}_{it}] - \mathbb{E}[\Omega_{it}\mathcal{E}_{it}|\mathcal{I}_{it}]F(X_{it}, K_{it})$$

with Λ_{it} being the Lagrangian multiplier, thus the cost of relaxing the quantity constraint. In this respect, Λ_{it} represents firm i's marginal cost. The first-order condition for the flexible input X_{it} is

$$[X_{it}]: \quad C_{it} = \Lambda_{it}\mathbb{E}[\Omega_{it}\mathcal{E}_{it}|\mathcal{I}_{it}]\frac{\partial F(X_{it}, K_{it})}{\partial X_{it}}$$

Denote $\mathcal{M}_{it}$ as the markup, P_{it} as output price, we can manipulate the first-order condition to yield

$$\begin{aligned} \mathcal{M}_{it} &= \frac{P_{it}}{\Lambda_{it}} \\ &= \mathbb{E}[\Omega_{it}\mathcal{E}_{it}|\mathcal{I}_{it}]\frac{\partial F(X_{it}, K_{it})}{\partial X_{it}}\frac{P_{it}}{C_{it}} \\ &= \frac{\mathbb{E}[\Omega_{it}\mathcal{E}_{it}|\mathcal{I}_{it}]}{\Omega_{it}\mathcal{E}_{it}}\frac{\partial F(X_{it}, K_{it})}{\partial X_{it}}\frac{\Omega_{it}\mathcal{E}_{it}X_{it}}{Q_{it}}\frac{P_{it}Q_{it}}{C_{it}X_{it}} \\ &= \frac{\mathbb{E}[\Omega_{it}\mathcal{E}_{it}|\mathcal{I}_{it}]}{\Omega_{it}\mathcal{E}_{it}}\frac{\partial F(X_{it}, K_{it})}{\partial X_{it}}\frac{X_{it}}{F(X_{it}, K_{it})}\frac{P_{it}Q_{it}}{C_{it}X_{it}} \\ &= \frac{\mathbb{E}[\mathcal{E}_{it}]}{\mathcal{E}_{it}}f_{it}^{X}\frac{R_{it}}{C_{it}X_{it}} \end{aligned}$$

where the last line comes from the fact that the firm knows part of its productivity before its input choice, $\frac{\mathbb{E}[\Omega_{it}\mathcal{E}_{it}|\mathcal{I}_{it}]}{\Omega_{it}\mathcal{E}_{it}} = \frac{\Omega_{it}\mathbb{E}[\mathcal{E}_{it}]}{\Omega_{it}\mathcal{E}_{it}} = \frac{\mathbb{E}[\mathcal{E}_{it}]}{\mathcal{E}_{it}}$, the definition of the output elasticity, $f_{it}^{X} = \frac{\partial F(X_{it}, K_{it})}{\partial X_{it}}\frac{X_{it}}{F(X_{it}, K_{it})}$, and revenues, $R_{it} = P_{it}Q_{it}$.

Defining now the expectation term as $b = \log \mathbb{E}[\mathcal{E}_{it}]$, and writing this result in logs gives us

$$\mu_{it} = \log f_{it}^{X} + r_{it} - c_{it}x_{it} + b - \varepsilon_{it} \tag{3}$$

That is, the log markup, μ_{it}, is the sum of the intermediate input's output elasticity, f_{it}^X, the intermediate input's log inverse cost share of revenue, $r_{it} - c_{it}x_{it}$, and the unplanned productivity term, $b - \varepsilon_{it}$. Because our derivation combines the ex-post price with the ex-ante marginal cost from the firm's planning problem, the resulting markup term μ_{it} correctly captures both the firm's planned markup and the price response to the ex-post shock. Equation 3 is the key relation that we use to estimate markups with our method.

2 Measuring Markups with Revenue Data

In this section, we describe our novel approach to estimate markups with revenue data and discuss the underlying assumptions. The aim is to identify physical output elasticities (and error terms) to measure markups from the first-order condition of the per-period static cost minimization problem.

We assume we observe data for a panel of firms $i = 1, 2, ..., N$ over periods $t = 1, 2, \ldots, T$. We let the data take a short panel form: the number of firms grows large for a fixed T. For each firm, we observe revenue $R_{it} = P_{it}Q_{it}$, expenditures on a competitively supplied flexible input X_{it} with cost C_{it}, and a vector of nonflexible inputs K_{it}. In practice, the choice of X_{it} might be energy, materials, labor, or some combination thereof. We recommend choosing the input that is most likely to satisfy the competitively supplied and flexible assumptions based on the empirical setting.

2.1 Problems with Revenue Data

Recovering markups from equation 3 requires an estimate of the output elasticity f_{it}^X. However, we cannot simply regress revenue on inputs to get this estimate for two critical reasons. The first reason is the omitted price bias emphasized in Klette and Griliches (1996): higher markups induce the firm to decrease input use, which increases prices and thus (all else equal) revenues. The second reason is the transmission bias emphasized in Marschak and Andrews (1944): higher physical productivity induces the firm to increase input use, which increases output and thus (all else equal) revenues.

One approach in the literature is to ignore the distinction between revenues and quantities, and estimating firm production function using common estimation techniques specifically designed to address the transmission bias. Specifically, the control function approach pioneered by Olley and Pakes (1996) and the dynamic panel methods proposed by Blundell and Bond (2000) are the most common methods. However, these existing methods to estimate production function parameters are designed for perfectly competitive environments and rely on physical quantity data. An empirical strategy that treats the revenue production function estimates as though they are physical production function estimates confounds demand with productivity (Foster et al., 2008) and results in markup estimates devoid of empirical content (Bond et al., 2021). Hence, we cannot use them in this context. Our method tackles this issue directly by including a control function approach in the spirit Olley and Pakes (1996) and allowing for imperfect competition. It requires only revenue data and is a generalization of the Klette and Griliches (1996)'s correction.

2.2 Our Approach: Controlling for Markups

At the core of our method there is the law of motion of productivity and a control function to proxy for markups. The former is common to the literature both for control function approach, that assumes a Markov process for physical productivity, and for dynamic panel methods, that assumes a linear process. Given that our aim is to provide an estimation method using revenue data, we focus instead on revenue productivity. The latter relates to the approach used in the control function literature and brings insights also from the share regression proposed by Gandhi et al. (2020). Given that our approach deals with markup estimation using only revenue data, we propose to control for markups directly rather than only for unobserved productivity.

Because firm-level prices are typically unobserved, the literature since (and including) Olley and Pakes (1996) has had to model the evolution of a proxy for the physical productivity primitive, ω_{it}. In revenue-based settings, the only commonly available proxy is revenue productivity, $\nu_{it} = p_{it} + \omega_{it}$. While treating this hybrid object as a state variable is a plausibly strong assumption given the endogeneity of prices, it is a standard and necessary approximation used in virtually all empirical applications, including the seminal works of Olley and Pakes (1996) (see footnote 3), Levinsohn and Petrin (2003), De Loecker (2011), and De Loecker and Warzynski (2012). This approach can be justified by the robust empirical finding that revenue productivity is highly persistent and well-approximated by a low-order autoregressive process (Foster et al., 2008).

Assumption 1 (The Stochastic Process of Revenue Productivity) *Let revenue productivity $\nu_{it} = p_{it} + \omega_{it}$ (the sum of prices p_{it} and physical productivity ω_{it}) follow a Markov process with additively separable mean-zero shocks η_{it}. Hence, $\mathcal{P}(\nu_{it}|\mathcal{I}_{it-1}) = \mathcal{P}(\nu_{it}|\nu_{it-1})$,* *and* $\mathbb{E}[\eta_{it}|\mathcal{I}_{it-1}] = 0$.

Following Assumption 1, we can write the stochastic process of revenues productivity as

$$\nu_{it} = g(\nu_{it-1}) + \eta_{it} \tag{4}$$

for some continuous function $g(\nu_{it-1}) = \mathbb{E}[\nu_{it}|\nu_{it-1}]$.

We discuss the microfoundations for the Markov assumption in Appendix A. We emphasize this is a standard maintained assumption in the literature and generalizing or improving it isn't our focus; our primary contribution is the development of a novel control function approach for markups to address the problem of how the omitted price bias propagates into production elasticities.

Now, we write firm revenues as

$$\begin{aligned} r_{it} &= p_{it} + q_{it} \\ &= p_{it} + f(x_{it}, k_{it}) + \omega_{it} + \varepsilon_{it} \\ &= f(x_{it}, k_{it}) + \nu_{it} + \varepsilon_{it} \end{aligned}$$

Where the second line comes from the definition of $f(x_{it}, k_{it})$, and the third line comes from the definition of the revenue productivity ν_{it}. Therefore, we can express the revenue production function as

$$r_{it} = f(x_{it}, k_{it}) + g(\nu_{it-1}) + \eta_{it} + \varepsilon_{it} \tag{5}$$

The latter can be used to estimate production function parameters leveraging assumptions on firms' information and timing to construct moment conditions. However, we first need to recover a measure of revenue productivity to operationalize it. To do so, we use our structural framework. Specifically, we interpret the first-order condition 3 as a markup function and use it to identify revenue elasticities.

Assumption 2 (The Markup Control Function.) *Let markups be a function of inputs, firm and time fixed effects ι_i and τ_t, and a vector of other firm-time varying observables relevant in determining markups $\mathbf{D}_{it}$*

$$\mu_{it} = h(x_{it}, k_{it}, \iota_i, \tau_t, \mathbf{D}_{it}) \tag{6}$$

We return to examples of possible observables in Section 3. One appealing feature is that in general models of competition, higher planned markups induce lower chosen intermediates; this suggests a straightforward way to microfound the markup control function in an input demand equation, as in Olley and Pakes (1996). Rewrite the first-order condition 3 as

$$c_{it}x_{it} - r_{it} = \log f_{it}^X - \mu_{it} + b - \varepsilon_{it} \tag{7}$$

$$= \log f_{it}^X - h(x_{it}, k_{it}, \iota_i, \tau_t, \mathbf{D}_{it}) + b - \varepsilon_{it} \tag{8}$$

The left-hand side is the intermediates log cost share of revenue. The term $\log f_{it}^X - \mu_{it}$ on the right-hand side is the log revenue elasticity with respect to input x_{it}, a mix of supply and demand parameters. As productivity is Hicks-neutral, the log elasticity term f_{it}^X is a function of inputs only: $f_{it}^X = f^X(x_{it}, k_{it})$. Combining the revenue elasticity terms into a single function $s(x_{it}, k_{it}, \iota_i, \tau_t, \mathbf{D}_{it}) = \log f^X(x_{it}, k_{it}) - h(x_{it}, k_{it}, \iota_i, \tau_t, \mathbf{D}_{it})$, our first stage estimating equation becomes:

$$c_{it}x_{it} - r_{it} = s(x_{it}, k_{it}, \iota_i, \tau_t, \mathbf{D}_{it}) + b - \varepsilon_{it} \tag{9}$$

To operationalize this equation, a researcher may (nonparametrically) regress the intermediates share of revenues on inputs, fixed effects, and a vector of markup determinants $\mathbf{D}_{it}$ to get an estimate of the revenue elasticity $\hat{s}_{it} = \widehat{\log f_{it}^X - \mu_{it}}$. This share regression estimates the specified markup control function: it describes the determinants of wedges between prices and marginal costs, and is similar to the share regression in Gandhi, Navarro, and Rivers (2020), but adapted for cases of imperfect competition with unobserved prices.

Importantly, it also recovers an estimate of the error $\hat{\varepsilon}_{it}$, and therefore $\hat{b}$. Estimating $\hat{\varepsilon}_{it}$ is the primary function of the first stage of proxy variable estimators (Olley and Pakes, 1996; Levinsohn and Petrin, 2003; Ackerberg et al., 2015). Estimating it here allows us to replace the physical productivity control function assumption of these models with a markup control function assumption. However, the share regression alone cannot separate the impact of markups from output elasticities, since it still contains the unknown f_{it}^X. We now return to the revenue production function to identify the physical elasticity f_{it}^X.

Specifically, we rewrite the output from estimating the first-order condition to isolate revenue productivity. The first stage gives the relationship:

$$c_{it}x_{it} - r_{it} = \hat{s}_{it} + \hat{b} - \hat{\varepsilon}_{it}$$

By definition, revenue is also $r_{it} = f(x_{it}, k_{it}) + \nu_{it} + \varepsilon_{it}$. Substituting this into the equation above and rearranging for ν_{it} yields:

$$\nu_{it} = c_{it}x_{it} - f(x_{it}, k_{it}) - \hat{s}_{it} - \hat{b} + (\hat{\varepsilon}_{it} - \varepsilon_{it})$$

As in any two-step estimator, we replace the unobserved true shock ε_{it} with our consistent first-stage estimate $\hat{\varepsilon}_{it}$ to form our operational proxy for revenue productivity for the second stage:[6]

$$\hat{\nu}_{it} = c_{it}x_{it} - f(x_{it}, k_{it}) - \hat{s}_{it} - \hat{b} \tag{10}$$

And we express the revenue production function as

$$r_{it} = f(x_{it}, k_{it}) + g(\hat{\nu}_{it-1}) + \eta_{it} + \hat{\varepsilon}_{it}. \tag{11}$$

Hence, we can directly estimate production function parameters from this equation.

2.3 Identifying Returns to Scale and Markups

A well-known limitation in the production function estimation literature is that it is exceedingly difficult to separately identify the flexible input elasticity and the returns to scale applying the proxy method on common datasets. In theory, Gandhi et al. (2020) show sufficient time-series variation in relative input prices can achieve identification, but also conclude this approach is largely impractical with common production datasets. Their Monte Carlo simulations, calibrated to match real-world data, show standard proxy estimators are significantly biased with wide standard errors even with long panels. Only when the variance of the input price innovation is amplified to ten times the empirically observed level do the estimates begin to converge. The practical challenge of using price variation is a long-standing issue in the literature, which has identified several difficulties,

[6] This procedure appeals to the properties of two-step GMM estimators, where consistency is established by the Law of Large Numbers and Slutsky's Theorem. While this yields consistent parameter estimates, correct inference requires adjusting the second-stage standard errors for the "generated regressor" problem, often via bootstrap or a corrected variance-covariance matrix.

including the lack of firm-specific price data and the concern that observed price variation may reflect unobserved quality differences rather than simple market-price differences (Griliches and Mairesse, 1995; Ackerberg et al., 2007). While recent work shows carefully employed price instruments can aid identification, this information is often unavailable (Doraszelski and Jaumandreu, 2013, 2018). Absent this additional source of identification, there exists a continuum of observationally equivalent production functions that satisfy the identification restrictions imposed in the proxy approach. Observing serially correlated, firm-specific input prices might solve this non-identification problem in specific cases, or imposing additional restrictions on the production function or productivity process.[7]

To avoid this non-identification issue, we follow the solution proposed by Flynn et al. (2019) which does not require additional data or assumptions on the evolution of the production function or productivity process. Specifically, they show it's possible to identify markups by setting the degree of returns to scale ex-ante. Their findings show that this approach drastically reduces the bias resulting from the non-identification result by up to twenty times. Therefore, we assume firm production function has constant returns to scale. Our estimators require known or externally identifiable returns to scale—constant or otherwise. Researchers could estimate the scale elasticity using external methods such as those in Basu and Fernald (1997) or Syverson (2004)—studies which also provide empirical evidence that constant returns is a reasonable approximation in many settings. Alternatively, as Flynn et al. (2019) show, it is possible to identify the scale elasticity structurally even when it is a function of non-flexible inputs like capital intensity. We proceed with constant returns for clarity, though our framework generalizes to any known or externally identified returns to scale.[8]

2.4 Estimation

Here we describe the steps to implement our suggested estimator.

[7] Flynn et al. (2019) provide a detailed discussion on the precise conditions under which the production function is point identified in the presence of markups.

[8] Our approach is in the tradition of Olley and Pakes (1996) and is distinct from the cost-share method. While markups can in principle be computed directly from cost shares under constant returns, this requires observing the full economic cost of all inputs—a demanding requirement when labor markets are imperfectly competitive or capital has adjustment costs. These measurement challenges were central to the intellectual history of production function estimation, motivating the field's shift away from cost-share methods toward the proxy variable approaches we build upon here (Griliches and Mairesse, 1995).

1. Estimate the share regression to recover revenue elasticities and firms' expectations on productivity. In practice, regress the intermediates log cost share of revenues ($c_{it}x_{it} - r_{it}$) on log inputs and markup determinants. The latter can include firm and time fixed effects, along with other observed variables included in the vector $\mathbf{D}_{it}$.[9] Use the predicted residual, $\hat{\varepsilon}_{it}$, to form $\hat{b} = \log \hat{\mathbb{E}}[\exp(\hat{\varepsilon}_{it})]$, and therefore recover $\hat{s}_{it} = \log \widehat{f_{it}^{X} -} \mu_{it} = c_{it}x_{it} - r_{it} - \hat{b} + \hat{\varepsilon}_{it}$.

2. Specify functional forms for the production function, $f(x_{it}, k_{it})$, and the Markov process, $g(\nu_{it-1})$, such as Cobb-Douglas technology with AR(1) productivity, or translog technology with quadratic Markov productivity. Note that, as explained above, $f(x_{it}, k_{it})$ must satisfy a scale elasticity assumption, such as constant returns to scale. In practice, the main concern is how to model the intermediates elasticity, because this term directly affects the markup as seen from the first-order condition 3.

3. Define the revenue productivity innovation, $\hat{\eta}_{it}$. First, define the proxy for revenue productivity using the results from the first stage and the definition of the production function:

$$\hat{\nu}_{it} = c_{it}x_{it} - f(x_{it}, k_{it}) - \hat{s}_{it} - \hat{b}$$

   The innovation is the part of current revenue productivity not predicted by its past, which we construct as:

$$\hat{\eta}_{it} = \hat{\nu}_{it} - g(\hat{\nu}_{it-1})$$

   This shock will be a function of the parameters in $f(x_{it}, k_{it})$ and $g(\nu_{it-1})$.

4. Finally, leverage the stochastic revenue productivity process to estimate the parameters of $f(x_{it}, k_{it})$ and $g(\nu_{it-1})$. Specifically, use the moment conditions formed by

$$\mathbb{E}[\hat{\eta}_{it} | k_{it}, \hat{s}_{it-1}] = 0$$

The specific moments depend upon the specifications of $f(x_{it}, k_{it})$ and $g(\nu_{it-1})$. For a translog production function, we add moments including squares and interactions of the

[9] Researchers can specify the information in $\mathbf{D}_{it}$ in line with their models. Examples include market shares, exporter dummies, location dummies, etc.

nonflexible inputs. We augment the revenue productivity process by adding lags and interactions of the revenue share vector, $\hat{s}_{it}$. For example, $\hat{s}_{it}^2$ controls for a second-order process. By construction, these additional moments are orthogonal to revenue productivity innovations. Since costs, demand, and conduct determine markups and co-evolve with the revenue productivity process, innovations to this process are orthogonal to any functions of costs, demand, or conduct embedded in the $\hat{s}_{it}$ vector. This two-step approach can also be implemented as a single-step GMM problem by jointly minimizing the residuals in the share regression with the moment conditions from the second step (Wooldridge, 2009).

2.5 Discussion of Identification Strategy

Production-based markup estimation requires either strong parametric assumptions about demand or semi-parametric exclusion restrictions. The literature since Klette and Griliches (1996) has often chosen the former, achieving identification by assuming constant elasticity of substitution (CES) demand with monopolistic competition. Under these assumptions, all firms face the same demand elasticity, and markups become a simple function of this single parameter—eliminating the heterogeneity in market power that researchers often seek to measure. This parametric approach undermines a key advantage of production-based methods, which is their ability to remain agnostic about market structure and conduct. We take the second approach. We generalize the control function logic of Olley and Pakes (1996) to settings with imperfect competition, using a flexible control function for markups. This approach nests the restrictive CES case (where our control function would reduce to an industry fixed effect) but allows for the richer heterogeneity in firm conduct that we observe in most real-world markets.

Our approach requires an exclusion restriction: there must be some source of variation in the markup control function, $h(\cdot)$, that does not directly affect the production technology, $f(\cdot)$. To see why, consider the revenue elasticity from our first-stage equation: $s(x_{it}, k_{it}, \iota_i, \tau_t, \mathbf{D}_{it}) = \log f^X(x_{it}, k_{it}) - h(x_{it}, k_{it}, \iota_i, \tau_t, \mathbf{D}_{it})$. The core identification challenge formalized by Gandhi et al. (2020) is that the observable revenue elasticity on the left-hand side is a function of two unobservables on the right-hand side: the physical output elasticity and the markup (embedded in $h(\cdot)$). Without additional information, there is a continuum of function pairs $(\log f^X, h)$ that are observationally equivalent.

Our exclusion restriction breaks this underidentification by ensuring that some arguments of $h(\cdot)$—namely ι_i, τ_t, or elements of $\mathbf{D}_{it}$—do not enter $f(\cdot)$. To see how this works, denote inputs as $W_{it} = (x_{it}, k_{it})$ and excluded variables as Z_{it} (which could be firm effects, time effects, or demand proxies). When we observe how revenue shares vary with Z_{it} while holding inputs constant, we learn about the markup function. Specifically, for any two values of the excluded variables, the difference in conditional expectations

$$\mathbb{E}[s_{it} \mid W_{it}, Z_{it} = z] - \mathbb{E}[s_{it} \mid W_{it}, Z_{it} = z'] = -\big(h(W_{it}, z) - h(W_{it}, z')\big)$$

identifies the markup function up to a normalization. Once we pin down $h(\cdot)$ by normalizing it at some baseline (say, $h(W_{it}, z_0) = 0$), we can recover the production function from the first-stage regression. Of course, this requires that Z_{it} actually varies conditional on inputs and that we impose a scale restriction like constant returns to avoid the flexible-input elasticity issue identified by Gandhi et al. (2020) and Flynn et al. (2019). But these are economic restrictions we can defend based on our institutional setting.

The identifying variation can come from three sources: firm fixed effects (ι_i), time dummies (τ_t), or observable firm characteristics in the vector $\mathbf{D}_{it}$. The choice depends on the economic setting. If firm fixed effects primarily capture persistent technical capabilities (some firms have better engineers), they belong in the production function and cannot serve as exclusions. But if these fixed effects mainly reflect persistent differences in market power—brand value, prime locations, established customer relationships—they provide valid identifying variation. These sources of market power affect the price-cost margin without changing how inputs transform into output. Similarly, time effects τ_t work as exclusions when they capture demand-side phenomena like business cycles that affect willingness to pay, rather than technology shocks that would alter production capabilities.

The researcher must justify these modeling choices based on institutional knowledge. In consumer goods industries, brand effects likely create persistent markup differences unrelated to production technology. In commodity industries, firm fixed effects might primarily reflect differences in extraction or processing efficiency and would need to be included in the production function. The key is being explicit about what drives the variation and defending why it belongs in markups rather than technology.

The Markov process for revenue productivity ν_{it} serves a specific statistical purpose. It isolates unexpected productivity shocks from predictable variation, ensuring that the innovations η_{it} are orthogonal to predetermined variables. This isn't a structural assumption about productivity evolution—it's a forecasting device that purges variation firms could anticipate when making input decisions. The actual separation between markups and physical productivity comes from our first-stage share regression, where the control function $h(\cdot)$ absorbs the markup variation. Once we control for markups in the first stage, the second-stage moments cleanly identify the production function parameters using only these orthogonal innovations.

Observable controls for $\mathbf{D}_{it}$ should be variables that plausibly affect firm conduct or demand conditions without directly entering the production function. Market concentration measures affect strategic interactions and pricing power without changing how a firm transforms inputs into output. Export status can shift the demand elasticity a firm faces without altering its production technology. Advertising expenditure creates market power through brand awareness rather than production efficiency. The panel structure offers additional opportunities: lagged market shares or competitors' characteristics can shift the residual demand curve without affecting the firm's own production capabilities.

Market interactions provide another source of identifying variation. In our model from Section 1, firms choose their optimal markup based on anticipated productivity and residual demand conditions before selecting inputs. Variables that shift the residual demand curve—lagged competitors' productivity realizations, predetermined regulatory changes affecting rivals—can influence a firm's input choices through their effect on expected market conditions, as long as they don't directly affect the firm's production technology.

A remaining concern is that markups depend on unobserved demand heterogeneity. Our markup control function $h(\cdot)$ addresses this by flexibly capturing how observables correlate with demand conditions and their effect on markups. By conditioning on firm fixed effects, time effects, and demand proxies in the first-stage share equation, we absorb much of this heterogeneity. The resulting productivity innovation is then orthogonal to predetermined variables, breaking the simultaneity between input choices and unobserved shocks. If demand shocks correlate with inputs in ways our control function cannot capture, some bias remains—a challenge shared by all production-based markup estimation methods (De Loecker and Warzynski, 2012; Doraszelski and Jaumandreu, 2018). Researchers can assess validity by including additional demand proxies and testing the orthogonality of their moment conditions when the model is overidentified.

We trade restrictive parametric assumptions about demand (like CES) for more flexible but still economically motivated exclusion restrictions. Rather than assuming all firms face identical demand curves, we allow heterogeneous market power while requiring researchers to specify and defend which variables affect markups versus technology. This approach extends the control function tradition in production function estimation to imperfectly competitive settings, providing a practical middle ground between fully parametric and fully nonparametric approaches.

3 Comparison to Related Literature

In this section, we detail how our estimator compares to commonly used methods to estimate production functions and then illustrate how to apply it to models with imperfect competition.

3.1 Comparison to Existing Competitive Models

We first discuss the proxy variable class, then the dynamic panel class. Both methods form moments from information and timing assumptions about the productivity process, but differ in other assumptions to achieve identification.

3.1.1 Proxy Variable Estimators

The proxy model of production (Olley and Pakes, 1996; Levinsohn and Petrin, 2003; Ackerberg, Caves, and Frazer, 2015; Gandhi, Navarro, and Rivers, 2020) identifies production functions using assumptions about the unobservable state. Proxy models assume that productivity is Markovian and that productivity can be written as a control function of observables. These two assumptions allow one to use lagged inputs to control for current productivity, solving the transmission bias. More formally, the assumptions of the proxy model are:

Proxy Variable Assumption 1: Physical productivity follows a first-order Markov process: $\omega_{it} = g(\omega_{it-1}) + \eta_{it}$.

Proxy Variable Assumption 2: Physical productivity is a control function of observables: $\omega_{it} = m(x_{it}, k_{it})$.[10]

Proxy Variable Assumption 2 ensures that past inputs can proxy for current productivity

[10] This is the standard assumption used in the proxy variable literature and represents the most direct application of the control function approach. It can be extended to include richer sets of controls—such as input prices or market shares.

through the productivity process. Then substitution yields

$$\begin{aligned} q_{it} &= f(x_{it}, k_{it}) + \omega_{it} + \varepsilon_{it} \\ &= f(x_{it}, k_{it}) + g(\omega_{it-1}) + \eta_{it} + \varepsilon_{it} \\ &= f(x_{it}, k_{it}) + g(m(x_{it-1}, k_{it-1})) + \eta_{it} + \varepsilon_{it} \end{aligned}$$

Identification proceeds by forming moments with the composite error term $\eta_{it} + \varepsilon_{it}$.

With revenue data, this derivation becomes

$$\begin{aligned} r_{it} &= f(x_{it}, k_{it}) + p_{it} + \omega_{it} + \varepsilon_{it} \\ &= f(x_{it}, k_{it}) + p_{it} + g(\omega_{it-1}) + \eta_{it} + \varepsilon_{it} \\ &= f(x_{it}, k_{it}) + p_{it} + g(m(x_{it-1}, k_{it-1})) + \eta_{it} + \varepsilon_{it} \end{aligned}$$

The appearance of prices on the right-hand side is the origin of the omitted price bias terminology.

Our approach is a direct modification of these two assumptions. For the first, we assume revenue productivity follows a first-order Markov process, instead of physical productivity. This assumption is consistent with Foster, Haltiwanger, and Syverson (2008), which finds that revenue productivity exhibits similar levels of persistence as physical productivity. It is also implicit in existing work that estimates revenue production functions, or explicit in existing work that attempts to correct for the omitted price bias with richer data or stronger parametric structure (Klette and Griliches 1996, De Loecker 2011, De Loecker and Warzynski 2012). For the second, we assume markups are a control function of observables, instead of physical productivity. Although markups and productivity are both unobservable, researchers typically impose some markup-setting process or rule in modeling (for example, monopolistic competitors facing a constant price elasticity of substitution demand system). Therefore, we view our assumption on markups as less limiting. More generally, one can think of our model as a version of a proxy variable estimator, in which we are proxying for markups instead of proxying for productivity. It is built to estimate markups, and also allows us to relax some of the physical productivity assumptions.

3.1.2 Dynamic Panel Estimators

The dynamic panel approach pioneered by Blundell and Bond (2000) is a commonly used alternative to proxy variable approaches. Dynamic panel models maintain the same basic structure of production. However, they impose linearity on the productivity process.

Dynamic Panel Assumption: Physical productivity ω_{it} follows an AR(1) process: $\omega_{it} = \rho\omega_{it-1} + \eta_{it}$.

The unobserved term η_{it} is uncorrelated with all past and future input choices. Then differencing the production function yields

$$\begin{aligned} q_{it} - \rho q_{it-1} &= f(x_{it}, k_{it}) - f(x_{it-1}, k_{it-1}) + \omega_{it} - \rho\omega_{it-1} + \varepsilon_{it} - \rho\varepsilon_{it-1} \\ q_{it} - \rho q_{it-1} &= f(x_{it}, k_{it}) - f(x_{it-1}, k_{it-1}) + \eta_{it} + \varepsilon_{it} - \rho\varepsilon_{it-1} \\ q_{it} &= \rho q_{it-1} + f(x_{it}, k_{it}) - f(x_{it-1}, k_{it-1}) + \eta_{it} + \varepsilon_{it} - \rho\varepsilon_{it-1} \end{aligned}$$

The appearance of lagged quantities on the right-hand side is the origin of the dynamic panel terminology. Identification proceeds by forming moments with the composite error term $\eta_{it} + \varepsilon_{it} - \rho\varepsilon_{it-1}$.

With revenue data, this derivation becomes

$$\begin{aligned} r_{it} - \rho r_{it-1} &= f(x_{it}, k_{it}) - \rho f(x_{it-1}, k_{it-1}) + p_{it} - \rho p_{it-1} + \omega_{it} - \rho\omega_{it-1} + \varepsilon_{it} - \rho\varepsilon_{it-1} \\ r_{it} - \rho r_{it-1} &= f(x_{it}, k_{it}) - \rho f(x_{it-1}, k_{it-1}) + p_{it} - \rho p_{it-1} + \eta_{it} + \varepsilon_{it} - \rho\varepsilon_{it-1} \\ r_{it} &= \rho r_{it-1} + f(x_{it}, k_{it}) - \rho f(x_{it-1}, k_{it-1}) + p_{it} - \rho p_{it-1} + \eta_{it} + \varepsilon_{it} - \rho\varepsilon_{it-1} \end{aligned}$$

Absent additional assumptions, we cannot proceed further without price data. However, combining this equation with a version of our earlier Markovian revenue productivity assumption ($\nu_{it} = g(\nu_{it-1}) + \eta_{it}$), we can make progress. Suppose our earlier assumption of Markovian revenue productivity holds, and further suppose it is linear, so that revenue productivity follows an AR(1) process: $\nu_{it} = \rho\nu_{it-1} + \eta_{it}$. We can then proceed as follows:

$$\begin{aligned} r_{it} - \rho r_{it-1} &= f(x_{it}, k_{it}) - \rho f(x_{it-1}, k_{it-1}) + p_{it} - \rho p_{it-1} + \omega_{it} - \rho\omega_{it-1} + \varepsilon_{it} - \rho\varepsilon_{it-1} \\ r_{it} - \rho r_{it-1} &= f(x_{it}, k_{it}) - \rho f(x_{it-1}, k_{it-1}) + \eta_{it} + \varepsilon_{it} - \rho\varepsilon_{it-1} \\ r_{it} &= \rho r_{it-1} + f(x_{it}, k_{it}) - \rho f(x_{it-1}, k_{it-1}) + \eta_{it} + \varepsilon_{it} - \rho\varepsilon_{it-1} \end{aligned}$$

This derivation does not require a markup control function, and therefore suggests a similar trade-off as in the competitive case: researchers may impose more structure (linear) on the productivity process to avoid assumptions that observables (inputs) and fixed effects span unobservables (markups).

3.2 Comparison to Imperfect Competition Models

Our solution generalizes much of the existing literature. In our earlier setup, we showed that our markup function identifies markups so long as $\mathcal{M}_{it}$ is determined partially independently from inputs. Now, we offer several commonly used parametric examples of markup functions.

3.2.1 Constant Markups: Monopolistic Competition and CES Demand

Suppose that firms are monopolistic competitors facing a constant price elasticity of substitution demand system. Suppose further that these firms compete in a number of industries j. In this environment, a firm i in industry j faces a demand curve given by $Q_{it} = Q_{jt}(\frac{P_{it}}{P_{jt}})^{-\sigma_j}$. Firm optimization implies that markups are constant within industries and given by $\mathcal{M}_{it} = \mathcal{M}_{jt} = \frac{\sigma_j}{\sigma_j - 1}$.

In the context of our markup function, assuming monopolistic competition with CES demand implies that $\mu_{it} = \mu_{jt}$: markups are fully determined by a constant within industry. One may recover markups and elasticities by simply including an industry fixed effect in the share regression.

This was originally noted in Klette and Griliches (1996). This paper shows that, in this case, one can estimate (industry-level) production functions by including controls for the industry quantity production. Klette and Griliches (1996) uses the estimating equation,

$$r_{it} = \beta_0 + \frac{\sigma_j - 1}{\sigma_j}(\beta_X x_{it} + \beta_K k_{it}) - \frac{1}{\sigma_j} q_{jt} + \nu_{it}.$$

Here, q_j is an industry-level price index which comes from the monopolistically competitive environment. The residual term ν_{it} collects firm-level productivity, specifically $\omega_{it} + \varepsilon_{it}$, scaled by the markup factor $\frac{\sigma_j - 1}{\sigma_j}$, consistent with the CES demand structure and constant-markup pricing rule. Estimation can then proceed using (observed) industry-level output to proxy for price variation across firms.

Gandhi et al. (2020) extends the Klette and Griliches (1996) model to allow for time-varying price elasticities of demand. We can easily accommodate such an extension by including a time fixed effect in the share regression.

In sum, our approach is a generalization of the Klette and Griliches (1996) correction that allows for more conduct and demand structures than monopolistic competition with CES demand. This generalization is especially important as it allows credible scaling of production-based markup estimators across many industries and time periods, where these earlier assumptions might be considerably off.

3.2.2 Variable Markups: Oligopolistic Competition and CES Demand

Suppose now that firms are oligopolistic competitors facing a variable price elasticity of substitution demand system. This is the case when demand is of the nested CES form: the final good is a CES aggregate from a continuum of sectors and each sector good is a CES aggregate of differentiated products (Atkeson and Burstein, 2008). In this setting, the elasticity of demand is a combination of the elasticity of substitution across sectors and within own sector, weighed by the market share of each firm.[11]

In Cournot competition, markups are a function of firm-specific demand elasticity: $\mu_{it} = \frac{\epsilon_{it}}{\epsilon_{it}+1}$, with the latter being

$$\epsilon_{it} = \left[\frac{1}{\eta}(1 - s_{it}) + \frac{1}{\theta}s_{it}\right]^{-1} \tag{12}$$

with s_{it} being the market share of firm i, and θ and η the elasticities of substitution across and within sector, respectively.

In the context of our markup function, assuming oligopolistic competition with nested CES demand implies that markups are fully determined by a combination of two constants and firms' market shares. One may recover markups and output elasticities by controlling for fixed effects and industry market shares in the share regression.

3.2.3 Variable Markups: Firm-Time Characteristics

We explored two specific models of imperfect competition so far and showed how our method can recover markups from these cases. In general, our method is flexible enough

[11] If firms compete in quantities, such combination takes the form of a harmonic mean. If firms compete in prices, it becomes a weighted mean.

to allow the researcher to adapt it to recover markups depending on the framework analyzed and the data available. If the researcher has data on firm-level characteristics which determine markups, then our model identifies market power and production elasticities by putting these into $\mathbf{D}_{it}$. For instance, advertising, managerial practices, research and development (Doraszelski and Jaumandreu, 2013), export status De Loecker and Warzynski (2012), or product mix De Loecker (2011) all might be associated with markups, and therefore added to the share regression. Any other observable market characteristics that vary by firm-time, such as combining geographic variation with consumer income variation, may also be added, depending on the researcher's model.

If revenue market shares determine markups, then in our model we have $\mathbf{D}_{it} = \frac{R_{it}}{R_{jt}}$, where R_{jt} is industry revenues. Researchers may define industries j in whatever way appropriate, such as common industry codes, or broadly or narrowly defined product markets. Then the markup control function may again be used to identify markups and output elasticities by adding market shares to each. Unlike models such as nested CES, this approach does not impose a parametric relationship between market shares and markups. Rather, the data determine the relationship.

In microfounding this control variable, typical models of competition such as homogenous product Cournot and differentiated product Bertrand result in a mapping from markups to quantity market shares, not revenue market shares. Of course, if we had quantity market shares, we would have quantity information, which would obviate the need for revenue data corrections.

4 Estimator Evaluation in Simulated Data

We evaluate our revenue-based estimator in simulated data environments with heterogeneous market conducts, structures and demand systems. We simulate data from four partial equilibrium models that share the same supply side but differ in demand structure. The first model features Cournot (quantity) competition, while the other three feature Bertrand (price) competition, allowing us to isolate the role of demand heterogeneity in driving estimation bias. For each model, we estimate markups using our method and compare the resulting elasticities with those obtained from three alternative estimators. Importantly, we intentionally constrain our method to use only minimal controls to establish conservative performance benchmarks, while noting that the framework's flexibility allows for substantial improvements when researchers can incorporate additional market-specific information.

4.1 Data Generating Processes

All models feature a fixed number of single-product firms in multiple industries producing differentiated goods. Firms produce output y_{it} with a Cobb-Douglas production function with idiosyncratic productivity ω_{it}, a flexible input x_{it}, and a fixed input k_{it},

$$y_{it} = \omega_{it} x_{it}^{\alpha} k_{it}^{1-\alpha}.$$

Firms observe their idiosyncratic productivity prior to choosing input levels, and optimally adjust the flexible input x_{it} each period to maximize profits. Productivity and capital evolve according to AR(1) processes following De Ridder et al. (0225). We simulate fifty periods. In the first model, we simulate 1440 firms across 180 sectors; in the others, 800 firms across 80 sectors. Further details on model characteristics, solution and calibration appear in Appendix B.

The models differ in their demand structure.

4.1.1 Nested CES

In the first framework, based on Atkeson and Burstein (2008), firms engage in Cournot competition with a nested CES demand structure. The output of N_s firms in each sector s is aggregated into sectoral goods y_s, which are then aggregated into final good via CES aggregators:

$$y_s = \left[\sum_{i=1}^{N_s} y_{is}^{\frac{\eta-1}{\eta}} \right]^{\frac{\eta}{\eta-1}}, \qquad y = \left[\int_0^1 y_s^{\frac{\theta-1}{\theta}} ds \right]^{\frac{\theta}{\theta-1}},$$

with $\eta = 10$ and $\theta = 1.1$ governing within- and across-sector substitution, respectively.

4.1.2 Logit

In the Logit framework, firms compete in prices, and a representative consumer chooses among J differentiated products. The indirect utility of consumer i from product j in sector s is

$$u_{ijs} = \phi p_{js} + x_{js}\gamma + \xi_{js} + \varepsilon_{ijs} \tag{13}$$

where $\phi < 0$, p_{js}, x_{js}, ξ_{js} are the price, observed and unobserved characteristics of the product j, common within a sector s. ε_{ijs} are idiosyncratic taste shocks distributed as the Type I Extreme Value distribution.

4.1.3 Nested Logit

The third framework extends the second by allowing for correlation in consumer unobserved utility across products within a *nest*, reflecting more realistic substitution patterns. Products are partitioned into G mutually exclusive nests. Consumers first choose a nest g, then a product $j \in g$. Utility is

$$u_{ijgs} = \delta_{js} + \sigma\psi_{gs} + (1-\sigma)\varepsilon_{ijgs}, \tag{14}$$

where $\delta_{js} = \phi p_{js} + x_{js}\gamma + \xi_{js}$, ψ_{gs} is the unobserved utility component common to all products in nest g and $\sigma \in [0, 1)$ captures within-nest correlation. When $\sigma = 0$, the model reduces to the standard logit model.

4.1.4 Random Coefficients

The random coefficients model extends the Logit specification by allowing for consumer heterogeneity in preferences over observable characteristics. Preferences are type-specific and drawn from a two-point distribution around a common mean, generating two consumer types with different sensitivities to product characteristics.

4.2 Competing Methods

We apply four estimators to each simulated dataset. All estimators use aggregate deflators to construct real revenues, reflecting standard practice when firm-level price indices are unavailable.

Measuring markups on revenue data (KMT) The KMT estimator adjusts for the use of revenue data by controlling for market shares and fixed effects in the estimation of the flexible input share.[12] It then recovers output elasticities using the stochastic evolution of revenue productivity, as in typical applications where only revenues and input costs are observed.

[12] In this exercise, we use firm and sector times year fixed effects. While we restrict attention to these minimal controls for comparability across models, the KMT framework allows researchers to incorporate additional demand and market structure variables into the D_{it} vector as discussed in Section 3.

Linear regression model (OLS) A simple log–log regression of deflated sales on inputs, which implicitly assumes constant markups and interprets revenues as quantities.

Standard production approach (ACF) The production approach outlined in De Loecker and Warzynski (2012) to measure markup recovering the output elasticities following Ackerberg et al. (2015). Specifically, treating deflated revenues as quantities, we estimate a first stage with a control function for productivity to purge output from measurement error. Then, we use the stochastic process of physical productivity to recover the variable input elasticity.

Dynamic panel method (BB) Based on Blundell and Bond (2000), this method estimates the production function in first differences using lagged deflated levels as instruments under an AR(1) assumption for productivity.

4.3 Estimation Results

Our baseline simulations deliberately adopt a conservative approach to demonstrate the robustness of the KMT estimator. We use only the most commonly available controls: revenue-based market shares and fixed effects (firm and sector-year), which represent the minimal information typically available to researchers working with financial data. This conservative specification establishes a lower bound on the method's performance. Researchers with access to additional information about demand conditions, product characteristics, or market structure can incorporate these variables into the D_{it} vector, potentially achieving even better results than those reported here.

Table 1 compares the estimated output elasticities of the flexible input across demand and competition environments. The true elasticity is $\alpha = 0.60$. In the Nested CES Cournot setting, KMT estimates 0.605 (bias $< 1\%$), while OLS, ACF, and BB overestimate by 23%, 33%, and 22% respectively. This bias arises because, in this highly structured demand system, a firm's high productivity is strongly correlated with its market share, which in turn is a key determinant of the market price index. Standard estimators, which are misspecified for this environment, misattribute this complex, equilibrium price-index effect to a higher-than-true output elasticity.

In the logit, nested logit, and random-coefficients logit settings, KMT produces estimates of 0.596, 0.577, and 0.619, with all biases under 4%. By contrast, OLS, ACF, and BB underestimate elasticities in these environments. The degree of bias varies with

Model	True	KMT	OLS	ACF	BB
Nested CES	0.600	0.605	0.738	0.797	0.730
Logit	0.600	0.596	0.512	0.554	0.589
Nested Logit	0.600	0.577	0.535	0.593	0.623
Random Coefficients Logit	0.600	0.619	0.512	0.562	0.579

Table 1: Output Elasticity Estimates Across Models

the demand system: ACF performs well in the nested logit (bias under 2%), while BB shows minimal bias in the simple logit. However, neither alternative method performs consistently across all environments, while KMT maintains accuracy throughout despite using only minimal controls.

Across all four environments, KMT maintains consistent accuracy despite differences in demand structure, while conventional estimators display systematic bias whose sign and magnitude depend on the demand system. In the structured Nested CES environment, traditional estimators overstate elasticities by 22-33%. In the Logit-family environments with more idiosyncratic demand shocks, they typically understate them, though the magnitude varies with the specific demand specification. These biases translate directly to the estimated markup distributions.

Table 2 reports the mean, median, and standard deviation of estimated markups, along with their bias relative to the true values, for each simulated environment.

In the Nested CES model with its structured demand system, KMT's mean estimate (1.266) deviates by less than 1% from the true value (1.252) with accurate dispersion. OLS, ACF, and BB all overstate average markups, with mean biases from 0.272 (BB) to 0.414 (ACF), reflecting their inability to account for the correlation between prices and inputs in this environment.

In the Logit model, KMT produces a mean bias of –0.007. OLS and ACF understate markups substantially, while BB performs better with a bias of –0.022. This pattern reflects how the Logit demand structure affects revenue elasticity estimation.

In the Nested Logit model, KMT's mean markup deviates by 0.048 from the true value. ACF produces the smallest bias (–0.015) in this setting, while BB overshoots by 0.048. BB's positive bias suggests its differencing strategy overcorrects when product characteristics are correlated within nests.

In the Random Coefficients Logit model, KMT's mean markup estimate is 0.039 above

Method	Mean	Median	SD	Mean Bias	Median Bias
Nested CES Model					
True	1.252	1.246	0.048	.	.
KMT	1.266	1.266	0.090	0.013	0.010
OLS	1.542	1.536	0.085	0.289	0.287
ACF	1.667	1.661	0.092	0.414	0.412
BB	1.525	1.519	0.084	0.272	0.270
Logit Model					
True	1.202	1.196	0.037	.	.
KMT	1.195	1.191	0.050	-0.007	-0.008
OLS	1.027	1.023	0.033	-0.176	-0.177
ACF	1.110	1.105	0.036	-0.093	-0.094
BB	1.180	1.176	0.038	-0.022	-0.023
Nested Logit Model					
True	1.272	1.262	0.058	.	.
KMT	1.225	1.216	0.074	-0.048	-0.049
OLS	1.135	1.127	0.053	-0.138	-0.139
ACF	1.258	1.249	0.059	-0.015	-0.015
BB	1.321	1.312	0.062	0.048	0.049
Random Coefficients Logit Model					
True	1.202	1.196	0.037	.	.
KMT	1.241	1.236	0.052	0.039	0.038
OLS	1.027	1.023	0.033	-0.175	-0.177
ACF	1.126	1.121	0.036	-0.077	-0.078
BB	1.160	1.155	0.037	-0.043	-0.043

Table 2: Markup Distribution and Bias Across Models

the true value, while OLS shows the largest underestimation (–0.175). KMT estimates higher markup dispersion than other methods, which may reflect markup heterogeneity across product–market pairs under preference heterogeneity.

These results demonstrate the robustness of the KMT estimator across diverse market environments. By directly addressing the use of revenue data and controlling for endogenous productivity and market shares, KMT consistently outperforms traditional approaches that misinterpret revenue as quantity and fail to correct for price endogeneity.

To illustrate the method's extendibility, we next examine performance when researchers have access to richer controls, following the discussion in Appendix A. While our baseline simulations use only revenue-based market shares and fixed effects—the minimal controls typically available—many empirical settings offer additional information. When

Model	True	KMT	Absolute Bias	Relative Bias
Nested CES	0.600	0.605	0.005	0.008
Logit	0.600	0.601	0.001	0.002
Nested Logit	0.600	0.600	0.000	0.001
Random Coefficients Logit	0.600	0.597	-0.003	-0.005

Table 3: Output Elasticity Estimates Across Models With Perfect Controls

we augment the control set by including the true price directly (representing an ideal scenario where researchers have detailed price data), as shown in Table 3, the estimator recovers the true elasticities with near-zero bias in all four models. This demonstrates that our framework can fully exploit additional control variables when available. Researchers with access to quantity-based market shares, input prices, demand shifters, or other market-specific variables can incorporate these into the D_{it} vector to further improve estimation.

The flexibility of our approach represents a key advantage: with only minimal controls (fixed effects and revenue-based market shares), it already substantially reduces the bias that plagues existing methods; with comprehensive controls tailored to the specific empirical setting, it can achieve unbiased estimates. This extendibility allows researchers to leverage their institutional knowledge and data availability to improve estimation beyond the conservative baseline demonstrated here.

5 Conclusion

This paper addresses the problem of estimating markups when only revenue data are available by proposing a method that works without information on prices. It recovers unbiased and consistent markup estimates using only common regression techniques and information available in most datasets. The method is based on a production function estimator that flexibly models markups as a function of observables and fixed effects and treats revenue productivity rather than physical productivity as the state variable. This approach solves the omitted price bias without imposing additional assumptions on demand or competition structure.

Modeling markups as a function of observable firm characteristics and fixed effects captures insights from recent macroeconomic and trade models featuring variable markups. These controls capture factors determining variable demand elasticities such as market complementarities or industry specific characteristics. In addition, the stochastic process for revenue productivity makes explicit assumptions underlying most recent work investigating markups and is consistent with evidence on the dynamics of revenue productivity.

Our method provides a simple and effective way to estimate markups using only revenue data. It has important implications for researchers and policymakers interested in understanding the dynamics of product markets and the impact of market power on economic outcomes. The approach can be extended to various empirical settings, making it a valuable tool for future research.

References

Ackerberg, D., L. Benkard, S. Berry, and A. Pakes (2007). Econometric tools for analyzing market outcomes. *Handbook of Econometrics 6*, 4171–4276.

Ackerberg, D., K. Caves, and G. Frazer (2015). Identification properties of recent production function estimators. *Econometrica 83*(6), 2411–2451.

Atkeson, A. and A. Burstein (2008). Pricing-to-market, trade costs, and international relative prices. *American Economic Review 98*(5), 1998–2031.

Autor, D., D. Dorn, L. Katz, C. Patterson, and J. Van Reenen (2020). The fall of the labor share and the rise of superstar firms. *Quarterly Journal of Economics 135*(2), 645–709.

Baker, J. and T. Bresnahan (1988). Estimating the residual demand curve facing a single firm. *International Journal of Industrial Organization 6*(3), 283–300.

Basu, S. and J. Fernald (1997). Returns to scale in us production: Estimates and implications. *Journal of Political Economy 105*(2), 249–283.

Basu, S. and J. Fernald (2002). Aggregate productivity and aggregate technology. *European Economic Review 46*(6), 963–991.

Bessen, J. (2020). Industry concentration and information technology. *Journal of Law and Economics 63*(3), 531–555.

Blundell, R. and S. Bond (2000). GMM estimation with persistent panel data: An application to production functions. *Econometric Reviews 19*(3), 321–340.

Bond, S., A. Hashemi, G. Kaplan, and P. Zoch (2021). Some unpleasant markup arithmetic: Production function elasticities and their estimation from production data. *Journal of Monetary Economics 121*.

Calligaris, S., C. Criscuolo, and L. Marcolin (2018). Mark-ups in the digital era. Working paper.

De Loecker, J. (2011). Product differentiation, multiproduct firms, and estimating the impact of trade liberalization on productivity. *Econometrica 79*(5), 1407–1451.

De Loecker, J., J. Eeckhout, and G. Unger (2020). The rise of market power and the macroeconomic implications. *Quarterly Journal of Economics 135*(2), 561–644.

De Loecker, J., P. Goldberg, A. Khandelwal, and N. Pavcnik (2016). Prices, markups, and trade reform. *Econometrica 84*(2), 445–510.

De Loecker, J. and F. Warzynski (2012). Markups and firm-level export status. *American Economic Review 102*(6), 2437–71.

De Ridder, M., B. Grassi, and G. Morzenti (20225). The hitchhiker's guide to markup estimation. *Econometrica*.

Decker, R., J. Haltiwanger, R. Jarmin, and J. Miranda (2016). Declining business dynamism: What we know and the way forward. *American Economic Review 106*(5), 203–07.

Diez, F., D. Leigh, and S. Tambunlertchai (2018). Global market power and its macroeconomic implications. Working paper.

Doraszelski, U. and J. Jaumandreu (2013). R&D and productivity: Estimating endogenous productivity. *Review of Economic Studies 80*(4), 1338–1383.

Doraszelski, U. and J. Jaumandreu (2018). Measuring the bias of technological change. *Journal of Political Economy 126*(3), 1027–1084.

Doraszelski, U. and J. Jaumandreu (2019). Using cost minimization to estimate markups. Working paper, Social Science Research Network.

Ericson, R. and A. Pakes (1995). Markov-perfect industry dynamics: A framework for empirical work. *Review of Economic Studies 62*(1), 53–82.

Flynn, Z., J. Traina, and A. Gandhi (2019). Measuring markups with production data. Working paper, Social Science Research Network.

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

Gandhi, A., S. Navarro, and D. Rivers (2020). On the identification of gross output production functions. *Journal of Political Economy 128*(8), 2973–3016.

Griliches, Z. and J. Mairesse (1995). Production functions: The search for identification. Working paper, National Bureau of Economic Research.

Grullon, G., Y. Larkin, and R. Michaely (2019). Are US industries becoming more concentrated? *Review of Finance 23*(4), 697–743.

Gutiérrez, G. and T. Philippon (2019). The failure of free entry. Working paper, National Bureau of Economic Research.

Hall, R. (1988). The relation between price and marginal cost in US industry. *Journal of Political Economy 96*(5), 921–947.

Hashemi, A., I. Kirov, and J. Traina (2022). The production approach to markup estimation often measures input distortions. *Economics Letters 217*.

Klette, T. J. and Z. Griliches (1996). The inconsistency of common scale estimators when output prices are unobserved and endogenous. *Journal of Applied Econometrics 11*(4), 343–361.

Levinsohn, J. and A. Petrin (2003). Estimating production functions using inputs to control for unobservables. *Review of Economic Studies 70*(2), 317–341.

Marschak, J. and W. Andrews (1944). Random simultaneous equations and the theory of production. *Econometrica*, 143–205.

Olley, S. and A. Pakes (1996). The dynamics of productivity in the telecommunications equipment industry. *Econometrica 64*(6), 1263–1297.

Petrin, A. and J. Sivadasan (2013). Estimating lost output from allocative inefficiency, with an application to chile and firing costs. *Review of Economics and Statistics 95*(1), 286–301.

Syverson, C. (2004). Market structure and productivity: A concrete example. *Journal of Political Economy 112*(6), 1181–1222.

Traina, J. (2018). Is aggregate market power increasing? Production trends using financial statements. Working paper, Stigler Center for the Study of the Economy and the State.

Wooldridge, J. (2009). On estimating firm-level production functions using proxy variables to control for unobservables. *Economics Letters 104*(3), 112–114.

Appendix

A Economic Content of the Revenue Productivity Process

This appendix provides microfoundations for Assumption 1 in Section 2, which states that revenue productivity $\nu_{it} = p_{it} + \omega_{it}$ follows a Markov process. A natural concern is why the sum of an endogenous price and physical productivity should follow such a process when prices are choice variables that respond to market conditions. The following analysis shows this Markovian structure can plausibly emerge from equilibrium behavior rather than being an arbitrary assumption.

As in the main text, let uppercase denote levels and lowercase logs. Revenue productivity is

$$\nu_{it} \equiv p_{it} + \omega_{it},$$

log price plus log physical productivity.

Consider the following economic environment consistent with a large class of dynamic oligopoly models with Markov-perfect equilibria in the spirit of Ericson and Pakes (1995). At each point in time, an industry state S_t captures all payoff-relevant information: firms' productivity levels, demand shifters, and other economic conditions. This state evolves according to a first-order Markov process:

$$S_t = F(S_{t-1}, \varepsilon_t),$$

where ε_t represents the vector of shocks. We assume competition within each period is static, so prices and quantities are fully determined by the current state without dynamic strategic effects across periods. Given state S_t, each firm chooses its price optimally according to the relevant equilibrium concept (monopolistic competition, oligopoly, etc.). Together with the firm's physical productivity ω_{it}, which is itself part of S_t, this determines revenue productivity through the equilibrium mapping:

$$\nu_{it} = V_i(S_t).$$

The stochastic process for ν_{it} follows directly from this equilibrium structure. Define the conditional expectation function $g(s) \equiv \mathbb{E}[V_i(S_t) \mid S_{t-1} = s]$, which represents the expected revenue productivity given yesterday's state. The innovation or surprise component is then $\eta_{it} \equiv \nu_{it} - g(S_{t-1})$. By construction, η_{it} has mean zero conditional on S_{t-1}. To connect this to the firm's decision problem, we assume that the firm's information set when making input choices at $t-1$ is no richer than the industry state:

$\sigma(\mathcal{I}_{it-1}) \subseteq \sigma(S_{t-1})$. This holds when firms observe the current industry state but not future shocks, or when any private information is already incorporated into S_{t-1}. Under this assumption, the law of iterated expectations gives us:

$$\mathbb{E}[\eta_{it} \mid \mathcal{I}_{it-1}] = \mathbb{E}[\mathbb{E}[\eta_{it} \mid S_{t-1}] \mid \mathcal{I}_{it-1}] = \mathbb{E}[0 \mid \mathcal{I}_{it-1}] = 0.$$

Therefore:

$$\nu_{it} = g(S_{t-1}) + \eta_{it}, \qquad \mathbb{E}[\eta_{it} \mid \mathcal{I}_{it-1}] = 0. \tag{A.1}$$

This equation establishes that revenue productivity follows a predictable process as a natural consequence of equilibrium behavior. To see why endogenous pricing doesn't destroy this structure, note that the equilibrium mapping $V_i(\cdot)$ incorporates both components: the firm's physical productivity (which is part of S_t) and its optimal pricing response to market conditions (also determined by S_t). Rather than treating prices and productivity as separate random processes that we hope combine nicely, the equilibrium approach shows they jointly evolve according to the industry state dynamics.

Implementation requires addressing the fact that the full state S_{t-1} is high-dimensional and unobservable. Assumption 1 in the main text imposes that lagged revenue productivity provides a sufficient statistic for forecasting—what we term *scalar sufficiency*:

$$\mathcal{P}(\nu_{it} \mid S_{t-1}) = \mathcal{P}(\nu_{it} \mid \nu_{i,t-1}). \tag{A.2}$$

Under this condition, the process simplifies to $\nu_{it} = g(\nu_{i,t-1}) + \eta_{it}$ where $g(\nu_{i,t-1}) = \mathbb{E}[\nu_{it}|\nu_{i,t-1}]$, yielding equation (4) of the main text. While convenient, scalar sufficiency can fail if other elements of S_{t-1} have independent persistence and predictive power for ν_{it} beyond what is captured by $\nu_{i,t-1}$.

Before examining when scalar sufficiency holds, we address a basic time-series concern. Even if prices and productivity each follow AR(1) processes, their sum is generally not AR(1). For AR(1) processes with parameters ϕ_p and ϕ_ω and serially uncorrelated innovations (allowing contemporaneous correlation):

$$(1 - \phi_p L)(1 - \phi_\omega L)\, \nu_{it} = (1 - \phi_\omega L) u_{it} + (1 - \phi_p L) w_{it},$$

where u_{it} and w_{it} are the respective innovations. The right side forms a bivariate MA(1) process, making ν_{it} generically ARMA(2,2). Special alignments (equal AR roots or proportional innovations) can yield ARMA(2,1), but AR(1) only emerges in knife-edge cases. This matters because an ARMA(2,2) process requires two lags of revenue productivity plus moving average terms for exact representation—complexity that a simple AR(1) specification would miss. However, when one persistence parameter dominates—particularly when $\phi_\omega \gg \phi_p$ because productivity shocks are more persistent than transitory price movements—an AR(1) approximation may capture most of the dynamics. When this approximation is inadequate, researchers must include additional lags and state variables: $g(\nu_{i,t-1}, Z_{i,t-1}, \nu_{i,t-2}, \ldots)$.

Three economically relevant cases yield exact scalar sufficiency. First, when a single persistent index drives all relevant variation and revenue productivity is a one-to-one function of this index. Specifically, if $U_t = \rho U_{t-1} + u_t$ and $\nu_{it} = \tilde{V}_i(U_t)$ where $\tilde{V}_i$ is strictly monotone, then knowing $\nu_{i,t-1}$ uniquely determines U_{t-1}, which suffices for prediction. This case requires both that all variation loads on a single factor and that the mapping preserves information about that factor.

Second, in linear–Gaussian environments where $S_t = AS_{t-1} + \varepsilon_t$ and $V_i(S_t) = c_i' S_t$, if c_i is a left eigenvector of A satisfying $c_i' A = \lambda c_i'$ for eigenvalue λ, then $\mathbb{E}[\nu_{it} \mid S_{t-1}] = c_i' A S_{t-1} = \lambda c_i' S_{t-1} = \lambda \nu_{i,t-1}$, yielding an exact AR(1) process. While this knife-edge condition rarely holds precisely, it illustrates how linear state dynamics can generate simple revenue productivity dynamics.

Third, when both markups and productivity respond to the same scalar persistent shifter, scalar sufficiency can hold. For instance, if a common demand or cost factor drives both market power and productivity decisions, and this factor follows an AR(1) process, then revenue productivity inherits this scalar structure. However, if markups respond to one persistent factor while productivity responds to a different one, scalar sufficiency fails and augmented controls become necessary.

More generally, when scalar sufficiency fails, researchers need multiple proxy variables to span the multidimensional state space. As emphasized in Ackerberg et al. (2007), when the true state is multidimensional, a single proxy cannot capture all relevant variation. The augmented specification becomes:

$$\nu_{it} = g(\nu_{i,t-1}, Z_{i,t-1}) + \eta_{it}, \tag{A.3}$$

where $Z_{i,t-1}$ includes additional state variables such as lagged capital, input prices, or market structure variables analogous to the controls $\mathbf{D}_{it}$ in the markup control function. When these variables together span the relevant predictive information from S_{t-1}, this specification recovers the true conditional expectation.

Researchers can test for failures of scalar sufficiency by examining whether additional lags or controls have incremental predictive power and verifying that estimated innovations $\hat{\eta}_{it}$ are orthogonal to observed predetermined variables in the researcher's instrument set. While we cannot test orthogonality to the unobserved information set $\mathcal{I}_{it-1}$ directly, we can test against observables that should be measurable with respect to it. Capital and input prices with independent dynamics, strategic interactions not captured by own past performance, or state-dependent measurement error can all violate scalar sufficiency.

For implementation, researchers construct the revenue productivity proxy using the procedure detailed in Section 2. After obtaining $\hat{s}_{it} = \log \widehat{f_{it}^X - \mu_{it}}$ and $\hat{\varepsilon}_{it}$ from the first-stage share regression:

$$\hat{\nu}_{it} = c_{it} x_{it} - f(x_{it}, k_{it}) - \hat{s}_{it} - \hat{b}_{it}.$$

This formalization extends the control-function tradition in production estimation (Olley and Pakes, 1996; Levinsohn and Petrin, 2003; Ackerberg et al., 2015) by making explicit why revenue productivity can serve as a state variable despite endogenous prices. It provides theoretical support for treating revenue productivity as Markovian—the "standard maintained assumption" referenced in the main text and applied throughout the literature. Although not the primary focus of our paper, by deriving conditions under which this structure emerges from equilibrium primitives, the analysis shows that endogenous pricing need not undermine the tractability required for estimation—providing a formal justification for Assumption 1 when these conditions are approximately satisfied.

B Simulation Details

This appendix provides additional information on the four simulated data-generating processes used in Section 4, including parameter values and distributional assumptions for productivity, input prices, and demand characteristics. Each model is solved for 50 periods. The Nested CES model features 1440 firms across 180 sectors (8 firms per sector), while the Logit, Nested Logit, and Random Coefficients models each contain 800 firms across 80 sectors (10 firms per sector).

B.1 Supply-Side

All models feature firms that produce output according to a Cobb-Douglas production function:

$$y_{it} = \omega_{it} x_{it}^{\alpha} k_{it}^{1-\alpha},$$

where x_{it} is a variable (flexible) input, k_{it} is a quasi-fixed input, and ω_{it} is firm-level productivity. The parameter $\alpha = 0.6$ represents the output elasticity of the flexible input.

Productivity follows an AR(1) process:

$$\log \omega_{it} = \rho_{\omega} \log \omega_{i,t-1} + \nu_{it}, \quad \nu_{it} \sim \mathcal{N}(0, \sigma_{\omega}^2).$$

Capital evolves as:

$$\log k_{it} = \rho_k \log k_{i,t-1} + \zeta_{it}, \quad \zeta_{it} \sim \mathcal{N}(0, \sigma_k^2).$$

Flexible input prices are idiosyncratic and follow:

$$\log Z_{it} = \rho_Z \log Z_{i,t-1} + \xi_{it}, \quad \xi_{it} \sim \mathcal{N}(0, \sigma_Z^2).$$

All models assume the same production parameters: $\alpha = 0.6$, $\rho_{\omega} = 0.7$, $\rho_k = 0.66$, $\rho_Z = 0.9$, and standard deviations $\sigma_{\omega} = 0.10$, $\sigma_k = 0.30$, and $\sigma_Z = 0.20$.[13]

[13]While this calibration is not linked to an empirical exercise, we take most of the parametrization

B.2 Nested CES

The environment features a competitive final-good sector, 180 industries indexed by s, and 8 heterogeneous firms $i = 1, \ldots, 8$ within each industry.[14] Firms produce differentiated intermediates using the Cobb–Douglas production function and engage in Cournot competition.

Let p_{is} denote the price set by firm i in sector s, and let p_s and p be the corresponding CES price indices at the sector and final-good levels. Demand for sector s is given by $y_s = (p/p_s)^\theta y$, while demand for firm i's product is $y_{is} = (p_s/p_{is})^\eta y_s$. The final-good price index satisfies $p = \left[\int_0^1 p_s^{1-\theta} ds\right]^{1/(1-\theta)}$, and the sector index is $p_s = \left[\sum_{i=1}^8 p_{is}^{1-\eta}\right]^{1/(1-\eta)}$.

Substituting across levels yields firm-level demand:

$$y_{is} = D p_{is}^{-\eta} p_s^{\eta-\theta}, \quad \text{where } D = p^\theta y.$$

Each firm's revenue-based market share is $s_{is} = (p_{is}/p_s)^{1-\eta}$, and the perceived elasticity of demand is:

$$\epsilon_{is} = \left[\frac{1 - s_{is}}{\eta} + \frac{s_{is}}{\theta}\right]^{-1}.$$

This elasticity averages between the within- and across-sector elasticities, and varies endogenously with market share.

Firms maximize static profits. Letting mc_{is} denote marginal cost, the optimal price satisfies:

$$p_{is} = \mu_{is} mc_{is}, \quad \text{where } \mu_{is} = \frac{\epsilon_{is}}{\epsilon_{is} - 1}.$$

Markups depend on perceived elasticity and hence vary with firm size.

Equilibrium. An equilibrium consists of firm-level variables $\{p_{is}, y_{is}, x_{is}, \mu_{is}, mc_{is}, s_{is}, p_s\}$ such that: firm output y_{is} is consistent with demand; the sector price index p_s aggregates firm prices; market shares s_{is} reflect firm prices relative to the sector index; perceived elasticities ϵ_{is} are implied by market shares; markups μ_{is} are chosen optimally; prices satisfy firms' first-order conditions; marginal costs mc_{is} are consistent with cost minimization given Cobb–Douglas technology; and input choices x_{is} are optimal given factor prices and firm productivities.

Parameter Values. We follow De Ridder et al. (0225) in setting the demand elasticities to $\eta = 10$ and $\theta = 1.1$, and in introducing time variation in demand through an AR(1) process for $\log D_t$.

from De Ridder et al. (0225).

[14] This framework is based on Atkeson and Burstein (2008) and implemented following De Ridder et al. (0225).

B.3 Logit

There is a continuum of consumers choosing among differentiated products offered by 10 single-product firms in each of 80 sectors. Firms compete in prices (Bertrand competition), recognizing the impact of their prices on market shares. The supply side features the same technology and input structure as the Nested CES model.

Each consumer j in sector s receives utility from product i according to:[15]

$$u_{jis} = \phi p_{is} + o_{is}\gamma + \xi_{is} + \varepsilon_{jis},$$

where p_{is} is the price of product i, o_{is} its observed characteristic, and ξ_{is} an unobserved demand shifter. The idiosyncratic taste shocks ε_{ijs} follow a Type I Extreme Value distribution.

Aggregating over consumers yields the standard Logit demand system. Denoting mean utility by $\delta_{is} = \phi p_{is} + o_{is}\gamma + \xi_{is}$, product i's market share in sector s is:

$$s_{is} = \frac{e^{\delta_{is}}}{1 + \sum_{k \in I} e^{\delta_{ks}}}.$$

The elasticity of demand is:

$$\epsilon_{is} = \frac{\partial s_{is}}{\partial p_{is}} \frac{p_{is}}{s_{is}} = \phi p_{is}(1 - s_{is}).$$

Using the Logit elasticity, the Bertrand pricing rule becomes:

$$p_{is} = \mu_{is} c_{is}, \quad \text{where} \quad \mu_{is} = \frac{1}{1 + \frac{1}{\epsilon_{is}}}.$$

Markups are increasing in demand elasticity and depend on both price sensitivity and market share.

Equilibrium. Given productivities ω_{is}, flexible input prices Z_s, and fixed inputs k_{is}, an equilibrium is a set of firm-level variables $\{p_{is}, y_{is}, x_{is}, \mu_{is}, c_{is}, s_{is}\}$ such that: prices p_{is} satisfy the firm's first-order condition; market shares s_{is} are consistent with the Logit demand system; demand elasticities ϵ_{is} reflect price and share; markups μ_{is} follow from the Bertrand rule; marginal costs c_{is} are implied by cost minimization; and input choices x_{is} are optimal given factor prices and productivity.

Parameter Values. In the Logit simulation, we set the price coefficient to $\phi = -1.6$ and the coefficient on observed characteristics to $\gamma = 1$, in line with values used in

[15] We restrict firms to be single-product, so each product corresponds to a firm.

structural demand estimation. Each sector has a fixed market size $M_s = 100$. The observed product characteristic o_{is} evolves over time according to an AR(1) process with persistence $\rho_o = 0.90$ and standard deviation $\sigma_o = 0.20$. The unobserved characteristic ξ_{is} is set to a constant value of 3 across all products and periods, which ensures a realistic share for the outside option.

B.4 Nested Logit

The nested Logit model generalizes the standard Logit framework by relaxing the Independence of Irrelevant Alternatives (IIA) assumption. It allows for correlation in unobserved utility across products within a *nest*, introducing more realistic substitution patterns. In our setting, each sector s contains 10 single-product firms, which we partition into $G = 2$ mutually exclusive nests (5 products per nest).

Consumers make choices in two stages: they first select a nest $g \in \{1, 2\}$ and then choose a product i within that nest. Consumer utility is given by:

$$u_{ijgs} = \delta_{is} + \sigma\psi_{gs} + (1-\sigma)\varepsilon_{ijgs},$$

where $\delta_{is} = \phi p_{is} + o_{is}\gamma + \xi_{is}$ is the mean utility, ψ_{gs} is a common shock shared across all products in nest g, and ε_{ijgs} follows a Type I Extreme Value distribution. The parameter $\sigma \in [0, 1)$ controls the correlation in unobserved utility within nests.

Product-level market shares decompose into two components. The *within-nest share* $s_{i|g,s}$ captures the probability of choosing product i conditional on choosing nest g, while the *nest share* $s_{g,s}$ reflects the probability of choosing nest g among all available nests (and the outside option). These are given by:

$$\begin{aligned}
s_{i|g,s} &= \frac{e^{\delta_{is}/(1-\sigma)}}{\sum_{k\in g} e^{\delta_{ks}/(1-\sigma)}}, \\
s_{g,s} &= \frac{\left(\sum_{k\in g} e^{\delta_{ks}/(1-\sigma)}\right)^{1-\sigma}}{1 + \sum_{h=1}^{G}\left(\sum_{k\in h} e^{\delta_{ks}/(1-\sigma)}\right)^{1-\sigma}}, \\
s_{isg} &= s_{g,s} \cdot s_{i|g,s}.
\end{aligned}$$

The own-price elasticity for product i in group g is:

$$\epsilon_{isg} = \frac{\phi p_{is}}{1-\sigma}\left(1 - \sigma s_{i|g,s} - (1-\sigma)s_{isg}\right).$$

As in the Logit case, profit-maximizing firms set prices according to the markup rule:

$$p_{is} = \mu_{is} c_{is}, \quad \text{where} \quad \mu_{is} = \frac{1}{1 + \dfrac{1}{\epsilon_{isg}}}.$$

Equilibrium. The definition of equilibrium is identical to the Logit case described. The only difference is the computation of demand elasticities, which now depend on both within- and across-nest substitution patterns.

Parameter Values. We set the number of nests to $G = 2$ and the nesting parameter to $\sigma = 0.5$, capturing moderate within-nest correlation in consumer preferences. Products are assigned evenly to nests, so each nest contains five products within a sector. All other parameters—including production and demand coefficients—are identical to those in the Logit specification.

B.5 Random Coefficients Logit

The Random Coefficients Logit model generalizes the standard Logit specification by allowing key demand parameters to vary across consumers. This introduces richer substitution patterns, further relaxing the IIA restriction. We model heterogeneity in preferences over observable product characteristics, splitting the population into $C = 2$ distinct consumer types.

Each consumer type $t = 1, 2$ is characterized by a taste coefficient γ_t that enters utility as follows:

$$u_{jis}^{(t)} = \phi p_{is} + o_{is}\gamma_t + \xi_{is} + \varepsilon_{jis},$$

where p_{is} is the price of product i in sector s, o_{is} is an observable product characteristic, ξ_{is} is an unobserved term common across consumers, and ε_{jis} follows a Type I Extreme Value distribution. Consumers choose the product that yields the highest utility.

Each type faces standard Logit demand, resulting in type-specific market shares:

$$s_{is}^{(t)} = \frac{\exp(\phi p_{is} + o_{is}\gamma_t + \xi_{is})}{1 + \sum_{k \in I} \exp(\phi p_{ks} + o_{ks}\gamma_t + \xi_{ks})}.$$

The aggregate share for product i is the population-weighted average across types:

$$s_{is} = \sum_{z=1}^{C} \pi_z s_{is}^{(z)},$$

where π_z represents the population share of consumer type z.

Equilibrium. The definition of equilibrium is identical to the Logit case. The only difference is that demand elasticities and market shares now reflect consumer heterogeneity and are computed by aggregating over types.

Parameter Values. We set the type-specific coefficients to $\gamma_1 = 0.5$ and $\gamma_2 = 1.5$, with equal population weights $\pi_1 = \pi_2 = 0.5$. All other parameters—including production and demand coefficients—are identical to those in the Logit specification.

Because the presence of consumer heterogeneity prevents a closed-form expression for own-price elasticities, we compute them numerically using finite differences.