Mechanically Determined Markups: A New Critique of the Production-Based Approach to Markup Estimation

Finley Lawlor-Mendez

A thesis submitted to

Auckland University of Technology

in partial fulfillment of the requirements for the degree of

Master of Business (MBus)

2026

Department of Economics and Finance

Table of Contents

List of Figures

List of Tables

Attestation of Authorship

I hereby declare that this submission is my own work and that, to the best of my knowledge and belief, it contains no material previously published or written by another person (except where explicitly defined in the acknowledgements), nor used artificial intelligence tools or generative artificial intelligence tools (unless it is clearly stated, and referenced, along with the purpose of use), nor material which to a substantial extent has been submitted for the award of any other degree or diploma of a university or other institution of higher learning.

[redacted]

Acknowledgements

I would like to express my sincere thanks to my supervisors, Matthew Ryan and Dr. Pik Yi Lydia Cheung, for their guidance and insight throughout my undertaking of this research. Their generous feedback and discussions were invaluable in shaping and refining this work.

Disclaimer

These results are not official statistics. They have been created for research purposes from the Integrated Data Infrastructure (IDI) and Longitudinal Business Database (LBD) which are carefully managed by Stats NZ. For more information about the IDI and/or LBD please visit https://www.stats.govt.nz/integrated-data/. The results are based in part on tax data supplied by Inland Revenue to Stats NZ under the Tax Administration Act 1994 for statistical purposes. Any discussion of data limitations or weaknesses is in the context of using the IDI for statistical purposes, and is not related to the data's ability to support Inland Revenue's core operational requirements.

Abstract

This paper identifies two critical flaws in the widely used production approach to markup estimation, illustrating them empirically with a new markup series constructed from New Zealand data. The first is the bias that occurs when markups are estimated from revenue (rather than output) data. While well-established in theory, the form and empirical implication of this bias are unclear. We derive an explicit expression for this empirical bias, showing that it can weaken or even invert true markup trends. We then uncover a second flaw: the level of markup estimates is determined mechanically by the researcher's definition of variable input. We establish this in theory before showing empirically that broad definitions, i.e., aggregations of multiple inputs, depress markup estimates, whereas narrow definitions inflate them. This reduces the estimated markup to an arbitrary value determined by data constraints and researcher choice.

1 | Introduction

Competition between firms is a vital element of an economy's health. In its absence, firms are incentivized to produce less, and charge more than they would in a competitive market. While this is to the great benefit of the firm, it comes at the expense of consumers, with the net effect being a loss of total societal welfare. It is therefore in the interest of society to regulate or intervene in markets to promote competition. To this end, the quantification of competition into a measurable, trackable figure is highly desirable, and this interest has led to an entire literature focused on developing, testing, and exploring the measurement of competition.

One of the most prominent competition measures is the markup, defined as a firm's ratio of price to marginal cost. Marginal cost is effectively unobservable and cannot be found in typical datasets, so researchers rely on an empirical estimation strategy established by Hall (1988), known as the "production approach". Under the assumption of cost-minimization, Hall shows that the markup (μ_{it}) can be expressed in a form derivable from widely available data:

$$\mu_{it} = \theta_{it}^{v} \frac{P_{it} Q_{it}}{P_{it}^{V} V_{it}}$$

where P_{it} and Q_{it} denote a firm's output price and quantity, and P_{it}^{V} and V_{it} denote the price and used quantity of a variable input (V). The specific variable input considered is a methodological choice. In theory, any flexible input or bundle — i.e., sum — of inputs will be equally valid. The term θ_{it}^{v} is the elasticity of output with respect to this chosen input. This is not directly observable and must be estimated econometrically — the primary empirical challenge of this approach.

This measure has recently garnered significant attention due to its central role in De Loecker et al. (2020), an influential paper that draws bold conclusions. Using the Hall method, De Loecker et al. find that average markups in the United States have grown steadily from 1.2 in 1980 to 1.6 in 2016, suggesting a pervasive and uninterrupted trend of growing market power and diminishing competition.

This is a provocative result with strong implications for the evolution and state of the economy. Subsequent papers in the literature have thus placed intense scrutiny on the methodology used to derive this trend, resulting in a re-evaluation of the feasibility of markup estimation and the veracity of prior papers relying on this method. This thesis contributes to this literature by arguing that empirical implementations of the production approach fail to recover true markups for two conceptually distinct reasons: trend distortions caused by the use of revenue data in estimation, and a mechanical relationship between defined inputs and the estimated markup's level.

The former distortion is the primary focus of the critical literature and regards the impossibility of recovering θ_{it}^{v} from firm revenue data. Proper estimation of output elasticities entails the estimation of a physical production function, i.e.:

$$Q_{it} = A_{it} V_{it}^{\theta^v} K_{it}^{\theta^k}$$

where K_{it} denotes capital, A_{it} denotes unobserved productivity, and Q_{it} denotes physical units of firm output. In practice, however, Q_{it} is rarely observed, as typical sources of firm data — financial statements and tax filings — are reported in terms of revenues and expenditures, rather than physical quantities of inputs and outputs.

This forces us to estimate a 'revenue function' with revenue — the product of output price (P_{it}) and quantity (Q_{it}) — rather than output, as the dependent variable. In theory, the estimates recovered from this function are revenue elasticities, i.e., the change in *revenue* resulting from a change in input. As we will cover in more detail later, identification issues prevent this, and so the recovered estimate — henceforth denoted by $\theta^{v,r}$ — will be neither a revenue elasticity nor an output elasticity. While the literature does not derive the exact form and properties of $\theta^{v,r}$, it is clear that it is not a substitute for θ^{v}. De Loecker et al. (2020) and many prior papers, however, use them as such, casting serious doubt on the validity of their results.

This methodological debate is the focus of several papers following De Loecker et al. Van Dijcke (2023) shows that output elasticities cannot be recovered from the revenue function. Hashemi et al. (2022) suggest that when $\theta^{v,r}$ is substituted for θ^{v} in the markup calculation, the result measures input distortions rather than markups, reinterpreting previous papers from this perspective. Bond et al. (2021), the most influential of these papers, show that if $\theta^{v,r}$ is equivalent to the theoretical revenue elasticity, estimated markups will be equal to 1, and are thus entirely uninformative about true markups. While identification issues prevent this theoretical elasticity from being empirically observed — hence the previous empirical findings of markups other than 1 — this is still a troubling criticism, suggesting that the derivation of markups from revenue data is fundamentally misguided, and that a new approach is required.

This conclusion was challenged, and a more positive result provided, by De Ridder et al. (2024) who propose that, while the *level* of estimated markups will be biased by the use of revenue data, *trends* can still be recovered. This would allow us to judge the evolution of markups over time, salvaging the core finding of De Loecker et al. — that markups have risen — even if their levels are mismeasured.

This debate about the recoverability of trends, and the broader point regarding the utility of revenue-derived markup estimates, has become the central focus of the markup literature. However, even if revenue-derived markups were undistorted, there is a separate, less discussed but equally problematic issue which undermines

the current production approach to markup estimation: the sensitivity of markup estimates to the researcher's choice of variable input.

This issue was first raised by Traina (2018), who argued that the upward trend in markups found by De Loecker et al. can be attributed to their use of 'cost of goods sold' as the variable input in their analysis. Cost of goods sold captures the direct costs associated with production, but excludes expenses such as marketing, insurance, and executive salaries which are growing increasingly relevant over time. When Traina re-estimates De Loecker et al.'s markup series with these costs included, the upward trend is neutralized.

This finding establishes that the choice of variable input is a critical determinant of markup estimates. We build on this intuition by identifying an econometric channel which creates a mechanical relationship between estimated markup levels and the choice of variable input: misspecification bias. This is a novel finding and one of the core contributions of this thesis.

To estimate a production function — as is necessary for the recovery of θ^v — we must make assumptions about its form. For example, when empirically estimating the generic Cobb-Douglas production function (above), it is typically assumed that machinery, land, and vehicles enter the function together as capital (K), rather than as separate inputs. Similarly, we must assume whether inputs such as purchases, contractor fees, marketing, etc., should be bundled within our chosen variable input (V). If this assumed form diverges from the true form of V,[1] then our constructed bundle will be a noisy approximation of the true V, and our estimated elasticity (θ_{it}^v) will be forced downward by misspecification bias.

Under misspecification bias, we find that elasticity estimates increase diminishingly as the number of inputs within the chosen bundle grows. The measured cost of that input bundle ($P_{it}^V V_{it}$), however, increases linearly, and so the markup ($\mu_{it} = \theta_{it}^v \frac{P_{it}Q_{it}}{P_{it}^V V_{it}}$) will decrease with the size of the bundle. This means that the level of the estimated markup is determined entirely by the choice of variable input, independent of the true markup and regardless of whether we observe output quantities or revenues.

Misspecification bias and the distortions introduced by revenue data both undermine the production approach to markup estimation. This thesis therefore addresses the question: do empirical implementations of the production approach recover the true levels or trends of markups? To this end we make three contributions to the literature. Firstly, we derive a precise expression for the revenue-derived elasticity estimate which shows that it often moves inversely to the true output elasticity, clarifying that time-trends *cannot* be recovered from revenue data. Secondly, we identify and explore the properties of misspecification bias — novel to the literature — and

[1] This would arise if, for example, we excluded machinery and land from capital.

prove that markup levels are mechanically determined by the methodological choices of researchers; namely, the choice of variable input. Thirdly, we construct a markup series for New Zealand using data from Stats NZ's Integrated Data Infrastructure. We use this series to confirm our findings on misspecification and revenue bias, demonstrating and explaining their impact on empirical estimates. We conclude that the standard implementation of the production approach cannot recover either the level or trend of true markups.

The paper proceeds as follows. Section 2 first provides the necessary technical background on the production approach. Section 3 then delves into the first major critique, reviewing the distortions introduced by revenue data and the proposition that they do not prevent the recovery of trends. Section 4 broadens the scope to assess whether even correctly estimated markups are valid measures of competition and evaluates an alternative metric. Building on these critiques, Section 5 presents the first of our major contributions: a formal derivation of the revenue elasticity estimate, and an exploration of the properties of this estimate in simulation. Section 6 follows with our second contribution: the novel identification of misspecification bias and the mechanical channel that it creates. We then move to our empirical work, with Section 7 providing an overview of the data, and Section 8 discussing our empirical results, which reinforce our findings on revenue distortions and misspecification bias. Finally, Section 9 synthesizes these findings and offers a concluding assessment.

2 | Fundamentals of Markups

This section introduces the production approach to markup estimation. We begin in Section 2.1 by motivating the use of markups and deriving the estimable form of the markup from production theory. Section 2.2 then outlines the core empirical strategy, giving particular attention to the challenge of controlling for unobserved productivity.

2.1 | Markup Theory and Intuition

2.1.1 | Market Power

The interest in competition — and measuring it — stems from the societal benefits brought by competitive interaction between firms. When firms are pressured by rivals, their only recourse is to make themselves more attractive to consumers, incentivizing the reduction of prices, the enhancement of product quality, and innovation. Firms unable to produce these consumer gains will be forced out of the market, freeing up market share to be captured by more efficient firms. Over time, therefore, we expect the competitive process to optimize resource allocation and usage, producing more efficient firms which create better products at lower prices.

The competitive interactions and pressures which drive this process are hard to quantify and measure on a wide scale. Instead, 'competition measurement' uses the analysis of firm-level characteristics to infer the degree to

which the competitive process is present. The characteristic in question is market power, which can be broadly understood as the ability of a firm to set a price that is higher than its production cost. This is a strong negative indicator of competition, capturing one of the main channels through which competitive interaction is converted into societal benefit. If the average firm in an industry is charging significantly beyond its production cost, we can deduce that they are unable — due to the presence of a dominant firm — or unwilling to compete on the basis of price, signifying insufficient competitive pressure.

This is the predominant approach of the competition literature: measure market power, infer competition. The specific method of measurement generally falls into two categories: direct and indirect measures.

Indirect measures, rather than capturing market power directly, reflect easily observable industry characteristics such as market share concentration. A classic example is the Herfindahl–Hirschman Index (HHI): the squared sum of firm market shares in an industry. High values of HHI indicate the concentration of market share, which is interpreted as high market power and a lack of competition.

The validity of indirect measures rests on this interpretive leap; however, it is not entirely justified. Papers as early as Demsetz (1973) have cautioned against equating concentration with market power, arguing that the dominance of efficient and competitive firms is a feature of competition. This is, in fact, precisely the mechanism that drives allocative efficiency, and so cannot be used as evidence of absent competition.

Direct measures avoid this weakness by quantifying market power directly, sidestepping interpretive assumptions. They use firm level data to estimate the relationship between price and marginal cost, thereby linking directly to definition of market power. The most widely used examples include the previously defined markup — which will be the focus of this paper — and the price-cost margin, defined as:

$$\frac{Y_{it} - C_{it}}{Y_{it}}$$

where Y_{it} is revenue and C_{it} is variable cost. These are quite similar both in intention and execution, with both using revenue-based data to approximate the gap between prices and marginal costs. This similarity implies the measures will be highly related; however, it also means that they share the same fundamental limitations. We later confirm this relationship empirically and show that both measures are dominated by cyclical economic trends in the short term.

2.1.2 | The Estimable Form of the Markup

We have described market power as a firm's ability to set a price higher than its production cost. This is mirrored perfectly by the theoretical markup, defined as a firm's ratio of price to marginal cost:

$$\mu_{it} \equiv \frac{P_{it}}{MC_{it}}$$

As marginal cost is not observed in existing datasets, we rely on the Hall (1988) method of estimation. The foundation of this method is the observation that the gap between price and marginal cost can be approximated from data on the output and input use of firms. This requires re-expressing the markup in terms of these observable variables. A precise understanding of this re-expressed form is essential, being the basis for our estimation strategy and later critiques. We now review the derivation of this form, following its treatment in De Loecker et al. (2020).

Consider a market of firms indexed by i over periods t. Each firm has a production function $Q_{it}(.)$, which is based on its productivity A_{it}, capital K_{it}, and input V_{it}:

$$Q_{it} = Q(\mathrm{A}_{it}, K_{it}, V_{it})$$

While input can be adjusted, productivity is intrinsic, and capital is assumed to be fixed for the relevant period. This means that the only output-shifting variable which can be changed by a firm during period t is input V_{it}. Firms will then naturally minimize their use of this input to achieve their desired level of output; this is the cost-minimization condition leveraged by Hall (1988) to recover markups. It should be reiterated that V_{it} can refer both to individual inputs (e.g., steel) or to a broad bundle of inputs such as 'intermediate inputs', or the more expansive 'cost of goods sold'. These are functionally equivalent in theory if bundled inputs are perfect substitutes. This is not the case empirically, but we defer the discussion of bundling and its implications to Section 6.

The cost minimization problem can be represented as a Lagrangian function:

$$\mathcal{L}(V_{it}, K_{it}, \lambda_{it}) = P_{it}^{V} V_{it} + r_{it} K_{it} + F_{it} - \lambda_{it}(Q(.) - Q_{it})$$

where P_{it}^{V} is the price of input V_{it}, r_{it} is the user cost of capital, Q_{it} is a targeted output level, F_{it} is fixed cost, and λ_{it} is the Lagrange multiplier. For this function, the first order condition with respect to input is:

$$\frac{\partial \mathcal{L}_{it}}{\partial V_{it}} = P_{it}^{V} - \lambda_{it} \frac{\partial Q(.)}{\partial V_{it}} = 0$$

From here, we can multiply all terms by $\frac{V_{it}}{Q_{it}}$ and rearrange to get:

$$\frac{\partial Q(.)}{\partial V_{it}} \frac{V_{it}}{Q_{it}} = \frac{1}{\lambda_{it}} \frac{P_{it}^{V} V_{it}}{Q_{it}}$$

The left-hand side of the equation is, by definition, the output elasticity of input V_{it} — denoted as θ_{it}^{v}. This is the percentage change in output that results from a percentage change in input. We can therefore rewrite this as:

$$\theta_{it}^{v} = \frac{1}{\lambda_{it}} \frac{P_{it}^{V} V_{it}}{\mathrm{Q_{it}}}$$

Rearranging and multiplying by P_{it} then gives us the ratio:

$$\frac{P_{it}}{\lambda_{it}} = \theta_{it}^{v} \frac{P_{it} Q_{it}}{P_{it}^{V} V_{it}}$$

Now, consider that, within our Lagrangian framework, the optimal amount of V_{it} is determined by the desired output.[2] We therefore have the cost function:

$$C(Q) = \mathcal{L}\big(V^{*}(Q),\, K^{*}(Q),\, \lambda^{*}(Q)\big)$$

By envelope theorem, the Lagrange multiplier is therefore equivalent to marginal cost:

$$C'(Q) = MC = \lambda$$

This means that the ratio $\frac{P_{it}}{\lambda_{it}}$ is identical to the markup, $\mu_{it} \equiv \frac{p_{it}}{mc_{it}}$, and so the final, estimable form of the markup is:

$$\mu_{it} = \theta_{it}^{v} \frac{P_{it} Q_{it}}{P_{it}^{V} V_{it}}$$

This expression tells us that two elements are required to find the markup; the inverted revenue share of the variable input, $\frac{P_{it} Q_{it}}{P_{it}^{V} V_{it}}$, and the output elasticity of the input, θ^{v} — note that from here onwards we will refer to the inverted revenue share of the variable input simply as the 'revenue share'. The only assumption we rely on to derive this expression is cost-minimization, i.e., that firms will produce their chosen quantity of output using the least amount of input possible (given their productivity and capital). The desire of firms to minimize costs can be assumed to apply regardless of the mode or level of competition, making it a suitable foundation for a widely applicable method of markup estimation.

2.1.3 | The Connection Between Markups and Market Power

We have now derived the estimable form of the markup, but the connection between this form and market power may not be immediately obvious. Before moving on to estimation we will therefore attempt to develop an intuitive understanding of how the estimable form reflects the power of firms.

[2] For any firm in the current period, all variables in (3) are given aside from $\mathrm{Q_{it}}$ and V_{it}, which depend on the decisions of the firm. With the other variables given, the level of input needed solely depends on the desired output — more output requires more input, less output requires less input.

Recall that the initial identity of the markup was given as: $\mu_{it} \equiv \frac{p_{it}}{mc_{it}}$. In a perfectly competitive environment, prices should be set equal to marginal cost, making the markup, as defined, equal to one. On the other hand, when a market is not perfectly competitive firms will be able to break this price setting condition, raising prices above marginal cost. In that case, the markup would rise above one. The ratio of price to marginal cost is thus a direct and intuitive measure of market power. When this ratio is high, it means that firms can charge well beyond their production cost without regard for competitive pressure, implying that their market power is also high.

The estimable form, though lacking the marginal cost term directly, captures this same dynamic. Consider that we can re-express it in the following manner:

$$\mu_{it} = \frac{\partial Q(.)}{\partial V_{it}} \frac{V_{it}}{Q_{it}} \frac{P_{it} Q_{it}}{P_{it}^V V_{it}} = \frac{\partial Q(.)}{\partial V_{it}} \frac{P_{it}}{P_{it}^V} = \frac{\partial Q(.)}{\partial V_{it}} \div \frac{P_{it}^V}{P_{it}}$$

where the first equality simply follows from the definition $\theta^v = \frac{\partial Q(.)}{\partial V_{it}} \frac{V_{it}}{Q_{it}}$. This balance of ratios, $\mu_{it} = \frac{\partial Q(.)}{\partial V_{it}} \div \frac{P_{it}^V}{P_{it}}$, reflects the same dynamic between production cost and price that is captured by the initial markup definition, just expressed in terms of input productivity and relative prices. For example, under perfect competition the markup must be one, and so it follows that $\frac{\partial Q(.)}{\partial V_{it}} = \frac{P_{it}^V}{P_{it}}$ or equivalently, that the ratio of output price to input price should be equal to the marginal productivity of that input. To remain perfectly competitive, if a firm's input becomes more productive, they must lower their output price. If the input becomes more expensive, the price will be raised, and vice versa. If this balance is broken — for example if inputs become more productive but output prices remain the same — then the firm will be gaining more revenue from additional output than it is spending on the additional input — in other words, charging a markup.

To take a simple numerical example, if a firm needs to use five additional units of input to produce one additional unit of output, then, under perfect competition, the price of the output must be no more or less than five times the price of the input. Under these conditions, $\frac{\partial Q(.)}{\partial V_{it}} = \frac{P_{it}^V}{P_{it}} = \frac{1}{5}$, the firm's revenue is equal to its input expenditure, and $\mu_{it} = 1$. If more input was needed to produce a unit, or the price of input was higher, then the firm would be taking a loss by producing the additional unit. Likewise, if less input was needed or the output price was higher, then the firm would be making a profit, and a more competitive firm would undercut them.

The estimable form of the markup is thus equivalent to the definitional markup. Both pick up imbalances between productivity and prices, which occur when firms have the power to disregard competitive pressures.

2.2 | Estimating Markups

This section details the empirical strategy for estimating markups. Our derived estimable markup consists of two components: the estimated output elasticity of the input, and the revenue share — i.e., the ratio of revenue to expenditure on the input. While the revenue share can be observed directly in data, recovery of the output elasticity entails estimating the production function of each industry — a non-trivial task and the primary empirical challenge of markup estimation. We therefore begin with a brief overview of production function and output elasticity estimation, before focusing on unobserved productivity, the primary obstacle to identification.

2.2.1 | Overview of Output Elasticity Estimation

The 'output elasticity' of an input is the percentage change in output that occurs in response to a percentage change in input, holding all other factors fixed. To recover this relationship, we must begin by specifying a form of the production function that can be empirically estimated.

The standard example is a Cobb-Douglas function with three productive factors, a productivity term, and an error term:[3]

$$Q_{it} = A_{it} V_{it}^{\theta^v} K_{it}^{\theta^k} L_{it}^{\theta^l} E_{it}$$

where A_{it} is a persistent, firm specific productivity shock known to the firm — this can be anything from geographic advantage to managerial expertise — V_{it} is the (variable) input, K_{it} is capital, L_{it} is labor, and E_{it} is an error term accounting for production deviations caused by transitory productivity shocks or measurement error. Estimating this multiplicative form is difficult, hence, we then apply the log-transformation:

$$q_{it} = \theta_t^v v_{it} + \theta_t^k k_{it} + \theta_t^l l_{it} + w_{it} + e_{it}$$

where the lowercase letters denote the logs of their uppercase counterparts, and w_{it} denotes the log of productivity. This additive form, suggested by Marschak & Andrews (1944), is much more tractable and suggests a straightforward estimation via ordinary least squares (OLS). This approach fails, however, because productivity is not a measurable, observable variable, but an unobserved abstraction which accounts for persistent productivity differences between firms that are not explained by heterogeneity in their productive factors (input, capital, labor, etc.). Since productive firms will likely use more input *and* produce more output, a simple regression of output on the observable factors that does not account for productivity will yield biased estimates. The following section details the nature and severity of this bias, which necessitates the more complex estimation strategy used in practice.

[3] Note that the firm subscript is here absent as we now refer to the industry-wide elasticity estimated in practice, rather than the theoretical firm/time specific output elasticity.

2.2.2 | Productivity and Simultaneity Bias

The core problem of production function estimation is that productivity — an unobservable output multiplier — simultaneously influences output and input decisions. This creates a simultaneity bias which is more severe than a typical omitted variable.

Firms with high productivity can produce more from a given set of inputs, incentivizing the expansion of output, the usage of more input, the hiring of additional laborers, and the purchase of capital. There are thus two separate channels relating input and output: the direct productive effect of the input which is constant across firms — i.e., the output elasticity of the input, which is our interest — and the tendency of high-productivity, high-output firms to use more input. If we do not distinguish between these channels by controlling for productivity, we will not be able to recover the output elasticity.

Consider an industry of ten farms. Each farm uses fertilizer to grow lettuce, applying it across their available land. Everything about these farms is identical, other than that one of them has a better crop growing technique, allowing it to produce more lettuce with a given amount of fertilizer than its peers. Because its yield per hectare and per unit of fertilizer is higher, this efficient firm will naturally be incentivized to expand their farm, purchasing more land and using more fertilizer. This increased use of land and fertilizer results directly from the farm's intrinsic productivity, but when we look at the data, we will only observe that the farm with more fertilizer produces more plants; when OLS 'looks' at the data, it will determine that fertilizer is more productive than it really is.

If productivity were not correlated with input — and so all farms used equal input — we would only see that one farm, despite using the same input as the others, produced more output. This would simply be treated as random variance.

On the other hand, when productivity is correlated with input usage, two biases are created: first, an omitted variable bias, because unobserved productivity is correlated with both input and output, and second, simultaneity bias, because input and output are simultaneously determined by productivity via the greater resource allocation given to firms with high output. The estimated elasticities of these resources — the factors of production — will then not only reflect their own contribution to output, and the elevated level of output brought by productivity, but also the strategic relationship between productivity and factor levels. This is simultaneity.

Controlling for productivity is therefore imperative, and to do so, we rely on the well-known Olley & Pakes (OP) method, described in Olley & Pakes (1996). In essence, this method approximates the inherent productivity of firms by capturing their resource allocation decisions through an investment variable. Firms receiving heavy

investment are likely more productive, and unlike productivity, investment can easily be quantified, making it an ideal control.

2.2.3 | Accounting for Productivity with the Olley & Pakes (1996) Method

The key intuition of the Olley and Pakes method is that firms' *observable* decisions can reveal information about their *unobservable* productivity. Specifically, a more productive firm, expecting higher future returns, will choose to invest more in capital today.

The Olley-Pakes method formalizes this by assuming that investment (ι_{it}) is a strictly increasing function of a firm's current productivity (w_{it}) and its capital stock (k_{it}):

$$\iota_{it} = f(k_{it}, w_{it})$$

where ι_{it} denotes investment. Because this function is strictly increasing in productivity, we can invert it to express the unobserved productivity term as a function of two observed variables: investment and capital:

$$w_{it} = f^{-1}(k_{it}, \iota_{it})$$

We can then substitute this expression for productivity directly into the log-form production function. While the exact form of f^{-1} is unknown, it can be approximated with a flexible polynomial in k_{it} and ι_{it}. In the case of a third-order polynomial, this leads to the following estimable first-stage equation:

$$q_{it} = \theta_t^v v_{it} + \theta_t^l l_{it} + \beta_{1,t}k_{it} + \beta_{2,t}\iota_{it} + \beta_{3,t}k_{it}^2 + \beta_{4,t}\iota_{it}^2 + \beta_{5,t}k_{it}\iota_{it} + \beta_{6,t}k_{it}^3 + \beta_{7,t}\iota_{it}^3 + \beta_{8,t}k_{it}\iota_{it}^2 + \beta_{9,t}\iota_{it}k_{it}^2$$

By estimating this equation via OLS, we obtain a consistent estimate of the output elasticity of the variable input, $\hat{\theta}_t^v$, which is insulated from simultaneity bias. A full technical derivation, including all underlying assumptions, is provided in Appendix A1.

2.2.4 | Extensions to the Olley & Pakes Method — Levinsohn and Petrin (2003)

The Olley and Pakes method assumes that investment (ι) is strongly connected to productivity (w_{it}). While there are reasonable justifications for this in theory, the connection may not be strong in empirical reality.

Much of this is due to the investment timing of firms. Typically, the investments (or dis-investments) made by firms in response to productivity and expected potential, are not made period-on-period. Consider a car dealership with excellent management and skilled salespeople — in other words, high productivity. They see great potential in the business and want to expand. Will they, at the end of a good year, sell off their lot and move into one 15% larger in a different area, repeating this year-on-year? Or will they simply open a second location, either after saving for some time, or borrowing, both of which substantially lessen investment in other periods?

While the former approach to investment is not impossible, the latter is more reasonable and common, and this will desynchronize investment and productivity. A firm is not necessarily hyper-productive during the period in which they make large investments, nor is their productivity necessarily flagging when, rather than purchasing capital, they save or repay borrowing costs.

A related issue is that, as a variable, investment is more likely to contain zero values than most. After log-transforming capital during the estimation process, these observations will be unusable, and so the Olley and Pakes method often requires one to discard large swathes of data. Levinsohn and Petrin (2003), for example, had over half of their sample report zero investment.

The Levinsohn-Petrin (LP) method solves this by using a firm's expenditure on *intermediate inputs* (m_{it}) — a candidate for the chosen variable input for markup estimation — as the proxy for productivity. The logic is parallel to that of Olley and Pakes: a more productive firm will use more input. This approach has an empirical advantage because intermediate inputs are rarely zero and adjust more fluidly with productivity.

Under the LP approach, the productivity control is thus:

$$w_{it} = f^{-1}(k_{it}, m_{it})$$

which is similarly inverted and approximated with a polynomial. In the context of markup estimation, however, this introduces an identification problem. Because the variable input (v_{it}) is often a bundle that includes, or is identical to, the intermediate inputs used as the proxy, we must separate the output elasticity of m_{it} from its role in controlling for productivity. A simple one-step regression is insufficient for this task. The LP method thus recovers the output elasticity through a two-stage estimation: the first stage recovers the output elasticity of the factors not used in the productivity control, and the second stage uses a Generalized Method of Moments (GMM) estimator to isolate the output elasticity of intermediate inputs.

Given its practical advantages in preserving data and avoiding the timing 'lumpiness' of investment, we adopt the Levinsohn-Petrin method for our baseline markup estimates. The technical details of the two-stage procedure are outlined in Appendix A2.

3 | Revenue-Derived Markups

This section discusses the complications that arise when revenue data is used in place of output data for output elasticity estimation. Section 3.1 establishes the core problem with a review of Bond et al. (2021) who provide the fundamental criticism of revenue-derived markups, suggesting that they are uninformative. Section 3.2 then covers the revenue-based empirical results of De Loecker et al. (2020), and attempts to interpret their findings in

light of Bond et al.'s criticisms. Finally, we examine De Ridder et al. (2024), who propose that markup trends can be recovered from revenue data, providing a hopeful perspective on the validity of De Loecker et al.'s results.

3.1 | Biases Introduced by the Use of Revenue Data

This subsection begins by describing the problem of revenue data in markup estimation, explaining why it is often unavoidable, and developing an intuition as to how it biases markup estimates. We then review and derive the main result of Bond et al. (2021), which suggests that revenue-derived markup estimates should always equal 1. Finally, we explain why this theoretical result is not observed empirically, and assess the implications this has for the severity of Bond et al.'s criticism.

3.1.1 | Omitted Price Bias

Firm-level data with scope suitable for markup estimation is typically collected for tax or financial reporting purposes. Such reporting is done in terms of revenues and expenditures, and so we will rarely observe the physical units of input and output needed to conduct markup estimation in a theoretically consistent manner. Instead of estimating a production function, with quantity as the dependent variable, we will instead be forced to estimate a revenue function, with revenue as the dependent variable and expenditures as the independent variables. The use of expenditures as input-quantity substitutes is a relatively minor issue; thus, we treat them as being equivalent. The difference between output quantity and revenue as dependent variables, however, is massive, representing a fundamental shift in what we estimate.

Consider that, by estimating the production function, we are trying to determine how output responds to changes in input quantity — if I use another unit of input, how much will output increase? The answer to this question, the output elasticity, will be determined entirely by technological factors and the inherent characteristics of that input, both of which are exogenous — i.e., imposed from outside the model.

With output data we can observe this relationship cleanly because output can only be increased via the physical transformation of units of input into units of output. Revenue, on the other hand, can be increased both by the conversion of input to output, or by increasing the price of the output, and thereby the revenue per input.

Without observing and controlling for price, we will be unable to separate the output variation caused by input variation — that is, the output elasticity — from the effect of price on revenue. Thus, the estimate recovered from the revenue function will *not* be an unbiased output elasticity.

There are two conceptual camps with regard to interpreting the recovered estimate. The first is that of De Loecker et al. (2020) and De Ridder et al. (2024), who regard it as a price-biased output elasticity. This view follows from

the fact that the revenue function is simply the production function with price added to both sides. Starting from a multiplicative production function, we have:

$$Q_{it} = A_{it} V_{it}^{\theta^v} K_{it}^{\theta^k} E_{it}$$

Multiplying by price then yields the revenue function:

$$P_{it} Q_{it} = R_{it} = A_{it} V_{it}^{\theta^v} K_{it}^{\theta^k} P_{it} E_{it}$$

where R_{it} is revenue. In log-form, this is:

$$r_{it} = \theta_t^v v_{it} + \theta_t^k k_{it} + p_{it} + w_{it} + e_{it}$$

If we do not observe price, then our elasticity estimate ($\hat{\theta}_t^v$) will be the sum of the true output elasticity (θ_t^v) and an omitted variable bias:

$$\hat{\theta}_t^v = \theta_t^v + \frac{cov(v_{it}, p_{it})}{var(v_{it})}$$

This is a hopeful perspective, suggesting that estimates of the revenue function produce output elasticities that are merely biased by price — a bias that may not entirely obscure the underlying relationship. De Ridder et al. even propose that, if the bias remains constant, variation in markups is fully recovered.

The second camp, consisting of Bond et al. (2021) and Hashemi et al. (2022), views the use of revenue data as a more fundamental shift from output elasticity estimation to revenue elasticity estimation. This is an equally justified perspective. If we estimate the revenue function without observing or controlling for price, i.e.:

$$r_{it} = \theta_t^{v,r} v_{it} + \theta_t^{k,r} k_{it} + w_{it} + e_{it}$$

then the elasticity estimates ($\theta_t^{v,r}$) should be interpreted as estimates of the relationship between inputs and revenue. These papers thus focus on the consequences of replacing output elasticity (θ_t^v) with revenue elasticity ($\theta_t^{v,r}$) in the construction of the markup.

These perspectives seem to imply different things, resulting in a divergence in the treatment of the revenue-derived estimate between the two camps. Theoretically and empirically, however, they are identical.

The revenue elasticity of an input can be understood as a combination of two relationships: the relationship between input and output, and the relationship between output and prices. The former represents the direct productive use of the input — an input is used to create an output — while the latter reflects the price setting and output decisions of firms. In the absence of perfect competition, a firm's price will be tied, via demand, to their output; greater output necessitating lower prices and vice versa.

We can derive this formally. Starting with a generic revenue function, we have:

$$R_{it} = P_{it}Q_{it}$$

The derivative of this function with respect to the variable input is:

$$\frac{\partial R_{it}}{\partial V_{it}} = \frac{\partial Q_{it}}{\partial V_{it}} P_{it} + \frac{\partial P_{it}}{\partial Q_{it}} \frac{\partial Q_{it}}{\partial V_{it}} Q_{it}$$

We can then form the definitional revenue elasticity by multiplying by $\frac{V_{it}}{P_{it}Q_{it}}$:

$$\frac{\partial R_{it}}{\partial V_{it}} \frac{V_{it}}{P_{it}Q_{it}} = \frac{\partial Q_{it}}{\partial V_{it}} \frac{P_{it}V_{it}}{P_{it}Q_{it}} + \frac{\partial P_{it}}{\partial Q_{it}} \frac{\partial Q_{it}}{\partial V_{it}} \frac{Q_{it}V_{it}}{P_{it}Q_{it}}$$

Simplifying yields:

$$\frac{\partial R_{it}}{\partial V_{it}} \frac{V_{it}}{P_{it}Q_{it}} = \frac{\partial Q_{it}}{\partial V_{it}} \frac{V_{it}}{Q_{it}} + \frac{\partial P_{it}}{\partial Q_{it}} \frac{\partial Q_{it}}{\partial V_{it}} \frac{V_{it}}{P_{it}}$$

The second term on the right-hand side is the price elasticity of the input: price is changed alongside output to fit demand, and output is determined by the use of input. The first term is the output elasticity of the input (θ_t^v). We can partially re-express this using our previous notation for the output and revenue elasticities:

$$\theta_t^{v,r} = \theta_t^v + \frac{\partial P_{it}}{\partial Q_{it}} \frac{\partial Q_{it}}{\partial V_{it}} \frac{V_{it}}{P_{it}}$$

This reveals the fundamental equivalence of the two perspectives: the omitted variable bias term ($\frac{cov(v_{it}, p_{it})}{var(v_{it})}$) is the empirical incarnation of the theoretical price elasticity component ($\frac{\partial P_{it}}{\partial Q_{it}} \frac{\partial Q_{it}}{\partial V_{it}} \frac{V_{it}}{P_{it}}$). The price-biased output elasticity considered by De Loecker et al. (2020) is thus conceptually equal to the theoretical revenue elasticity referenced by Bond et al. (2021). The critiques levied by Bond et al. at the use of revenue elasticities therefore apply directly to revenue-based empirical strategies, even if they are framed as recovering price-biased output elasticities.

3.1.2 | Bond, Hashemi, Kaplan, and Zoch (2021)

The core finding and criticism of Bond et al. (2021) is that the endogeneity introduced by price — through the simultaneous setting of price and output according to demand — prevents the recovery of *any* information about markups. More specifically, if firms are profit-maximizing — which should be true in most cases — the theoretical markup will be exactly 1.

This is derived as follows. First, we define the revenue-derived markup as a markup which uses revenue elasticity ($\theta_t^{v,r}$) in place of output elasticity (θ_t^v):

$$\mu_{it}^{r} = \theta_{t}^{v,r} \frac{P_{it}Q_{it}}{P_{it}^{V}V_{it}}$$

It follows from the definition of the revenue elasticity that this is equal to:

$$\mu_{it}^{r} = \frac{\partial R}{\partial V_{it}} \frac{V_{it}}{P_{it}Q_{it}} \frac{P_{it}Q_{it}}{P_{it}^{V}V_{it}}$$

Using the chain rule and cancelling the price terms gives:

$$\mu_{it}^{r} = \frac{dR}{dQ_{it}} \frac{\partial Q}{\partial V_{it}} \frac{V_{it}}{Q_{it}} \frac{Q_{it}}{P_{it}^{V}V_{it}}$$

The definition of the standard markup is $\mu_{it} = \theta^{v} \frac{P_{it}Q_{it}}{P_{it}^{V}V_{it}}$, hence $\frac{\mu_{it}}{P_{it}} = \frac{\partial Q}{\partial V_{it}} \frac{V_{it}}{Q_{it}} \frac{Q_{it}}{P_{it}^{V}V_{it}}$, and so we can re-express this as:

$$\mu_{it}^{r} = \frac{dR}{dQ_{it}} \frac{\mu_{it}}{P_{it}}$$

The theoretical markup is $\mu_{it} \equiv \frac{P_{it}}{MC_{it}}$, so we can again rewrite as:

$$\mu_{it}^{r} = \frac{dR}{dQ_{it}} \frac{P_{it}}{MC_{it}} \frac{1}{P_{it}} = \frac{dR}{dQ_{it}} \frac{1}{MC_{it}}$$

and given that $\frac{dR}{dQ_{it}}$ is an expression of marginal revenue, we can finally express this as:

$$\mu_{it}^{r} = \frac{MR_{it}}{MC_{it}}$$

The profit maximizing condition for firms with market power is $MR_{it} = MC_{it}$; thus, we will observe $\mu_{it}^{r} = 1$ for all firms, regardless of their true markup. This renders the revenue-derived markup entirely uninformative, and useless as a measure of competition.

Why then, do people bother to estimate markups? Why do De Loecker et al. (2020), and others, not empirically observe this predicted markup of 1? This apparent discrepancy between theory and practice arises because Bond et al.'s derivation rests on the assumption that we use the *theoretical* revenue elasticity — or in other words, the firm's true revenue elasticity — in the markup calculation. As we discuss in the next subsection, this is almost impossible to satisfy empirically. The revenue-derived estimate we recover is not the theoretical revenue elasticity (nor the output elasticity) but a vague combination of relationships, the form of which has yet to be explicitly derived. In empirical applications, therefore, we will not satisfy $\hat{\theta}_{t}^{v,r} = \theta_{t}^{v,r}$, and estimated markups will diverge from 1.

This does not temper Bond et al.'s criticism, however. The production of estimates that differ from 1 is not a successful bypassing of this problem, but merely the result of an empirical failure to extricate the revenue elasticity from our revenue-specified production function. Without observing and controlling for price, output elasticities cannot be recovered from revenue data, and so this approach is fundamentally misdirected.

3.1.3 | Identification Issues

The identification issues which prevent recovery of the revenue elasticity — and thus the empirical observation of $\mu_{it}^{r} = 1$ — also pose a problem for the identification of the output elasticity. These issues are well documented in the literature, with an excellent description being given in Ackerberg et al. (2015). We will briefly review the fundamental problem here before extending the discussion to the revenue-specific case.

In Section 2.1.2, we derived the estimable form of the markup. The premise of this derivation was cost-minimization: that firms use the least amount of input to generate their target output — determined by profit maximization — given their production technology, capital, productivity, etc.

This creates an identification issue. If the variable input usage of firms is determined entirely by their levels of capital, labor, etc. — the independent variables in production function estimation — then the variable input will be a function of the productive factors, and we will not be able to separately identify their effects on output. For firms with the same levels of these factors, we will not observe any variation in input use, likewise any input variation we do observe will be accompanied by corresponding variation in the productive factors. This is described in Ackerberg et al. (2015) as an issue of functional dependence.

Identification of the output elasticity of the variable input will therefore only be possible if there are additional sources of variation which cause firms with the same technology and productive factors to use different amounts of variable input. One such source of variation, discussed by Gandhi et al. (2020), is the input price. Firms facing different input prices will naturally have different solutions to the cost-minimization problem, creating identifying variation. Another source is the demand faced by firms, which allows them to vary their output price and will thus similarly cause the profit-maximizing output of firms to differ.

With such sources of variation, identification of output elasticities becomes possible, but the same is not true for revenue elasticities. As we derived in 3.1.1, the reaction of a firm's revenue to input changes is driven simultaneously by two channels: the output elasticity of the input, and the *price* elasticity of the input. The latter elasticity, however, is impossible to identify.

In order to capture the price channel of the revenue elasticity, we would need to observe and compare multiple firms facing the same demand — i.e., possessing the same market power — with varying levels of input. This is infeasible. First, there is no reason for profit-maximizing, cost-minimizing firms with the same market power,

capital, labor, etc. to vary their input usage. In the output case, this identification problem was largely solved by variation in market power, but this channel is now endogenous to the model, with price variation being absorbed by revenue as the dependent variable. Second, and more fundamentally, holding market power fixed — so that we can estimate a given firm's revenue elasticity — requires us to *observe and control for market power*. This is an inherently contradictory requirement, as if we were able to control for market power, we would not need to estimate markups.

As we cannot control for market power, we will not be able to recover revenue elasticities. What is produced instead is an estimate of the combined relationships between output and input, and price and input. The form of this empirical estimate, and its implications for the recovery of markups, has not been fully explored in the literature. In Section 5 we will do just this, deriving an explicit expression for the revenue-derived estimate which clarifies its empirical properties and relation to the true markup. For now, however, it will suffice to say that revenue-derived markups will diverge from the uniformity predicted by theory.

3.2 | De Loecker et al. (2020): Empirical Trends in US Markups

With identification issues preventing recovery of the theoretical revenue elasticity, it is not surprising that De Loecker et al. (2020) find markups that diverge from 1. What is surprising, however — given Bond et al.'s theory suggests revenue-derived markups should be uninformative — is that De Loecker et al.'s estimates seem to capture a coherent trend, and that their levels are economically plausible.

De Loecker et al. (2020) estimate markups for a dataset of US firms over the period 1950-2016. They estimate a revenue-derived elasticity ($\theta_{jt}^{v,r}$) for each industry/year, with markups being calculated as:

$$\mu_{it} = \theta_{jt}^{v,r} \frac{P_{it} Q_{it}}{P_{it}^{V} V_{it}}$$

These are then aggregated into a revenue-weighted average for each period:

$$\mu_t = \sum_i \rho_{it} \mu_{it}$$

where ρ_{it} is the ratio of firm i's revenue to the total revenue of the sample in year t. Their central finding is that this weighted markup has maintained a steady upward trend for more than three and a half decades, rising from 1.21 in 1980 to 1.61 in 2016. This increase is notable not only for its magnitude — a 33% growth in the weighted average markup — but also because of the relentlessness of the observed trend. Over the 36-year period, markups constantly climb upward, with no significant drops or reversals.

De Loecker et al. decompose the change in the aggregate markup into three sources: evolution of the revenue-derived elasticity, changes in the revenue share ($\frac{P_{it}Q_{it}}{P_{it}^{V}V_{it}}$), and changes in the market shares of high and low-markup firms.

The first source — evolution of the revenue-derived elasticity — is eliminated as a potential driver of the trend, with De Loecker et al. finding that their estimated elasticity varies little over time. This is so much the case that they find almost no difference in their markup series after replacing their estimated elasticity with a constant of 0.85. There are two possible reasons for this: either the true output elasticity for their chosen variable input has remained static over the period of their sample, or the distortions introduced by revenue data make the estimated elasticity insensitive to the true elasticity. Either way, with changes in elasticity ruled out, the upward trend must be explained by the remaining two sources.

The authors attribute roughly two-thirds of the evolution of aggregate markups to the reallocation of economic activity to firms with high *estimated* markups. This gives high-markup firms more weight in the aggregate markup calculation, thus driving it up over time.

The remaining third of the markup trend is attributed to a trend in revenue shares. That is, a rise in revenue relative to expenditure on the chosen variable input over time. This results in an upward trend in $\frac{P_{it}Q_{it}}{P_{it}^{V}V_{it}}$ that, again, drives the aggregate markup upward.

From the constancy of the estimated elasticities, we can conclude that De Loecker et al.'s observed trend is not the product of unstable and noisy elasticities. The question remains however: why does this trend arise, and what do the estimated markups represent?

A possible answer is that the aggregate estimate reflects the true movement of markups in the US economy. Given the implications of Bond et al. (2021), however, this does not seem an entirely plausible and satisfactory explanation. Traina (2018) puts forth a more convincing interpretation, attributing the trend to De Loecker et al.'s choice of variable input (V_{it}).

De Loecker et al. take a cost of goods sold measure from Compustat as their variable input. This is a bundle of inputs which primarily includes the direct costs associated with production (wages, utilities, intermediate inputs, etc.). These costs are highly relevant to firms that deal in physical goods and their production goods — manufacturing, agriculture, wholesale, etc. As Basu (2019)[4] points out,however, cost of goods sold fails to capture the primary expenses of many modern industries. What, for example, is the cost of goods sold of YouTube? How much do material inputs contribute to the activities of Disney, or Myriad Genetics INC? These

[4] https://pubs.aeaweb.org/doi/pdfplus/10.1257/jep.33.3.3

firms, among others, are contained in the sample used by De Loecker et al., but their chosen variable input does not capture their salient costs.

Traina (2018) thus proposes that the observed trend in markups simply reflects the diminishing relevance of De Loecker et al.'s chosen input. Because modern firms increasingly utilize indirect inputs — such as marketing, insurance, administrative expenses, or recruitment — and interact less with physical production processes, revenues will naturally rise relative to cost of goods sold over time. This does not represent a true rise in markups, but a shift in the resources used by modern firms. A shift which is not observed when we consider only direct costs.

Traina (2018) re-estimates De Loecker et al.'s series using a much broader bundle that aggregates both cost of goods sold and sales, general, and administrative expense[5]. They find that revenue shares do *not* increase for this bundle, and the upward trend in markups is thus eliminated, remaining stably in the 1.1-1.15 range over the 1950-2016 period.

The comparison between the markup series of De Loecker et al. (2020) and Traina (2018) is interesting, and their choice of variable input plausibly explains their respective observed trends. However, two questions remain unanswered. First, if cost of goods sold is of decreasing relevance to the productive processes of modern firms, why does De Loecker et al.'s estimated elasticity remain constant across the 66-year period of 1950-2016? We can guess that this is related to the issues associated with revenue-derived elasticity estimated, but a formal exploration of these mechanisms is missing.

Second, why is Traina's aggregate markup estimate so much lower than De Loecker et al.'s? The former is persistently lower than the latter across the entire estimated period, with this gap widening as the series progresses. In 2016, Traina finds a markup of roughly 1.17, whereas De Loecker et al.'s estimate is 1.61. This wide gap at the later stages might be partially attributable to distortions in the revenue-derived elasticity — i.e., overestimation — but this does not account for the persistent gap between the two series.

We will answer these questions in Sections 5 and 6 respectively, precisely deriving the revenue-derived elasticity and identifying a channel that mechanically relates variable input choice to the level of estimated markups. Before presenting these solutions, however, we first examine De Ridder et al.'s more optimistic interpretation, which argues that revenue-derived markups can recover information about real markup trends.

[5] This encompasses the 'indirect inputs' — i.e., marketing, executive salaries, legal fees, insurance, etc.

3.3 | Recovering Markup Trends from Revenue Data

De Ridder et al. (2024) propose that revenue-derived markups, though biased in level, are highly correlated with true markups. They thus argue that such estimates can be used to find the "dispersion of markups across firms and trends in markups over time". This contrasts with the assertion of Bond et al. (2021) that revenue-derived markups are uninformative and, if true, would imply that the general trend of markups can be identified from revenue data. In this section, we review the derivations leading to this conclusion, the restrictive assumptions under which it holds, and its ultimate utility given these required assumptions.

3.3.1 | De Ridder et al. (2024): Framework and Estimators

The foundational logic of De Ridder et al.'s approach is familiar: a production function is estimated to recover an elasticity, which is then multiplied by revenue share to construct the markup. There is a key departure, however, in their estimator. Rather than the productivity-controlled regression that we and De Loecker et al. (2020) use, De Ridder et al. construct their own instrumental variable - generalized method of moments estimator (IV-GMM). This begins with a log-form Cobb-Douglas production function, which is simplified by omitting capital, labor, and the error term:

$$q_{it} = \theta^v v_{it} + w_{it}$$

Here, the lower-case variables represent *log deviations from the mean* (e.g., $v_{it} = \log V_{it} - \mathbb{E}[\log V_{it}]$). This facilitates the construction of De Ridder et al.'s IV-GMM estimator. As in previous sections, the deviations of output quantity, output elasticity of the variable input, input, and productivity are denoted by q_{it}, θ^v, v_{it}, and w_{it}, respectively.

This sets the stage for the IV-GMM estimator, which we derive more explicitly in appendix B. The critical identifying assumption is that productivity is independently and identically distributed (I.I.D.), implying that current productivity is uncorrelated with past input choices: $E[w_{it}v_{it-1}] = 0$. Under this assumption, the IV-GMM estimator which recovers the output elasticity is the value of $\hat{\theta}^v$ satisfying:

$$\left(\theta^v - \hat{\theta}^v\right)\mathbb{E}[v_{it}v_{it-1}] + \mathbb{E}[w_{it}v_{it-1}] = 0$$

Solved for $\hat{\theta}^v$, this is:

$$\hat{\theta}^v = \theta^v + \frac{\mathbb{E}[w_{it}v_{it-1}]}{\mathbb{E}[v_{it}v_{it-1}]}$$

Thus, the estimated output elasticity ($\hat{\theta}^v$) converges to the true output elasticity θ^v under the assumptions that productivity is I.I.D — and thus $\mathbb{E}[w_{it}v_{it-1}] = 0$ — and that there is persistence in input use, i.e., $\mathbb{E}[v_{it}v_{it-1}] \neq 0$.

With the IV-GMM estimator defined, De Ridder et al. turn to the core issue: the use of revenue data. Revenue is introduced into the production function by adding the log-deviation of price on both sides:

$$q_{it} + p_{it} = y_{it} = \theta^v v_{it} + w_{it} + p_{it}$$

The *revenue* IV-GMM estimator is then defined as the value $\hat{\theta}^v$ that satisfies:

$$(\theta^v - \hat{\theta}^v)\mathbb{E}[v_{it}v_{it-1}] + \mathbb{E}[p_{it}v_{it-1}] = 0.$$

Re-solving for $\hat{\theta}^v$ now yields:

$$\hat{\theta}^v = \theta^v + \frac{\mathbb{E}[p_{it}v_{it-1}]}{\mathbb{E}[v_{it}v_{it-1}]}$$

This is similar in structure to the output-based estimator, with the estimate being equal to the sum of the true elasticity and a ratio of expected values. Unlike the output-based estimator, however, we do not expect the numerator of the ratio ($\mathbb{E}[p_{it}v_{it-1}]$) to vanish. In imperfectly competitive markets, a firm's price will be determined simultaneously with its output (and thereby input use) via their demand curve. Thus, the output elasticity estimate will be persistently biased by $\frac{\mathbb{E}[p_{it}v_{it-1}]}{\mathbb{E}[v_{it}v_{it-1}]}$. This can be considered another empirical expression of the price-bias ($\frac{cov(v_{it},p_{it})}{var(v_{it})}$) derived in Section 3.1.1.

3.3.2 | Perfect Correlation Between True and Estimated Markups

De Ridder et al. (2024) claim that this persistent bias only affects the *level* of the estimated markup, and does not distort its variation. They proceed to demonstrate this explicitly by substituting the generic demand function:

$$p_{it} = -\sum_j d_{ijt}y_{jt}$$

where d_{ijt} is the cross elasticity of firm i's price with respect to firm j's quantity, into the revenue IV-GMM estimator. After some re-arranging, this yields the expression:

$$\hat{\mu}^r_{it} = \mu_{it}\left(1 - \frac{\sum_j \mathbb{E}\left[d_{ijt}(v_{jt} + \frac{w_{jt}}{\theta^v})v_{it-1}\right]}{\mathbb{E}[v_{it}v_{it-1}]}\right)$$

This states that the revenue-derived markup estimate ($\hat{\mu}^r_{it}$) is equal to the true markup (μ_{it}), multiplied by a complex bias term. Crucially, the bias term does *not* vary over time, thus the true markup and the revenue-derived estimate will be perfectly correlated with each other. This suggests that revenue-derived markups accurately reflect true markup trends, and consequently, that De Loecker et al.'s finding of a rise in markups may represent the true state of competition in the US.

3.3.3 | Connection to Bond et al. (2021)

While this is a powerful and suggestive result, it seems to contradict the core finding of Bond et al. (2021) that revenue-based markups are uninformative. De Ridder et al. attempt to reconcile these opposing conclusions by deriving the conditions under which $\hat{\mu}_{it}^{r} = 1$.

This derivation begins by considering a static oligopoly, a special case of demand where a firm's price depends solely on its own output, i.e.:

$$p_{it} = -d_{it} y_{it}$$

where d_{it} is the elasticity of firm i's price with respect to its own output. Under this structure, and with the profit-maximizing condition $\mu_{it} = \frac{1}{1-d_{it}}$, they derive two key expressions. The first is a re-expression of $\hat{\mu}_{it}^{r}$:

$$\hat{\mu}_{it}^{r} = \mu_{it} \mathbb{E}\left[\mu_{it}^{-1} \frac{v_{it} v_{it-1}}{\mathbb{E}[v_{it} v_{it-1}]}\right]$$

This again shows that the revenue-derived markup estimate is equal to the true markup multiplied by a constant, and that markup variation is thus recoverable. The second is an expression of the *expected value* — i.e., the level — of the markup estimate:

$$\mathbb{E}[\hat{\mu}_{it}^{r}] = \mathbb{E}[\mu_{it}] E[\mu_{it}^{-1}]$$

By Jensen's inequality, $\mathbb{E}[\mu_{it}^{-1}] \neq \mathbb{E}[\mu_{it}]^{-1}$. There will be an arbitrary discrepancy between the two, which dictates the value of $\mathbb{E}[\hat{\mu}_{it}^{r}]$. This has nothing to do with the true markup; hence, as predicted by Bond et al., the *level* of the estimated markup is indeed entirely uninformative. This expression also reveals the condition under which the estimated markup will be 1: constant true markups.

If markups are constant, we have $\mathbb{E}[\mu_{it}] E[\mu_{it}^{-1}] = \mu_{it} \mu_{it}^{-1}$, bypassing Jensen's inequality and giving us:

$$\mathbb{E}[\hat{\mu}_{it}^{r}] = \mathbb{E}[\mu_{it}] E[\mu_{it}^{-1}] = \mu_{it} \mu_{it}^{-1} = 1$$

3.3.4 | The Limitations of Perfect Correlation

While the expressions in the preceding subsection mathematically connect De Ridder et al. (2024) and Bond et al. (2021), they do not resolve the fundamental problem posed by Bond et al.: how can revenue-derived markups become informative given the failure to recover true output elasticities?

The answer is: they can't. De Ridder et al.'s argument rests on the assumption that output elasticities do not vary in a way that affects trends, and can therefore be ignored. The key mechanism enabling their derivations is the assumption that the output elasticity (θ^{v}) is a time and firm invariant constant. Under this assumption, we only need to estimate a single, non-time-varying elasticity for each industry, meaning that there will only be a single,

non-time-varying bias. It follows that estimated elasticities and markup will perfectly correlate with their true counterparts.

De Ridder et al.'s conclusion is therefore the artifact of a strong assumption. By estimating a single, time and firm invariant elasticity, we mechanically impose a constant bias term. The empirical validity of the perfect correlation hinges on this untestable assumption being correct. If output elasticities *do* vary over time — as would be the case if, as Traina (2018) argues, technological progress rendered 'direct costs' less relevant — the correlation is broken, and the estimated trend will be driven by a bias of unknown form.

This weakness underscores the need for a precise derivation of the revenue-derived estimate's form. Without it, the bias induced by revenue distortions is a black box, rendering empirically recovered trends uninterpretable.

4 | The Utility of Markups: Conceptual Issues and Alternative Measures

The prior sections have proceeded under the implicit assumption that the primary challenge of markup-based analysis is accurate estimation. In this section, we step back from methodological debate to ask a more fundamental question: if accurately estimated, do markups effectively capture competition in modern economies? Section 4.1 explores this question directly, pointing out conceptual weaknesses in the markup. Section 4.2 then considers an alternative measure — profit elasticity — which was developed largely in response to the limitations of markups.

4.1 | Do Markups Matter?

The conceptual core of the markup is the idea that, under perfect competition, the price charged by a firm should be exactly equal to marginal cost. While this may be a useful benchmark, the rigid treatment of divergence from this ideal as a lack of competition risks conflating market power with the legitimate rewards of the competitive process. This conflation arises, in part, because marginal cost — as the sole dimension of cost considered — fails to capture a critical channel of welfare improvement and competition: investment.

Firms invest in capital primarily to enhance product quality, increase production capacity, and reduce marginal costs through returns to scale, all of which are a net benefit to societal welfare. Yet, because these investments are typically fixed costs, they are invisible to the markup, which only observes the resulting marginal cost. Furthermore, the recoupment of these welfare-enhancing investments often requires that prices be maintained — preventing them from falling to match lower marginal costs — or even increased. This necessarily inflates the markup, potentially misrepresenting societally beneficial investment as anti-competitive pricing power.

This problem is compounded when fixed costs constitute not just strategic investments, but the essential infrastructure of modern firms. Echoing the insight of Traina (2018) regarding 'indirect' costs, a structural shift from variable to fixed expenditures will undermine the validity of measures which focus exclusively on the former. The billions of dollars that Google invests in servers and data infrastructure, for instance, do not result in lowered costs *per se*. Rather, investments in data storage and processing power are fundamental to their operations, functioning as quasi-inputs that remain invisible to a variable cost-based perspective.

By taking the $MR = MC$ condition as the competitive ideal, we therefore risk imposing an interpretive framework that is blind to welfare gains and the realities of modern firms. Berry et al. (2019), for example, review several industry-specific markup papers, finding that the relationship between welfare, markups, and fixed costs is circumstantial, and without a clearly defined relationship.

Similarly, Ganapati (2024) studies wholesalers, and finds that, while the industry has experienced concentration and markup inflation, it has also enjoyed overall growth. The expansion of the industry's dominant firms was accompanied by increased technological investment and the opening of new warehouses in convenient locations, driving an increase in demand and the development of the industry as a whole.

These increases in technological investment and convenience are not a form of benevolence on the part of larger firms. Rather, they represent dimensions of competition which are not observed through the channels of price and marginal cost.

This reveals a paradox unsolvable by markups: we can easily imagine a market that is highly competitive along the fixed-cost dimension, yet exhibits high markups. In the long term, firms constantly innovate and invest heavily in technology or assets which put them ahead of their competitors for some period. These competitors may not be able to undercut them in the short term — resulting in an unchallenged markup — but this reflects the dynamics of competition via high-cost investment, rather than a lack of competition altogether. If these investments are associated with high markups, then a markup measure will always tell us that the market is not competitive, but this will only be true in a constrained sense.

Of course, this is an idealized scenario, and does not reflect the reality behind every high markup industry. Berry et al. (2019) also give examples of industries where the relationship between markups and product improvement is not concrete: airline markups are sometimes associated with better products, sometimes not, while hospital markups are generally unrelated to product quality altogether.

Therefore, while markups have their role to play as an important diagnostic, they must be interpreted within the context of industry-specific dynamics. This discussion does not reject their utility but reframes their purpose: an estimated markup is a piece of evidence, rather than a self-contained conclusion.

4.2 | Profit Elasticity: A Potential Alternative

Given the conceptual limitations of markups — described in the preceding section — and the empirical complications associated with estimating them, the literature has room for alternative 'direct' measures of competition. Profit elasticity, proposed by Boone et al. (2007), presents itself as such an alternative, claiming to avoid many of the markup's issues.

Empirically, profit elasticity requires only a regression of profit on marginal cost, seemingly eliminating the requirement for output data. Conceptually, the relationship between profit and marginal cost that this estimates sidesteps the interpretive issues connected to increasing capital-dependence and investment-based competition: if innovation and investment increase efficiency, this dimension of competition will be accurately represented.

Profit elasticity is thus positioned as a promising alternative to the markup. In this section, we evaluate whether it delivers on this promise.

4.2.1 | The Weakness of Markups: Reallocation Effects

Boone et al. motivate the need for profit elasticity by highlighting a key weakness of markup measures: their inability to capture competition that occurs through aggressive firm interaction and the consequent reallocation of market share toward competitive firms. They frame competition as acting through two channels. The first is the reduction of entry barriers, which increases the number of competitors in a market. The second is the intensification of 'direct' competition — i.e., price wars, innovation, or improvements in product quality.

If competition is increased via firm entry, markups function correctly: weighted markup measures, such as the one used in De Loecker et al. (2020), will accurately pick up an increase in competition. But if competition is increased via direct methods, a weighted aggregate markup will assign increasingly more weight to the winners of this competition as their market shares increase. This is what the weighting is designed to accomplish, but if competitive firms are associated with higher markups, this reallocation of market share will be misinterpreted as a welfare-harming concentration of market power, and an increase in the aggregate markup.

Such a situation is rather plausible. We have already discussed fixed costs as an avenue for the connection between competitive firms and markups. Another possibility, raised by the authors, is firm efficiency. If some firms are more efficient, they can match their competitors' prices while producing at a lower cost, capturing the difference as profit and a high markup. While this rewards efficiency at minimal cost to consumers, it will be reflected as a high markup. Firms that derive their markups in this way will likely be advantaged in competition — being more efficient — and so competitive interactions will result in these high-markup firms gaining market share.

The critical point is that when competition is intensified, there is both a direct and indirect effect on the aggregate markup. The direct effect is obvious; firms will compete, pushing down the markup and eliminating inefficient firms. The indirect effect is less obvious. While the winners of the competitive process may have lowered their markups, if their markups are higher than those of the eliminated firms, the increased weight given to the winners will increase the weighted markup. The reallocation effect counterbalances the direct decrease in markups, and so the observed markup may remain stable or even increase, falsely indicating a reduction in competition.

4.2.2 | The Solution: Profit Elasticity

Profit elasticity attempts to solve this by controlling for the reallocation effect and measuring only the direct effect. The measure itself is simple, being a regression of profit on marginal cost:

$$\ln(\pi_i) = \alpha + \beta \ln(C_i)$$

where π_i is the profit of firm i, C_i is their marginal cost, and β is the estimate of the market profit elasticity, representing the percentage change in profits that occurs when the marginal costs of firm *i* increase by 1% (holding the marginal costs of other firms fixed). When the change in profits is negative, it indicates that firms with high marginal costs (i.e., inefficient firms) are punished with lowered profits, and that the market is therefore competitive. This premise is similar to that of markups. If a firm can maintain a high markup without being undercut, the market is uncompetitive. If a firm can maintain high profits while being less efficient than the rest of the market, then again, the market is uncompetitive.

The difference between the measures — allowing profit elasticity to see through the reallocation effect — is that profit elasticity captures the *relationship between firms*, rather than being a somewhat binary indicator of the state of the market. If competition is intensified via the punishing of inefficient firms, the profit-marginal cost relationship between firms will strengthen, regardless of their markups and market shares.

It is worth noting that this is vastly different from taking an unweighted average markup, which would severely misrepresent the competitive reality of an industry. With an unweighted average, a few small, struggling firms with low markups can completely counterbalance a high markup firm with 99% market share. Profit elasticity, on the other hand, is only sensitive to whether these firms are rewarded for efficiency. If the large, high markup firm makes more profit than the small firms despite having higher marginal costs, then profit elasticity will be low, indicating an uncompetitive market; if the opposite is true, then profit elasticity will be high, indicating competition.

4.2.3 | When Profit Elasticity Fails

Both Boone et al. (2007) and De Loecker et al. (2020) give strong evidence that the evolution of markup measures over time is heavily influenced by the reallocation effect; thus, profit elasticity seems an appealing tool for empirical work. Unfortunately, several problems undermine its utility, potentially rendering it even less relevant as a competitive metric than estimated markups.

Firstly, while the relational nature of profit elasticity minimizes reallocation effects, it also causes the measure to be insensitive to competitive changes that do not create clear winners and losers. For example, suppose competition in a market intensifies, causing firms to cut their prices such that profits are uniformly reduced by 10%. Competition has clearly increased, but estimated profit elasticity will be unchanged. This is because while firms are more competitive across the board, the relative, between-firm relationship of profit and marginal cost has not changed. With no relative change, the relationship between profit and marginal cost will remain the same, and profit elasticity will detect nothing. In this respect, the measure is much inferior to markups, which we would observe decreasing in aggregate.

Secondly, and even more problematic, is that the validity of the profit-marginal cost relationship as a competitive measure is predicated on the assumption that profits can only be generated through competitive means. If more profitable firms are simply more efficient, then this will hold, but in that case, profits should always be correlated with efficiency, and it is unclear how an uncompetitive, low profit elasticity scenario could arise. This is troubling given that the primary purpose of a competitive measure is to detect uncompetitive markets; if all profit is derived from efficiency and competitiveness, why do we need to formulate competitive measures at all?

The distortive effect of this flaw can be revealed through a Cournot equilibrium exercise. Consider a market for cell phones facing linear demand, $P(Q) = 100 - Q$, supplied by three firms: firms A and B (efficient, $C = 10$) and firm C (less efficient, $C = 12$). The standard Cournot solution gives the following competitive equilibrium:

Firm	Marginal Cost	Profit
A	10	529
B	10	529
C	12	441

Regressing profit on marginal cost, we estimate the profit elasticity of this industry to be -1.01, meaning a 1% increase in marginal cost results in a 1.01% decrease in profit. This correctly reflects a competitive market where efficiency is rewarded.

Now, suppose that firm A engages in anti-competitive behavior, imposing vertical restraints by acquiring a controlling stake in the sole supplier of microchips, and capping rivals' output at $Q_B = Q_C = 20$. Firm A, now an unconstrained monopolist over the remaining demand, increases its own output to $Q_A = 25$. The market price after these changes rises to $P = 35$, resulting in the following market outcome:

Firm	Marginal Cost	Profit
A	10	625
B	10	500
C	12	460

Relative to the inefficient firm c, firms a and b are now capturing an even greater share of the market profit. Of course, the entirety of this increase is being reaped by firm A, and only because of their anti-competitive obstruction of the other two firms, but this is not visible in the data. In the data, the relationship between profit and marginal cost has strengthened, hence re-estimation of the profit elasticity yields an estimate of -1.46. Following the intended interpretation, this suggests the industry has grown significantly more competitive.

The imposition of vertical restraints by firm A — which, in reality, would likely violate competition law — may seem like an unrealistically extreme example, but that is exactly the point. For a firm with relatively low marginal cost, any increase in profit will strengthen the simple relationship between profit and marginal cost. The way this profit is achieved is completely irrelevant. The firm's methods may be blatantly anti-competitive, involving collusion with suppliers or other firms, but this does not matter. Profit elasticity, by considering only the final outcome, systematically mistakes the leverage of market power for the rewards of efficiency.

4.2.4 | Profit Elasticity with Average Variable Cost

This confusion between market power and well-rewarded efficiency is worsened by the fact that, empirically, we do not observe marginal cost.

Boone et al. suggest that marginal cost can be approximated with average variable cost (AVC). By itself, this is reasonable. Both are cost measures which are increasing in output, and one of the determinants of AVC is marginal cost. In fact, holding all else equal, lower marginal cost will result in lower AVC, meaning the two variables will move together.

Average variable cost is defined as the ratio of total variable cost (TVC) to output, that is:

$$AVC = \frac{TVC}{Q_i}$$

Boone et al. implicitly assume that firms have constant marginal costs. Thus, $TVC = C_i Q_i$, and we can rewrite AVC as:

$$AVC = \frac{C_i Q_i}{Q_i}$$

This form of AVC is also unobservable, however. As was so problematic for the markup estimates, profit elasticity too suffers from the fact that output and prices are rarely observed separately, forcing us to rely on revenue-based data. Boone et al. thus use a revenue-based approximation of AVC, which itself is meant to approximate marginal cost. This is:

$$AVC = \frac{C_i Q_i}{P_i Q_i}$$

With this, the profit elasticity regression becomes:

$$\ln(\pi_i) = \alpha + \beta \ln\left(\frac{C_i Q_i}{P_i Q_i}\right)$$

This specification relies on the ratio $\frac{C_i Q_i}{P_i Q_i}$ being analogous in some way to firm efficiency in the same way as marginal cost. This is plausible in a situation where market power is limited. In that case, if a firm has low variable cost and high revenue, it must be because they are producing efficiently. In an uncompetitive market, however, a firm can achieve a low cost-to-revenue ratio simply by using market power to raise its prices. This creates a spurious negative relationship where the firm with the most market power appears to be the most 'efficient'. In an uncompetitive industry, we would therefore expect the profit elasticity to be strongly negative. This implies a high level of competition and exactly the wrong result.

This empirically convenient approximation thus fundamentally distorts the validity of profit elasticity. Under the theoretical specification — which uses marginal cost — profit elasticity is only distorted to the extent that there are efficient but uncompetitive firms. The profit of those firms — derived through uncompetitive means — is incorrectly interpreted as evidence of a competitive process.

The AVC specification, however, completely waives the requirement that the high-market power firms be efficient. To strengthen the estimated relationship, firms no longer need to have high profit and low marginal cost, only high profit and low ‘average variable cost’. The group which satisfies this condition will include not only efficient and successful firms, but also inefficient firms which leverage their market power to extract profit. Hence, even an industry dominated by the *least* efficient firm will be described as competitive, and the higher that firm’s market power, the more competitive the industry will seem. This is a complete inversion of what a competitive measure should seek to capture.

4.2.5 | Empirical Implementations of Profit Elasticity: Fixed Effects

The preceding subsections established that profit elasticity fundamentally misinterprets competition. This issue is regularly exacerbated, however, by a methodological choice common in the literature: the use of fixed effects.

Examples of fixed-effects implementations of profit elasticity include Boone et al. (2007) — the progenitors of the measure — and Fabling and Maré (2019) who estimate profit elasticities from New Zealand data. Boone et al. explain that fixed effects are meant to control for inflation, 'cyclical effects', and, primarily, observational errors in the profit and cost data which may arise due to misclassification of a firm's products (consider that the revenues from Coca-Cola merchandise such as clothing would be classified as beverage manufacturing revenue).

Accounting for these imperfections is a reasonable goal; however, this may be a case where the cure is worse than the disease. Fixed-effects — also known as the 'within' estimator — completely undermine the purpose and foundational logic of profit elasticity as an estimate of the reward for efficiency *between* firms.[6]

Consider that a firm fixed-effects profit elasticity specification estimates the effect of deviations from firm-average marginal costs on deviations from firm-average profit. That is:

$$\ln(\pi_{it}) - \ln(\bar{\pi}_i) = \alpha + \beta(\ln(c_{it}) - \ln(\bar{c}_i))$$

where c denotes marginal cost. What does this regression tell us? Assume that we find a large, negative profit elasticity coefficient. This means that within a given industry, when a firm's marginal costs rise above their average, their profit falls below their average. This superficially aligns with the interpretation of profit elasticity as an estimate of the punishment for inefficiency — in this case, indicating that inefficient firms are indeed punished. In reality, however, there is a vast difference between finding that the profit of firms decreases as their marginal costs increase, and finding that high-marginal cost firms are punished with low profit. The latter is a direct indicator of the current competitive state of an industry, whereas the former can be true independent of the presence of competition.

The fundamental issue is that the fixed-effects estimator only uses the internal period-to-period variation. Consequently, it produces an estimate of the 'within' effect that tells us about the relationship between profit and cost *within each firm over time*, but reveals nothing about the competitive state of the market at any given moment. It is entirely possible for every firm to have a negative, within-firm profit elasticity, while, in every period, the least efficient firms reap the highest profits.

[6] Also note that the use of time and firm fixed effects necessitates that our profit elasticity estimate will cover multiple periods, i.e., the profit elasticity for industry A over the period of 2005-2010. Longer periods are preferred as they will provide more within-firm variation to base the estimates on. This makes year-to-year analysis impossible.

This divergence between what profit elasticity should measure — the reward for efficiency in a market — and what fixed effects measure in actuality, can be demonstrated with a simple simulation. Consider the following two-firm, two-period market:

Firm	Period	Marginal Cost	Profit
A	1	100	10000
A	2	110	9500
B	1	25	2000
B	2	30	1800

In both periods, firm A has a higher marginal cost than firm B but also generates more profit. In other words, the least efficient firm in this market is the most successful. By profit elasticities' standard of competition — in which efficient firms should be the most profitable — this is not a competitive industry.

This is reflected in the between-firm regression. The estimated profit elasticity in period 1 is 1.66, and in period 2 it is 1.90, with the positive coefficients implying that the market is uncompetitive. The fixed effects regression, on the other hand, yields an estimate of -1.08. Contrary to our intuition and the between-firm estimate, this suggests a high degree of competition.

Thus, by using fixed effects, we will accurately detect the firm-specific negative relationship between profit and marginal cost over time. This is not consistent with the original intention to measure the *between firm* reward for efficiency, however, and cannot be interpreted as such. A negative fixed-effects estimate merely tells us that firms will be more profitable than usual when their marginal costs are reduced. If this is not a tautology, it is very close.

This dissonance is only worsened by the use of average variable cost as a substitute for marginal cost. Under an AVC, fixed-effects specification, the only condition for profit elasticity to be negative is that, when a firm's revenue increases relative to its costs, or its costs decrease relative to its revenue, its profit will increase. This is almost true by definition, rendering the common empirical implementation of profit elasticity uninformative as a measure of competition.

4.3 | Summary

The preceding sections have established the theory and empirics of markup estimation, and have reviewed the complex literature discourse regarding the validity of markups. The core methodological debate revolves around De Loecker et al. (2020)'s results and their use of revenue data. While Bond et al. (2021) demonstrate that

revenue-derived elasticities are uninformative about true markups, De Ridder et al. (2024) contend that trends may remain recoverable despite biased levels. On the fringes of this debate is Traina (2018) who suggests that De Loecker et al.'s choice of variable input is consequential. Additionally, the comparison between the series of Traina and De Loecker et al. leads to two crucial questions which are not resolved by the current literature: why do De Loecker et al.'s estimated elasticities not respond to the (apparent) diminishing productivity of cost of goods sold, and why do Traina's estimates lie persistently below De Loecker et al.'s?

The following sections address these questions by developing comprehensive theoretical frameworks for two biases. The first is revenue bias, which is well understood in theory but lacks an empirically relevant explanation. The second is misspecification bias, which we novelly identify.

We derive precise expressions for these biases in Sections 5 and 6 respectively. We give them interpretable forms, revealing their effects on estimated markups, and explore these effects in simulation. In Section 8, we then review our empirically estimated markup series, constructed from New Zealand data. These estimates verify our theoretical hypothesis, leading to a strongly evidenced conclusion on the validity of empirical markups.

5 | Empirical Revenue Bias

The literature well establishes that revenue-derived markup estimates will be distorted. The empirical form of this distortion, however, is not provided, making it difficult to discern what revenue-derived level and trend estimates actually capture. In this section, we fill this gap by deriving an explicit expression of the revenue-derived elasticity estimate, and exploring its relation to the levels and trends of true markups. The analysis proceeds in four parts. Section 5.1 begins by describing our configuration of a simulated market. In Section 5.2, we use this simulation to test the efficacy of our empirical strategy, outlined in Section 2.2, and the relation between output and revenue-derived markups. Section 5.3 presents our first core theoretical contribution: the derivation of a precise expression for the revenue-derived elasticity estimate. Finally, Section 5.4 tests the properties of this expression under simulated trends of true productivity and market power, allowing us to observe how estimated trends form in relation to underlying movements in competition and productivity.

5.1 | Simulated Market Configuration

5.1.1 | Market Setup and Equilibrium

We simulate 1500 firms over 10 periods, summing to 15,000 observations. Each firm is assigned values of capital, productivity, etc., which are then used, alongside a production function and demand framework, to solve for their profit-maximizing levels of production and revenue.

We start with a generalized Cobb-Douglas production function:

$$Q_{it} = A_{it}K_{it}^{1-\theta_t}M_{it}^{\theta_t}E_{it}$$

where M is intermediate input and θ is the output elasticity of the intermediate input. Note that θ is constant across firms within a period but can vary over time.

The demand for the firm's output is specified by the price function:

$$P_{it} = T\Omega_{it}Q_{it}^{-\frac{1}{\eta_{it}}}, \quad \eta_{it} = 1 + \frac{1}{\Omega_{it}}$$

where T is a demand shifter (i.e., market size) and Ω is a market power shifter, affecting the price charged by firms both directly and by determining the demand elasticity they face. This is a non-standard demand specification in that it is not derived from a particular utility function; however, it accomplishes two things very well. Firstly, and most importantly, variation in Ω_{it} enables identification of the production function through estimation. If Ω_{it} were constant across firms or was not included in the demand specification at all, then the equilibrium output of firms would be determined solely by optimization of each firm's factors of production. For a given level of productivity and capital, there will be an optimal level of output and intermediate input usage. Hence, conditional on A_{it} and K_{it}, there is no independent variation in M_{it}. And with no independent variation, there is no way to separately identify the elasticities of capital and the intermediate inputs.

Exogenous variation in Ω_{it} solves this problem by shifting the equilibrium output of firms. Even for firms with identical bundles of productive factors, firms with high market power will push production and input usage beyond that of firms with lower market power.

The second benefit of this demand specification is that it allows us to control markups precisely. Markups under this specification are given by:[7]

$$\mu_{it} = 1 + \Omega_{it}$$

and so we can directly determine the average simulated markup by tweaking market power. This is useful later when we try to ascertain how well markup trends are captured by revenue-derived markup estimates.

Regardless, with defined production and price functions, we can construct the profit function that will be used to determine the equilibrium output of the firms. This is given by:

[7] We derive this explicitly in Appendix C.

$$\pi_{it} = Revenue - Cost = Q_{it} * T\Omega_{it} Q_{it}^{-\frac{1}{\eta_{it}}} - rK_{it} - \frac{P^M Q_{it}^{\frac{1}{\theta_t}}}{\left(K_{it}^{1-\theta_t} E_{it} A_{it}\right)^{\frac{1}{\theta_t}}}$$

where r is the cost of capital, P^M is the price of the intermediate input, and the rightmost term is variable cost, derived by solving the production function for M_{it} and multiplying by P^M. We can then determine each firms profit maximizing output by setting the derivative of profit with respect to output to zero, and solving for Q_{it}. The cost of capital does not change with output, so this is equivalent to the derivative of revenue minus the derivative of variable cost (with respect to output):

$$\frac{\partial \pi_{it}}{\partial Q_{it}} = \frac{\partial R_{it}}{\partial Q_{it}} - \frac{\partial VC_{it}}{\partial Q_{it}}$$

The derivative of revenue is:

$$\frac{\partial R_{it}}{\partial Q_{it}} = T\Omega_{it} Q_{it}^{-\frac{1}{\eta_{it}}} - \frac{1}{\eta_{it}} T\Omega_{it} Q_{it}^{-\frac{1}{\eta_{it}}-1} Q_{it}$$

which we get by applying the product rule to our revenue expression $R = Q_{it} * T\Omega_{it} Q_{it}^{-1/\eta_{it}}$. The derivative of variable cost is:

$$\frac{\partial VC_{it}}{\partial Q_{it}} = \frac{P^M}{\theta_t (K^{1-\theta_t} EA)^{\frac{1}{\theta_t}}} * Q^{\frac{1}{\theta_t}-1} = b_{it} Q^{\frac{1}{\theta_t}-1}$$

With all terms in the ratio being fixed parameters, we let this ratio equal b_{it} for simplicity. This allows us to represent the derivative of firm profit with respect to output as:

$$\frac{\partial \pi_{it}}{\partial Q_{it}} = T\Omega_{it} Q_{it}^{-\frac{1}{\eta_{it}}} - \frac{1}{\eta_{it}} T\Omega_{it} Q_{it}^{-\frac{1}{\eta_{it}}-1} Q_{it} - b_{it} Q_{it}^{\frac{1}{\theta_t}-1} = 0$$

This does not have a closed-form solution which can be determined from the values of the productive factors. We instead implement a numerical solver for the first order condition:[8]

$$T\Omega_{it} Q_{it}^{-\frac{1}{\eta_{it}}} - \frac{1}{\eta_{it}} T\Omega_{it} Q_{it}^{-\frac{1}{\eta_{it}}-1} Q_{it} = b_{it} Q_{it}^{\frac{1}{\theta_t}-1}$$

This gives us the equilibrium output of each firm, from which we can derive price, revenue, intermediate usage, and expenditures — everything necessary for the estimation of elasticities and markups.

[8] Refer to Appendix D for more detail.

5.1.2 | Parameter Generation

In our base specification we simulate 1500 firms over 10 periods for a total of 15,000 observations. The outcome of the simulation and of the firms is determined almost entirely by the variables of the production function, as well as market power. Table 1 contains the summary statistics for these variables. Productivity and the production function error are distributed normally across all firm/year observations with arbitrary variances[9]. Capital and market power, on the other hand, are generated with more intention.

Table 1 — Simulated Market Summary Statistics

Variable	Mean	St. Dev.	Median	10th Pct.	90th Pct.	Observations
Capital	60.06	50.2	42.7	10.6	144	15,000
Mkt. Power	0.185	0.135	0.175	0	0.368	15,000
Productivity	1.000	0.202	0.999	0.740	1.26	15,000
Error	1	0.03	1	0.962	1.04	15,000

Table 1: Summary statistics for simulated capital, market power, productivity, and error.

For capital, firms are first assigned a size (small, medium, large, very large), and capital is then normally distributed among firms within the same size-group. Having distinct groups of varying capital sizes is necessary because we do not simulate firm investment. For the firms in each size group, capital in the first period is generated as a normal distribution, and capital in the subsequent periods is simply re-generated using the same distribution. Hence, while a firm's capital may vary from year to year, there is no trend or evolution, nor a connection to period-on-period investment. For our second-stage GMM to be effective, however, we require the lag of capital to be highly correlated with current-period output. Having groups of varying mean capital is a simple solution, generating enough variation to identify the capital coefficient, while also having the lag of capital be a good indication of current output.

Additionally, while we do not simulate investment, it is a necessary element of the control function strategy, acting as a control for productivity. We therefore generate an investment variable which is arbitrarily related to productivity and capital by the equation:[10]

$$I_{it} = 200w_{it} + 100w_{it}^2 + k_{it}$$

Finally, the two key parameters in our simulation are market power and the output elasticity of the variable input — Ω_{it} and θ_t. In our baseline specification, the output elasticity is a constant 0.5, while market power is normally

[9] Recall that these variables will not be individually identified in our estimation of the production function but rather act as noise which must be eliminated through our control method.

[10] This arbitrary relation ensures that productivity and investment are directly related. There is thus no need to generate our productivity variable according to a Markov process to satisfy the requirements of the control method.

distributed around 0.175, with a cutoff at 0, resulting in a mean of 0.185. Note that this entails an average markup of 1.185. We then test several other specifications which allow Ω_{it} to increase by 0.025 per year, θ_t by 0.02 per year, or both at the same time. This clarifies how revenue-derived markup estimates behave relative to the true markup when true markups and elasticities are variable, allowing us to evaluate the viability of recovering markup trends from revenue data.

5.2 | Simulation Results

We now present the results of the simulation. Section 5.2.1 compares the levels of the revenue and output-derived elasticity estimates, confirming the severe distortion introduced by revenue data. Section 5.2.2 then analyzes their correlation, assessing whether, despite this level bias, revenue-derived markups can recover trends.

5.2.1 | Failure to Recover Output Elasticity

We begin by applying the standard control function approach — outlined in Section 2.2 — to the simulated physical output and revenue data. This yields an output-derived elasticity, representing the theoretically correct approach, and a revenue-derived elasticity, which we expect to be distorted. Table 2 compares these estimates.

Table 2 — Simulated Elasticity Estimates

Variable	True Value	Output Derived	Revenue-Derived
θ^v	0.5	0.499	1.044

Table 2: Estimated elasticity of the intermediate input for the base specification (static market power and true elasticity). 'True value' denotes the true output elasticity as defined in our model. Output and revenue-derived are the estimates produced from the output and revenue data respectively. We estimate elasticities separately by year, then take the average.

Our estimation strategy successfully recovers the true output elasticity when using physical output data. However, the estimate derived from revenue data is severely biased (1.044 vs 0.5). Using the revenue-derived output elasticity as a substitute for the output-derived elasticity will thus result in markup estimates that are meaningless in level.

5.2.2 | Correlation Between True and Estimated Markups

De Ridder et al. (2024) claim that, in spite of the bias in the revenue-derived elasticity estimate, markup trends can be recovered from revenue data. This is predicated on strong or even perfect correlation between the true and revenue-derived elasticities, which itself is predicated on the assumption that the true output elasticity is constant. Table 3 tests this claim by examining this correlation under differing market specifications, including those which break the constant elasticity assumption.

Table 3 — Correlations Between True and Revenue-Derived Markups

Condition	Base Specification	Trending Market Power	Trending Elasticity
Within Year	1.00	1.00	1.00
Overall	0.99	0.91	0.96
Yearly Average	0.91	-0.19	-0.13

Table 3: Correlation between true and revenue-derived markups. The 'condition' column describes the set of elasticities used for comparison. 'Within Year' denotes the correlation between the true and revenue-derived markups across firms within a given year. 'Overall' lifts this restriction and gives the correlation across all years and firms. 'Yearly Average' uses only the average markup of each year.

Superficially, Table 3 appears to support De Ridder et al.'s claims. Correlations are high under the "Within Year" and "Overall" conditions, mirroring their own finding of a high correlation (0.93) which they interpret as evidence that revenue-derived markups are "highly informative of true markups."

This high correlation, however, is a mechanical artifact. It arises only when the true output elasticity is restricted such that it has no or very little variation. Under the base specification, the true elasticity is constant, satisfying the constant elasticity assumption. On the other hand, under the "Within Year" and "Overall" conditions for trending specifications, the true elasticity is held constant *within* each year, and the massive number of cross-sectional firm observations dominates the limited time-series variation, effectively masking the trend.

With the constant elasticity assumption satisfied, we have mechanically ensured that none of the variation in markups can be attributed to output elasticities. We will have the true markup:

$$\mu_{it} = \theta \frac{P_{it}Q_{it}}{P_{it}^{M}M_{it}}$$

and the estimated markup:

$$\hat{\mu}_{it} = \hat{\theta} \frac{P_{it}Q_{it}}{P_{it}^{M}M_{it}}$$

Where both θ and $\hat{\theta}$ are constant. All of the variation in μ_{it} and $\hat{\mu}_{it}$ will be generated by the common term $\frac{P_{it}Q_{it}}{P_{it}^{M}M_{it}}$, mechanically forcing perfect correlation regardless of the values of the constant elasticities.

As Traina (2018)'s finding strongly implies the evolution of output elasticities over time is empirically likely. Thus, this result of 'perfect correlation' — reliant on the mechanical elimination output elasticity variation — tells us little about the relationship between true markups and estimates in empirical contexts.

The critical test is the "Yearly Average" correlation, which isolates the time series variation used to assess aggregate trends. Here, under more realistic and empirically relevant conditions of trending market power or elasticity, the correlation not only breaks down but inverts. This demonstrates that the claim that revenue data

can reliably capture markup trends is not robust to violations of the constant elasticity assumption, providing no evidence that time-trends can be recovered.

5.3 | The Empirical Revenue-Derived Elasticity

The results so far demonstrate that revenue-derived markup levels are meaningless, and their trends are unreliable. This leaves us with no framework to interpret empirical estimates, making it impossible to discern whether they reflect true markups or statistical noise.

To resolve this dead-end, we derive the precise expression for the revenue-derived elasticity estimate. This serves two crucial purposes: it formally explains the correlative failure of the non-constant elasticity specifications observed in Section 5.2.2, and provides an analytical framework through which to understand what the revenue-derived estimate actually captures.

We can motivate the derivation with some intuition. In Section 3.1.1, we described the revenue-distortion as an omitted-variable-bias on the estimates of the production function:

$$\hat{\theta}_t^v = \theta_t^v + \frac{cov(v_{it}, p_{it})}{var(v_{it})}$$

While this expression is too simplistic to capture the full complexity of the bias, it provides useful intuition. In general, the bias on the revenue-derived elasticity estimate will depend on the relationship between prices and inputs, which will itself depend on the market power of firms. When market power is high, firms act like monopolists, contracting output and raising prices. This creates a strong negative correlation such that biases estimated downward. Conversely, when market power is low, firms must keep prices relatively constant regardless of their output, weakening the correlation and reducing the bias. Under perfect competition, for example, firms face a constant market price independent of their output, reducing the covariance to 0.

This pattern emerges clearly in our simulated market. Under our evolving market power specification, the correlation between price and input for firms with markups below the sample median of 1.3 is -0.075. For firms over the median, this increases sharply in magnitude to -0.71.

The omitted-variable-bias framework thus provides a vague directional prediction. When market power is low, true elasticity trends will be recovered. When market power is high, estimates will begin to diverge. This is a useful heuristic, but fails to capture the full complexity of the distortion.

We now formally derive the probability limit of the revenue-derived elasticity estimate, starting from the underlying economic model of supply and demand.

Our production function, in logs is:

$$q_{it} = \omega_{it} + (1 - \theta_t)k_{it} + \theta_t m_{it} + e_{it}$$

and the price function is:

$$p_{it} = t + \ln(\Omega_{it}) - \frac{1}{\eta_{it}}\, q_{it}, \quad \eta_{it} = 1 + \frac{1}{\Omega_{it}}$$

The revenue function, in logs, is simply $r_{it} = p_{it} + q_{it}$, so we can substitute both our price and production functions into it:

$$r_{it} = t + \ln(\Omega_{it}) - \frac{1}{\eta_{it}}\,(\omega_{it} + (1 - \theta_t)k_{it} + \theta_t m_{it} + e_{it}) + \omega_{it} + (1 - \theta_t)k_{it} + \theta_t m_{it} + e_{it}$$

$$= \left(1 - \frac{1}{\eta_{it}}\right)(\omega_{it} + (1 - \theta_t)k_{it} + \theta_t m_{it} + e_{it}) + t + \ln(\Omega_{it})$$

Now recall that $\eta_{it} = 1 + \frac{1}{\Omega_{it}}$, and our simulated markup is given by $\mu_{it} = 1 + \Omega_{it}$, hence:

$$1 - \frac{1}{\eta_{it}} = \frac{1}{1 + \Omega_{it}} = \frac{1}{\mu_{it}}$$

We can then rewrite the revenue function as[11]:

$$r_{it} = \frac{\omega_{it}}{\mu_{it}} + \frac{(1 - \theta_t)}{\mu_{it}}k_{it} + \frac{\theta_t}{\mu_{it}}m_{it} + t + \ln(\Omega_{it}) + \frac{e_{it}}{\mu_{it}}$$

This is simply a restatement of the revenue function, but it reveals that the revenue function's coefficients are the output elasticities *scaled by the inverse of the firm's own markup*.

In Appendix E, we complete the derivation to show that the probability limit of the revenue-derived estimate is:

$$plim\ \hat{\theta}_t^r = E\left[\frac{\theta_t}{\mu_{it}}\right] + \frac{cov\left(\frac{\theta_t}{\mu_{it}}, m_{it,\perp}^2\right)}{var(m_{it,\perp})} + \frac{cov\left(m_{it,\perp}, d_{it,\perp}\right)}{var(m_{it,\perp})}$$

where $d_{it} = \ln(\Omega_{it}) + \frac{e_{it}}{\mu_{it}}$, and $\perp$ is used to denote the residuals of a variable after having regressed on our controls for capital and productivity — i.e., $m_{it,\perp}$ is the variation in intermediate input usage that is not explained by k_{it} or ω_{it}.

This expression reveals why revenue-derived estimates are fundamentally uninformative about true markups:

[11] The exact expression for the revenue function will depend on the assumed structure of demand. Hence, this specific expression is tied to our model. But something similar can be derived for other demand structures. Bond et al. (2021), for example, assuming CES demand, derive a revenue function which is approximately:
$r_{it} = t + \frac{\beta_k}{\mu}k_{it} + \frac{\beta_l}{\mu}l_{it} + \frac{\beta_m}{\mu}m_{it} + [\frac{\omega_{it}}{\mu_{it}} + e_{it}]$. This is very similar to our expression.

First, $\hat{\theta}_t^r$ is a complex non-linear function — $\hat{\theta}_t^r(\Omega_{it})$ — of markups. Rather than reflecting the true elasticity (θ_t) the estimate depends on three terms that each respond differently to changes in market power. The first term is the average of $\frac{\theta_t}{\mu_{it}}$, which generally decreases with markups. The second term is similar to an inverted $cov(v_{it}, p_{it})$, being negative at low levels of market power and positive at high levels. The third term meanwhile has an ambiguous relationship with market power, but generally starts high before converging to 0 as market power increases.

Second, the non-linearity of $\hat{\theta}_t^r(\Omega_{it})$ makes its relation to the true markup a matter of happenstance. As shown in Figure 1, $\hat{\theta}_t^r(\Omega_{it})$ is non-monotonic with several turning points. Whether estimated markups rise and fall with the true markup depends entirely on which term of the expression dominates at a given level of market power — a factor that cannot be controlled or observed by researchers.

If markups are increasing and the sample happens to be at a level where the covariance terms dominate, estimated markups may coincidentally correlate with true markups. At a different level of market power, the decrease in $E\left[\frac{\theta_t}{\mu_{it}}\right]$ may dominate, causing estimates to trend downward even as true markups rise. This explains the negative correlations observed in Section 5.2.2.

Crucially, even when true and revenue-derived markups move in the same direction, we are merely benefiting from coincidence rather than capturing the true relationship. A similarly valid method would be to determine trend direction by coin flip. The direction of movement in revenue-derived markups is fundamentally unpredictable and independent of true elasticity trends.

5.4 | Simulated Markup Trends

Figure 1 illustrates the non-linear relationship between true and revenue-derived markups by plotting their yearly averages. We simulate a steady increase in market power over 100 periods, starting from 0 and growing by 0.025 each period. Initially, the markups move together, but then several turning points cause their correlation to fluctuate between positive and negative, eventually stabilizing with inverse movement.

Figure 1 — Movement of True and Revenue-Derived Markups When Markups Trend

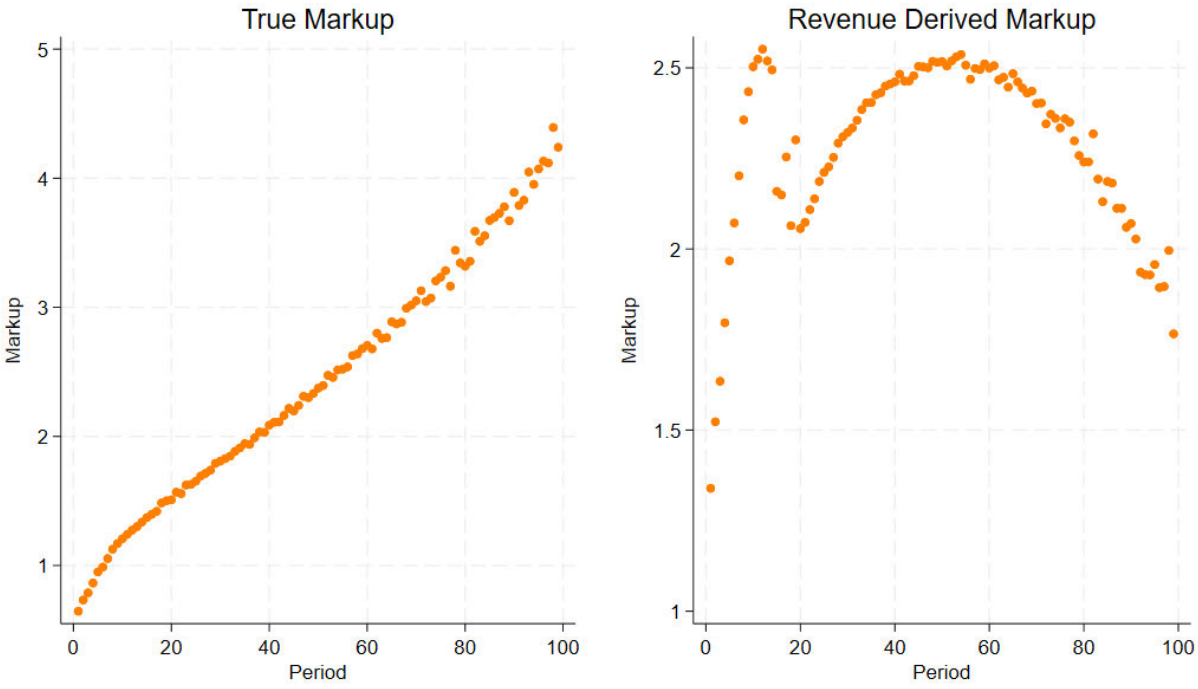

Figure 1: True and revenue-derived markups over 100 periods. Market power starts at 0, then increases by 0.025 per period. True output elasticity remains static.

This unpredictability extends to trends in the true output elasticity, as shown in Figure 2. When the true output elasticity increases, the first term in our derived expression, $E\left[\frac{\theta_t}{\mu_{it}}\right]$, increases. This is offset, however, by an increase in $var(m_{it,\perp})$, which results from firms producing more output and thus further leveraging productivity and capital differences. These opposing forces are roughly equal, leaving the estimated revenue elasticity ($\hat{\theta}_t^r$) close to constant. Meanwhile, the revenue-to-cost ratio ($\frac{P_{it}Q_{it}}{P_{it}^M M_{it}}$) decreases as firms expand output and reduce prices. With the ratio decreasing while the estimated elasticity stays roughly the same, the estimated markup falls even as the true markup remains static.

Figure 2 — Movement of True and Revenue-Derived Markups When Elasticity Trends

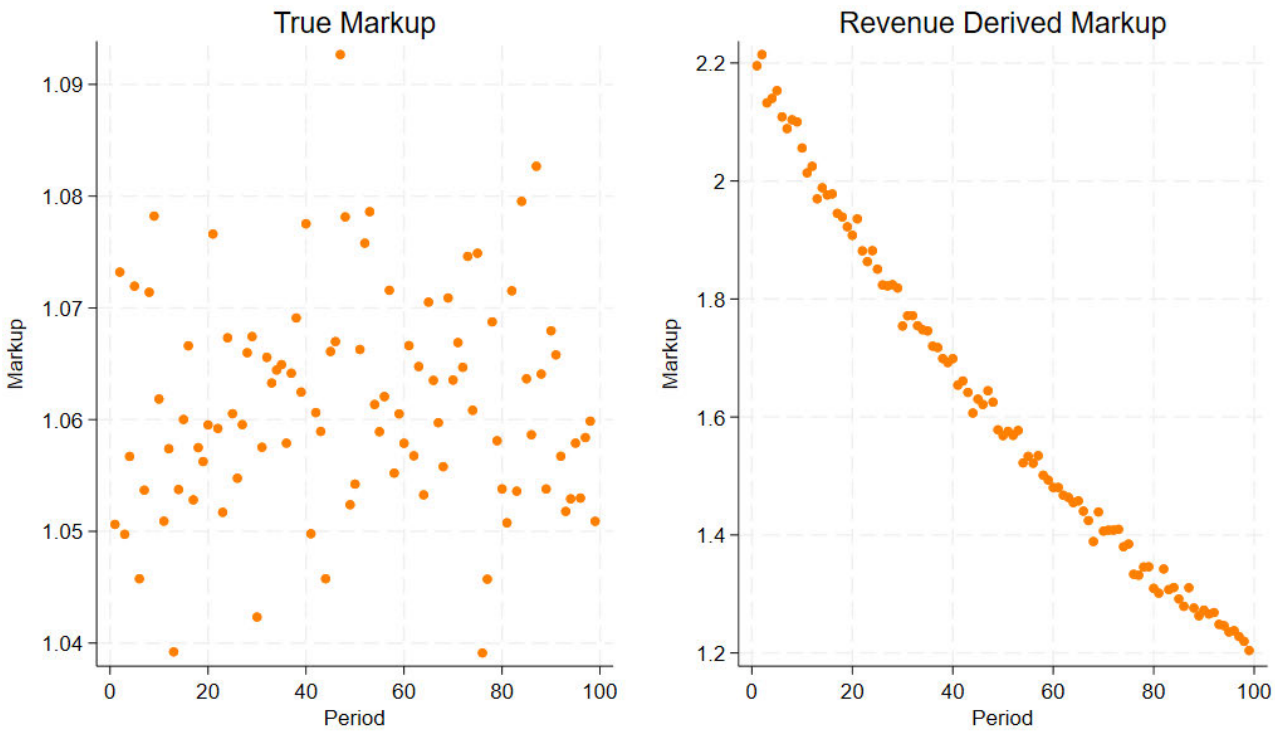

Figure 2: True and revenue-derived markups over 100 periods. True output elasticity starts at 0.5, then increases by 0.0035 per period. Market power remains static.

We therefore conclude that revenue-derived markup estimates cannot reliably recover the trend of true markups. The observed trends in revenue-derived markups may sometimes coincide with true trends, but this is both coincidental and indeterminable, making it indistinguishable from spurious correlation.

Moreover, the constancy of estimated revenue elasticities does not imply constancy of the true elasticities. As Figure 2 demonstrates, a constant $\hat{\theta}_t^r$ is perfectly consistent with a strongly trending true elasticity. The finding of constant elasticities in De Loecker et al. (2020) may be therefore be misleading, and provides no justification for De Ridder et al.'s constant elasticity assumption.

In summary, the revenue-based approach to elasticity estimation does not merely provide a biased estimate; it does not estimate the output elasticity at all. The object it recovers is an artifact — the product of a web of relationships influenced separately by market power and weighted by the unobservable distribution of markups. Furthermore, the inability of revenue-derived elasticities to capture true-elasticity trends is a fatal backdoor, rendering them completely unable to account for progress in firm processes. As shown by Traina (2018) this leads to exaggerated or even false trends in markups driven by shifts in input productivity. Without the ability to impose constant elasticities in empirical environments, we cannot recover long-term markup trends.

6 | Input Bundling and Misspecification Bias

The previous section has shown how the use of revenue data biases output elasticity estimates, making the veracity of the revenue-derived markup questionable. We now propose that there is a second source of bias in the literature-standard approach to markup estimation that arises even when prices and output quantities are observed: input bundling. This refers to the aggregation of individual inputs into bundles such as 'cost of goods sold' or 'material inputs' which are then taken as the cost-minimized input upon which the markup estimate is based. In this section we will describe the biases induced by input bundling. In Section 6.1 we set up the basic theoretical framework through which to understand input bundling and describe the conditions under which it is valid. In Section 6.2, we present our second core contribution: the identification of misspecification bias. We derive a precise expression of the output elasticity estimate of a bundle and discuss its implications. Finally, in Section 6.3, we return to our simulated equilibrium market to demonstrate the consequences of misspecification on output-derived markups.

6.1 | Bundling in Theory: Markups and Elasticities

This section develops the theory of input bundling for markup estimation. We begin by defining input bundles and the critical substitutability assumption in Section 6.1.1. We then derive the markup formula for a bundle of substitutable inputs in Section 6.1.2. Finally, we explore the consequences of bundling non-substitutable inputs in Section 6.1.3.

6.1.1 | The Substitutability Assumption

We define an input bundle B of cardinality N as the sum of some sub-inputs X_j:

$$B_N = X_1 + X_2 + \cdots + X_N$$

This could be, for example, the 'cost of goods sold' variable which De Loecker et al. (2020) use in their analysis. When the elasticity of this bundle is estimated it is implicitly assumed that the true production function takes a form which includes this bundle. For the following exposition, we will use the example of an assumed Cobb-Douglas form with constant returns to scale, but the main result — that unobserved compositional heterogeneity in bundles introduces bias — does not rely on a particular functional form. For now, suppose we assume the true production function to be:

$$Q = AK^{1-\theta}B_N^{\theta} = AK^{1-\theta}(X_1 + X_2 + \cdots + X_N)^{\theta}$$

This is not a trivial assumption, as bundling necessitates that each sub-input be perfectly substitutable with each other. The values of each individual X_j are thus inconsequential, and only the total value of B_N has an effect on

output. For some bundles, this may be a reasonable assumption, however the bundles observed in existing datasets have been constructed with financial reporting in mind and are therefore aggregated somewhat arbitrarily from an industrial-organization perspective. To continue with the cost of goods sold example, for this to enter the production function as a distinct variable is to assume that lease expenses and labor are perfect substitutes, or non-income tax expense and input purchases, all of which and more are included within the cost of goods sold variable provided by Compustat and used by De Loecker et al.

For now, we will proceed as though this assumption is correct, and will demonstrate that there is no problem, in theory, with estimating markups from a perfectly substitutable bundle of inputs. We will then clarify the consequences of violations to the assumed substitutability of bundled inputs, before moving onto a discussion of misspecification bias, which will be problematic regardless of whether or not the substitutability assumption holds.

6.1.2 | The Estimable Form of the Markup with Bundled Inputs

The cost function for the production function with bundled sub-inputs is:

$$C = rK + \sum_{j=1}^{N} p_j X_j$$

where p_j is the price of sub-input j. Following our initial derivation of the markup, we construct the Lagrangian function for this production function as:

$$\mathcal{L} = \sum_{j=1}^{N} p_j X_j + rK + \lambda(Q - AK^{1-\theta}(X_1 + X_2 + \cdots + X_N)^{\theta})$$

The derivative with respect to X_j is:

$$\frac{\partial \mathcal{L}}{\partial X_j} = p_j - \lambda\theta AK^{1-\theta}(X_1 + X_2 + \cdots + X_N)^{\theta-1}$$

And solving for λ yields:

$$\lambda = \frac{\sum_{j=1}^{N} p_j X_j}{\theta Q}$$

Recalling that $\lambda = MC$ and $\mu = \frac{P}{MC}$, we can formulate the markup as:

$$\mu = \theta \frac{PQ}{\sum_{j=1}^{N} p_j X_j}$$

This says that the markup is equal to the elasticity of the sum of inputs, multiplied by the ratio of revenue to expenditure on the bundled inputs. Functionally, this is exactly the same as the markup derived from a single variable input. The only difference being that we now refer to the elasticity of, and expenditure on, the bundle, rather than a single input. Hence, bundling, in and of itself, is theoretically sound, so long as the bundled inputs are substitutable.

6.1.3 | Consequences of Bundling Non-Substitutable Inputs

When bundled inputs are not perfectly substitutable, problems arise. Intuitively, this is because their non-substitutability implies that they should enter the production function separately, much like capital and labor would. Each non-substitutable input will have a distinct effect on output beyond what they contribute to the magnitude of the bundle. The output elasticity of the bundle, then, depends on its composition, which we do not observe.

Consider the production function:

$$Q = AK^{1-\theta}\prod_{j=1}^{N} X_j^{\theta_j}, \qquad \sum_j \theta_j = \theta$$

This is a standard Cobb-Douglas production function where each non-substitutable sub-input enters separately. Given that B_N is the sum of the sub-inputs, we can also express each sub-input as a proportion of the bundle:

$$X_j = S_j B_N, \qquad S_j = \frac{X_j}{B_N}, \qquad \sum_j S_j = 1$$

Substituting this into our production function:

$$Q = AK^{1-\theta}\prod_{j=1}^{N}\left(S_j B_N\right)^{\theta_j} = AK^{1-\theta}B_N^{\theta}\prod_{j=1}^{N} S_j^{\theta_j}$$

Hence, the sub-inputs affect output through two channels: their sum, i.e., the value of the bundle B_N^{θ}, and their proportions relative to each other, i.e., the composition $\prod_{j=1}^{N} S_j^{\theta_j}$. In most empirical contexts, however, we only observe the overall value of the bundle. With the composition unobserved, we will not be able to properly capture the effects of individual sub-inputs on output.

There are two scenarios in which this ceases to be a problem. Firstly, if bundled sub-inputs are perfectly substitutable, the composition of the bundle is of no consequence. Dropping the composition term gives us:

$$Q = AK^{1-\theta}B_N^{\theta}$$

which is exactly the earlier specified production function for substitutable sub-inputs. Alternatively, if the proportion of the bundle made up by each sub-input is fixed, the composition term will be a constant with no implications on the estimation of θ. To show this more explicitly, we can express the derivative of output with respect to the bundle B_N as:

$$\frac{\mathrm{dln}\, Q}{\mathrm{dln}\, B_N} = \theta + \sum_{j=1}^{N} \theta_j \frac{\mathrm{dln}\, S_j}{\mathrm{dln}\, B_N}$$

Thus, if the composition of the bundle does not change, the elasticity of the bundle will be θ. If it does, however, then there will be an additional composition effect which we will be unable to account for without directly observing the individual sub-inputs.

We conclude with a numerical example to highlight the severity of mis-specifying the production function by bundling non-substitutable inputs in this way. Consider a firm with the simplified production function:

$$Q = B_2^{\theta} = (X_1^{0.4} X_2^{0.6})^{\theta}$$

B_2 is a bundle of two inputs, X_1 and X_2, with elasticities $\theta_1 = 0.2$, $\theta_2 = 0.3$, and $\theta = \theta_1 + \theta_2 = 0.5$. We observe two differently specified bundles of these inputs. Firstly, an elasticity weighted, multiplicative bundle, which represents the correct way to bundle inputs with separate elasticities:

$$B_{2,mult} = X_1^{\frac{0.2}{0.5}} X_2^{\frac{0.3}{0.5}} = X_1^{0.4} X_2^{0.6}$$

The second bundle we observe is simply the sum of the two inputs:

$$B_{2,add} = X_1 + X_2$$

We can then consider the changes in these bundles for some change in the underlying inputs. Suppose the firm starts with the bundle (250,400) before increasing both inputs by 10%.

Bundle	(250,400)	(275,440)	%change
$B_{2,mult}$	331.445	364.59	10%
$B_{2,add}$	650	715	10%

In this case, the bundle has been increased proportionally by 10%, meaning the composition of the bundle did not change. Hence, there is no difference between one bundle or the other for the purposes of elasticity estimation. This is not the only way to change the inputs, however.

Now, consider that a 10% increase to $B_{2,add}$ represents an increase of 65 physical units. Above we have distributed this change proportionally, but we can also channel the entire increase through X_2:

Bundle	(250,400)	(250,465)	%change
$B_{2,mult}$	331.445	362.78	9.45%
$B_{2,add}$	650	715	10%

Now, $B_{2,mult}$ increases by less than 10%, properly reflecting the diminishing returns of X_2, whereas $B_{2,add}$ still changes by exactly 10%. Going further, we can increase X_1 by 65 units, while simultaneously distributing 300 units of X_2 to X_1:

Bundle	(250,400)	(615,100)	%change
$B_{2,mult}$	331.445	206.8	-37.6%
$B_{2,add}$	650	715	10%

In this case the mass re-distribution has considerably decreased the effective productivity of the inputs, which is reflected in $B_{2,mult}$. On the other hand, $B_{2,add}$ has again increased by 10%. It does not differentiate between any bundle of 715 units. Consequently, we will be unable to link the expected decrease in output to any change in $B_{2,add}$, massively decreasing its power as an explanatory variable, and impairing our ability to recover the correct elasticity estimate.

6.2 | Misspecification Bias

Under the assumption that all inputs included in B_N are perfectly substitutable, its composition will not matter, and it will be possible to estimate its elasticity using only observations of the sum of the sub-inputs. There is, however, a separate and more severe issue: bundle misspecification — the mismatch between how inputs truly enter production and how we observe them in data. This section identifies the consequences of this misspecification, showing that it creates a channel of bias through which the levels of estimated markups become mechanically determined.

We begin in Section 6.2.1 by formalizing the difference between the *true* bundle that enters the production function and the bundle that we observe. Then we derive the exact form of misspecification bias in Section 6.2.2. We characterize the movement of this bias in Section 6.2.3, before revealing how it creates a mechanical relationship between bundle size and markup estimates in Section 6.2.4.

6.2.1 | The True and Observed Bundles

However inputs enter the true production function, we must observe and estimate them in the same form. If, for example, the production function is a hybrid of bundled and non-bundled inputs:

$$Q_i = A_i K_i^{\theta_k} (X_1 + X_2)^{\theta} X_3^{\theta_3} X_4^{\theta_4}$$

then, to recover θ_v, we must observe $B_{i2} = (X_1 + X_2)$ and use it as an independent variable in our productivity-controlled regression. If we instead observe only $B_{i1} = X_1$ our estimate will be biased. Likewise, if our data is aggregated such that we only observe $B_{i3} = (X_1 + X_2 + X_3)$, our estimate will also be biased. This latter case is particularly relevant given the methodologies of De Loecker et al. (2020) and Traina (2018). These papers use broad input bundles containing disparate inputs that are not perfect substitutes, and are thus unlikely to enter the true function as a bundle.

We generalize this intuition using a hybrid production function that allows both bundled and separate inputs

$$Q_i = A_i K_i^{\theta_k} B_{iN}^{\theta} \prod_{j=1}^{N} X_{ij}^{\theta_j}$$

In log form this is:

$$q_i = \omega_i + \theta_k k_i + \theta b_{iN} + \sum_{j=1}^{N} \theta_j x_{ij}$$

We will then further reduce this to:

$$q_i = \theta b_{iN} + Z\beta$$

where $Z\beta$ is a vector of controls containing the productivity control, capital, and all inputs not included in B:

$$Z = \omega_i + \theta_k k_i + \sum_{j \notin T} \theta_j x_{ij}$$

Now, let the true bundle be $B_{iN} = \sum_{j \in T} X_{ij}$, where T denotes the set of inputs that truly enter the production function as a bundle. We define the observed bundle B_{iN}^{*} as the sum of inputs in some observed set O, which may differ from T.

We can express the relationship between the observed and true bundles with a general proportion factor:

$$B_{iN}^{*} = B_{iN} * R_{iN}$$

where:

$$R_{iN} = \frac{\sum_{j \in O \cap T} X_{ij}}{\sum_{j \in T} X_{ij}} + \frac{\sum_{j \in O \setminus T} X_{ij}}{\sum_{j \in T} X_{ij}} = \frac{\sum_{j \in O} X_{ij}}{\sum_{j \in T} X_{ij}}$$

In log-form this is:

$$b^*_{iN} = b_{iN} + r_{iN}$$

The function we empirically estimate — i.e., the observed function — is thus:

$$q_i = \theta b^*_{iN} + Z^* \beta + \varepsilon_i = \theta(b_{iN} + r_{iN}) + Z^* \beta + \varepsilon_i$$

where ε_i is the error term of the empirical model, and Z^* is the vector of all controls not included within the observed bundle. Note that this entails a direct relationship between b^*_{iN} and Z^* in that all observed variables are included in one or the other. If a variable is included within Z^* then it is absent from b^*_{iN} and vice versa. The allocation of variables between b^*_{iN} and Z^* will thus move the proportion factor (R_{iN}).

This framework handles three cases of interest. First, the under-inclusion of inputs, such that our observed bundle is a subset of the true bundle ($O \subset T \; and \; R_{iN} < 1$). Second, perfect measurement, where our observed bundle is exactly equivalent to the true bundle ($O = T \; and \; R_{iN} = 1$). And third, over-inclusion, where our observed bundle contains more inputs than the original bundle ($O \supset T \; and \; R_{iN} > 1$).

The critical insight is that when $R_{iN} \neq 1$ — that is, whenever the observed bundle differs from the true bundle — estimates suffer from misspecification bias. This occurs because our observed bundle consists of both the true bundle (b_{iN}) and noise (r_{iN}). If there is any variance in how firms allocate their expenditures across a given bundle — for example, if wages make up a greater proportion of cost of goods sold for some firms and less for others, there will be variance in r_{iN}, and b^*_{iN} thus becomes a noisy approximation of the true bundle. While this resembles classical measurement error (typically causing attenuation), we will show that over-inclusion can produce upward bias, making "misspecification bias" the more accurate description.

6.2.2 | Deriving Misspecification Bias

We now explicitly derive the estimated elasticity of an observed bundle, revealing the exact form of misspecification bias. To do this, we rely on the Frisch-Waugh-Lovell theorem[12], which shows that for the model:

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_n x_{in} + \varepsilon_i$$

the estimate of β_n produced by the multivariate regression of y on all independent variables $\{x_1, \ldots, x_n\}$, will be the same as the estimate produced by a regression of $y_{i,\perp}$ on $x_{in,\perp}$, where $y_{i,\perp}$ denotes the residuals from a regression of y on all independent variables other than x_n, and $x_{in,\perp}$ denotes the same but for a regression of x_n.

[12] See Basu (2023) for an overview.

In other words, if we are only interested in the estimate β_n, we can partial out the effects of all other independent variables on y and x_n, then estimate:

$$y_{i,\perp} = \beta_n x_{in,\perp} + \varepsilon_i$$

and this will be equivalent to the full model. The ordinary-least-squares estimate of this linear model will be:

$$\hat{\beta}_n = \frac{\sum(x_{in,\perp} - \bar{x}_{n,\perp})(y_{i,\perp} - \bar{y}_{\perp})}{\sum\left(x_{in,\perp} - \bar{x}_{n,\perp}\right)^2}$$

But unless we deliberately omit the constant from our regressions, the residuals $x_{in,\perp}$ and $y_{i,\perp}$ will be constructed such that their expected value is zero, hence:

$$\hat{\beta}_n = \frac{\sum(x_{in,\perp} * y_{i,\perp})}{\sum\left(x_{in,\perp}\right)^2}$$

We can therefore express the estimated elasticity of an observed bundle as:

$$plim\ \hat{\theta} = \frac{E[b^*_{iN,\perp} * q_{i,\perp}]}{E[(b^*_{iN,\perp})^2]}$$

where $\perp$ denotes the residuals after regressing on the controls Z^*. Then given:

$$b^*_{iN,\perp} = b_{iN,\perp} + r_{iN,\perp}$$

and, from the true model:

$$q_{i,\perp} = Z_\perp \beta + \theta b_{iN,\perp} + \varepsilon_{i,\perp} = \theta b_{iN,\perp} + \varepsilon_{i,\perp}$$

we can rewrite the estimator as:

$$plim\ \hat{\theta} = \frac{E[(b_{iN,\perp} + r_{iN,\perp}) * (\theta b_{iN,\perp} + \varepsilon_{i,\perp})]}{E[(b_{iN,\perp} + r_{iN,\perp})^2]}$$

which expands to:

$$plim\ \hat{\theta} = \frac{\theta E[b^2_{iN,\perp}] + E[b_{iN,\perp}\varepsilon_{i,\perp}] + \theta E[b_{iN,\perp} r_{iN,\perp}] + E[r_{iN,\perp}\varepsilon_{i,\perp}]}{E\left[b^2_{iN,\perp}\right] + E\left[r^2_{iN,\perp}\right] + 2E[b_{iN,\perp} r_{iN,\perp}]}$$

We assume exogeneity, i.e., that the error term ε_i is uncorrelated with b, r, or Z. It follows that $E\left[b_{iN,\perp}\varepsilon_{i,\perp}\right] = 0$ and $E\left[r_{iN,\perp}\varepsilon_{i,\perp}\right] = 0$. Additionally, since we have residualized each variable, and the expected value of a residual is 0 by construction, we can express each expectation as a variance or covariance:

$$plim\ \hat{\theta} = \theta\ \frac{var(b_{iN,\perp}) + cov(b_{iN,\perp} r_{iN,\perp})}{var(b_{iN,\perp}) + var(r_{iN,\perp}) + 2cov(b_{iN,\perp} r_{iN,\perp})}$$

This expression states that the estimated coefficient of the partial bundle will be equal to the true elasticity of the bundle, multiplied by a ratio of the variances and covariances of $b_{iN,\perp}$ and $r_{iN,\perp}$ — this is the bias.

6.2.3 | The Direction of Misspecification Bias

This expression reveals that $\hat{\theta}$ depends on the relationship between the true bundle and the observed bundle through the variances and covariances of $b_{iN,\perp}$ and $r_{iN,\perp}$. To understand the practical implications, we examine how this bias evolves as we systematically expand the observed bundle.

Case 1: Under Inclusion ($O \subset T$)

When the observed bundle is a subset of the true bundle, $cov(b_{iN,\perp} r_{i,N,\perp}) > 0$. The residualized r represents firm-specific variation in the proportion of the true bundle that is observed. When a firm uses more of the observed inputs relative to firms with similar production characteristics, we would also expect higher usage of the true bundle. This positive covariance makes the bias ratio less than 1, resulting in $\hat{\theta} < \theta$ — classical attenuation bias.

Case 2: Perfect Measurement ($O = T$)

When the observed bundle exactly matches the true bundle, $r_{iN,\perp} = 0$, $var(r_{iN,\perp}) = 0$, and $cov(b_{iN,\perp} r_{iN,\perp}) = 0$, thus collapsing the bias ratio to 1. We thus recover the true elasticity: $\hat{\theta} = \theta$.

Case 3: Over-inclusion ($O \supset T$)

When the observed bundle contains extraneous inputs not in the true bundle, $cov(b_{iN,\perp} r_{i,N,\perp}) < 0$. Here, r reflects the proportion of extraneous inputs relative to the true bundle. Firms with higher true bundle usage typically have a smaller proportion of extraneous inputs, creating negative covariance.[13] This can make the bias ratio exceed 1, resulting in $\hat{\theta} > \theta$.

In both under-inclusion and over-inclusion, adding inputs to the observed bundle systematically affects $\hat{\theta}$. Moving from under-inclusion toward perfect measurement reduces attenuation, increasing $\hat{\theta}$. Continuing into over-inclusion can further increase $\hat{\theta}$ due to negative covariance effects.

6.2.4 | The Mechanical Channel: From Bias to Determinism

The existence of misspecification bias is concerning, but the core problem runs deeper: it creates a *mechanical relationship* between bundle size and markup estimates that operates independently of true markups.

---

[13] This follows directly from definition of R in Section 5.2.1. For example, if we have the production function: $Pants = Dye * Cloth$ — that is, dye and cloth separately enter the production function to create pants — but we observe the bundle $B^* = Dye + Cloth$, we will have $R = \frac{dye+cloth}{dye} = 1 + \frac{dye}{cloth}$. Given that the true bundle is simply dye ($B = Dye$), the covariance term will thus approximate: $cov(dye, 1 + \frac{cloth}{dye})$ which should naturally be negative.

This occurs through the combination of two channels that evolve predictably with bundle size. The first is the estimated elasticity. As shown above, $\hat{\theta}$ changes systematically with bundle composition. When moving from under-inclusion to perfect measurement, $\hat{\theta}$ increases due to the reduction of attenuation bias. Continuing into over-inclusion typically maintains or further increases $\hat{\theta}$ due to negative covariances.

At the same time, expenditure on the observed bundle ($\sum_{j\epsilon O} p_j X_j$) grows linearly with bundle size. Each input adds to total expenditure regardless of its productive relevance, thereby depressing the revenue share $\left(\frac{PQ}{\sum_{j\epsilon O} p_j X_j}\right)$.

The markup formula then combines these two channels:

$$\hat{\mu} = \hat{\theta} \frac{PQ}{\sum_{j\epsilon O} p_j X_j}$$

Thus, when we add inputs to the observed bundle, there are only two possibilities. The growth of $\hat{\theta}$ will outpace the growth of expenditure, causing the estimated markup to increase, or expenditure will dominate $\hat{\theta}$, pushing the markup down.

In general, we will find the latter to be true, as while $\hat{\theta}$ increases with bundle size, this growth typically occurs at a *diminishing rate* due to the non-linear nature of the bias ratio. Meanwhile, expenditure grows *linearly*. The net effect is that markup estimates ($\hat{\mu}$) systematically decrease as inputs are added to the observed bundle.

Regardless, in either case this mechanism guarantees that $\hat{\mu}$ becomes a direct function of bundle size rather than an estimate of true market power. The methodological choice of which inputs to include mechanically determines the resulting markup estimate.

6.3 | Simulated Exploration of Misspecification Bias

We now return to the simulated equilibrium market to confirm the mechanical relationship between bundle size and markup estimates predicted by our theoretical framework. Using the simulation, we are able to show that both under-inclusion and over-inclusion of inputs creates a systematic bias such that estimated markups are reduced to arbitrary values determined by the observed bundle, rather than by market power.

6.3.1 | The Model

The simulated market is set up as before. We generate 15,000 firm observations, all with the same production function, and with assigned values of capital, productivity, etc. The primary change is to our true production function, which is now a hybrid that includes a bundle of 15 perfectly substitutable inputs, as well as 30 non-substitutable inputs which enter separately:

$$Q_{it} = A_{it}K_{it}^{0.125}B_{it15}^{0.75}\prod_{j=16}^{45}X_{ij}^{\frac{0.125}{30}}E_{it} = A_{it}K_{it}^{0.125}(X_{it1} + X_{it2} + \cdots + X_{it15})_{it}^{0.75}\prod_{j=16}^{45}X_{ij}^{\frac{0.125}{30}}E_{it}$$

As in the prior simulation, we generate values for the non-variable inputs — which now include inputs X_{16}- X_{45} — then use the production function to solve for each firm's equilibrium output and input usage. In this case, the relevant variable input is B_{it15}, i.e., the sum of inputs X_1-X_{15}. After solving for the aggregate bundle, we then define each individual input in the bundle as a share (S_{itj}) of the full bundle (B_{it15}):

$$for\ j = 1, \ldots, 15, \qquad X_{itj} = S_{itj}B_{it15},\ \ S_{itj} = \frac{X_{itj}}{B_{it15}}\ \ , \sum_{j=1}^{15} S_{itj} = 1$$

In practice, we accomplish this by generating a share variable — representing S_{itj} — for each input, according to a specified distribution. The share variables are then normalized such that they sum to 1 for each firm. The result is that B_{it15} varies both in magnitude — from the cost-minimization process — and composition — from the distribution of shares.

We test four different distributions — uniform, Dirichlet, censored normal, and log-normal — which are used for the generation of both input shares S_{itj}, and the non-bundled inputs (X_{16}- X_{45}). The choice of distribution matters because it determines the variance-covariance patterns that drive how elasticity estimates evolve with bundle size. With homogeneous input shares across firms, estimates follow a smooth trajectory. But when firms differ substantially in their input allocations, the path becomes more extreme.

Figure 3 visualizes how the firm-specific share of a given input (S_{itj}) is distributed across firms. The shares of the first bundled input (X_1) is used as an example, but all bundled inputs are mean $\frac{1}{15}$ and are thus roughly equivalent. Similarly, the non-bundled inputs merely shift the mean to 0.5, otherwise following the same distribution.

Figure 3 — Distributions of S_{it1} (True Bundle Shares of Input X_1)

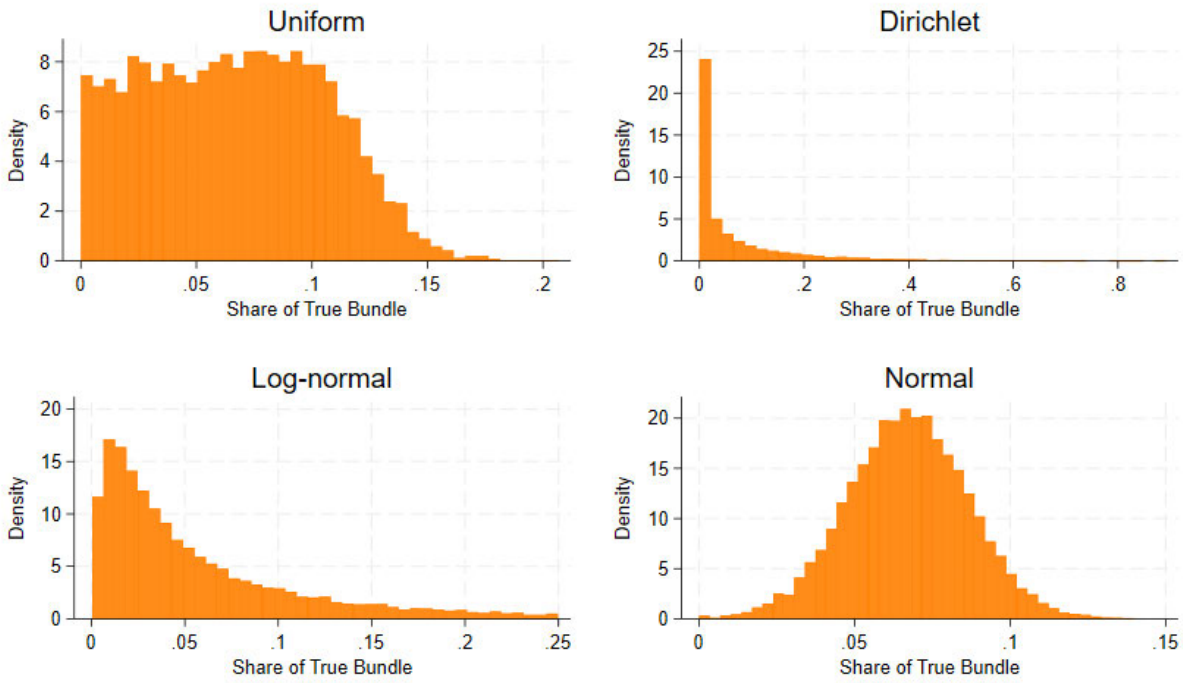

Figure 3: Distributions of the share of input X_1 in the total bundle.

The uniform distribution generates input shares that are equally likely across a fixed range, producing highly variable bundle compositions where no single input tends to dominate. This pattern emerges when inputs are readily substitutable and firms face few constraints in their input mix, such as with generic raw materials.

The Dirichlet distribution produces extreme heterogeneity, with most firms allocating negligible shares to a given input while a few concentrate heavily. This reflects specialized inputs used by few firms.

The normal and log-normal distributions generate concentrated shares clustered around the mean, resulting in homogeneous bundle compositions across firms. This pattern characterizes standardized inputs that serve common functions across all production processes — i.e., electricity, labor, etc., that firms use in consistent proportions.

6.3.2 | Simulated Markup Estimates: Under and Over-Inclusion

Having solved for each firm's equilibrium output (Q_{it}) and variable input use (B_{it15}), and with each bundled and non-bundled X_{itj} generated, we are able to estimate markups and elasticities. Note that these are all output-derived, and thus the demonstrated biases are independent of revenue bias.

We use the 45 substitutable and non-substitutable inputs to construct 45 different bundles for which we estimate the elasticity and calculate the corresponding markup. Each bundle is the sum of the first N inputs, i.e.:

$$B^*_{itN} = \sum_{j=1}^{N} X_{itj}$$

The first bundle, for example, is $B^*_{it1} = X_{it1}$, while the last is $B^*_{it45} = X_{it1} + \cdots + X_{it45}$. All inputs not contained in the observed bundle implemented as individual controls, preventing omitted variable bias. This design allows us to examine the three cases described in Section 6.2.3. Bundles 1-14 represent under-inclusion, where our observed bundle is a subset of the substitutable inputs contained in the true bundle. Bundle 15 is perfect measurement, and should recover the true markup. Finally, bundles 16-45 are over-inclusive, containing extraneous inputs that are not present in the true bundle.

The over-inclusion case is of particular relevance to empirical work as perfectly substitutable inputs are theoretical rarities in practice. Even when such inputs exist, data of reasonable scope will not be aggregated at a fine enough level to observe them separately. Consequently, virtually every input variable in empirical work represents an over-included bundle. De Loecker et al. (2020)'s 'cost of goods sold', for example, is an accounting aggregate combining wages, materials, depreciation, and other economically distinct elements. Even in more ideal cases where data is less aggregated, seemingly fine-grain categories such as 'intermediate inputs' inevitably bundle non-substitutable components. Thus, a proper understanding of over-inclusion bias is critical — both for interpreting existing estimates and for evaluating the very feasibility of the production approach to markups.

Figure 4 presents our elasticity estimates for the 45 bundles, showing precisely how bundle construction determines the estimate we obtain. Note that for this exercise, we have set the true output elasticity of the bundle to 0.75, and the true markup to 1.2.

Figure 4 — Output Elasticity Estimates of Bundles 1-45

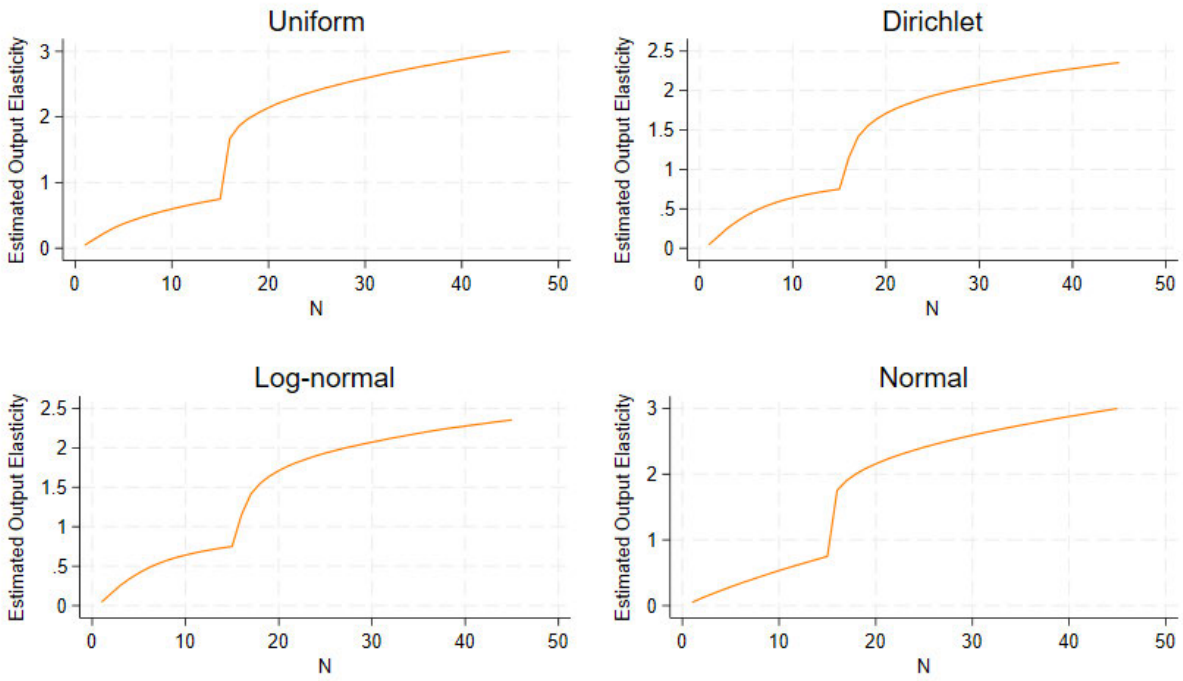

Figure 4: Output-derived elasticity estimates of bundles containing the first N inputs.

The estimates exhibit three characteristics of note. Firstly, in the under-inclusion case, the reduction of attenuation bias given by the addition of inputs is generally diminishing, though this depends on the distribution. Secondly, the correct output elasticity is recovered at bundle 15. Thirdly, and most noticeably, the estimated elasticity spikes exaggeratedly at bundle 16, with the extremity of the spike depending on the distribution.

This latter observation coincides exactly with the movement predicted by our derived form of the estimate. At bundle 16, we introduce our first extraneous — i.e., non-substitutable — input. This causes the residual covariance between the true bundle and the proportion factor ($cov(b_{iN,\perp} r_{iN,\perp})$) to instantly shift from 0 — as $r_{iN,\perp}$ is a constant in the perfect measurement case — to negative.

Recall that the proportion factor is:

$$R_{itN} = \frac{\sum_{j \in O} X_{itj}}{\sum_{j \in T} X_{itj}}$$

In the case of over inclusion, this represents the use of extraneous inputs relative to the inputs in the true bundle. In our simulation, this has a mechanical relationship with the variable bundle. Because we treat the extraneous inputs as fixed — generating them prior to cost-minimization — firms possessing higher levels of the extraneous inputs will also use less variable input. This is in the same way that cost-minimizing firms with high capital will

use less input, all else equal. This accounts for the extremity of the jump. In empirical contexts, the transition will likely be smoother.

Regardless, both before and after the jump, the elasticity estimates increase as inputs are added to the observed bundle. Expenditure on the observed bundle also necessarily increases as inputs are added, and one will outpace the other, creating the channel through which bundle size determines markups.

In the simulation, the direction of the channel depends on the price we define for the extraneous inputs. In general, we discard the possibility of the elasticity outpacing cost — thereby inflating the markup — as this only occurs when we define the price as being unreasonably low.[14] Going forward, we thus assume that the markup is pushed down as the observed bundle grows — an assumption we later validate empirically.

Figures 5 and 6 present the estimated markups. Figure 5 shows the full evolution of the markup, while Figure 6 focuses on bundles 15-45, representing the transition from perfect measurement to the empirically inevitable over-inclusion.

Figure 5 — Average Estimated Markup

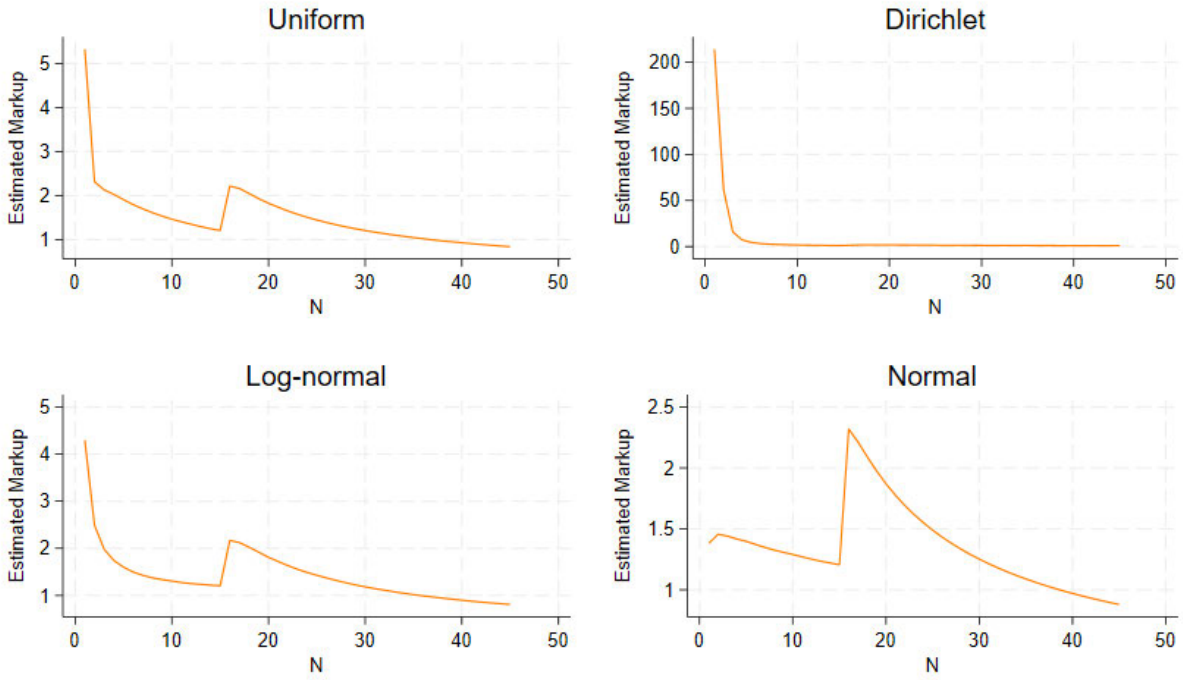

Figure 5: Output-derived markup estimates of bundles containing the first N inputs.

[14] Elasticity growth just barely ekes out expenditure when we set up the model such that the average cost of an extraneous input is 15x lower than the average cost of a substitutable input.

Figure 6 — Average Estimated Markup (bundles 15-45)

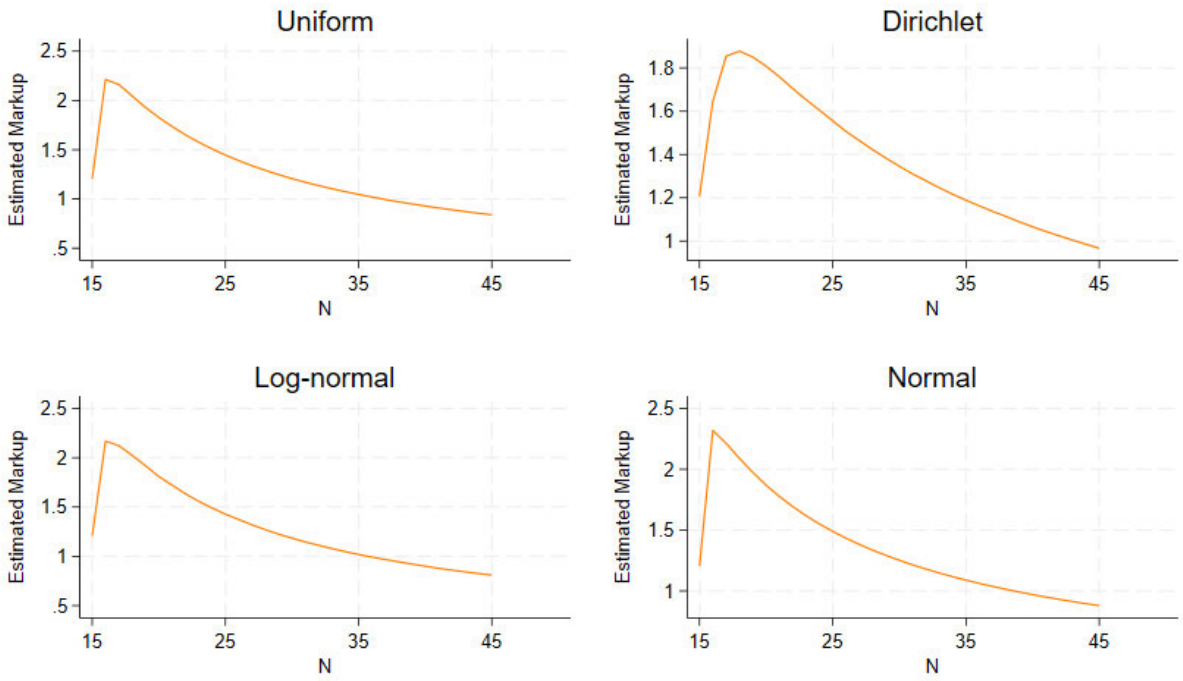

Figure 6: Output-derived markup estimates of bundles containing the first N inputs. Estimates for bundles 15-45 are displayed.

Figure 5 shows that as attenuation bias is reduced, the markup estimate converges to the true markup (1.2), with perfect recovery of the markup occurring at bundle 15. We then observe the spike again, which is exaggerated by the strictness of the simulated covariance structure. The more critical insight for empirical work, is that further additions to the observed bundle *mechanically depress the markup*. In Figure 6, this depression results in a final markup estimate of 0.88 — significantly below the true markup — however this is only limited by the number of added inputs and their costs.

If we, for example, double the cost of the extraneous inputs, the final markup estimate drops to 0.5, as shown in Figure 7.

Figure 7 — Average Estimated Markup with Doubled Extraneous Input Cost

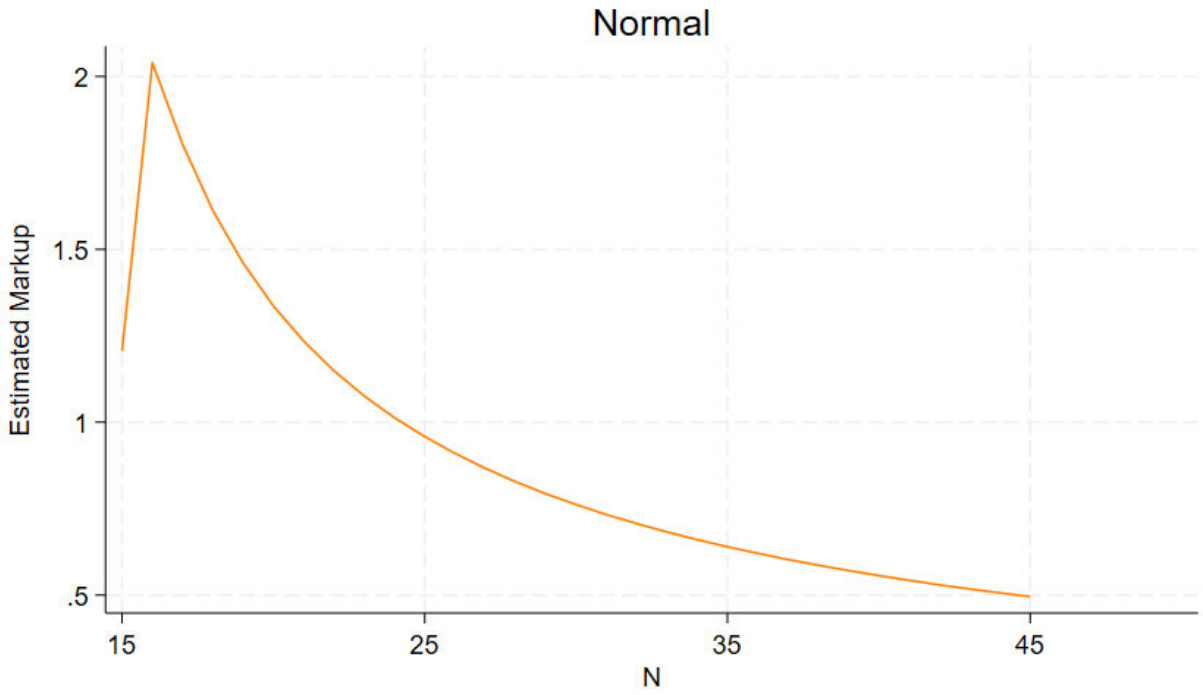

Figure 7: Output-derived markup estimates of bundles containing the first N inputs when the cost of extraneous inputs (X_{15}-X_{45}) is doubled. Estimates for bundles 15-45 are displayed.

As this demonstrates, the downward bias we can introduce into the markup estimate by bundling non-substitutable inputs is only constrained by the total expenditure of firms. Furthermore, once we have included even a single extraneous input, the true markup becomes unrecoverable. This is demonstrated in Figure 8, which examines mixed bundles where substitutable and extraneous inputs are added in alternating sequence.[15] Each additional extraneous input permanently reduces the maximum achievable markup estimate. When attenuation bias from under-inclusion combines with the bias from over-inclusion, the resulting estimates fall substantially below those from the clean sequential path. Nevertheless, despite these different trajectories, all bundle specifications ultimately converge to the same final markup estimate of 0.88.

[15] More, specifically, we alternate, adding two extraneous inputs before a substitutable input.

Figure 8 — Average Estimated Markup with Alternating Inputs

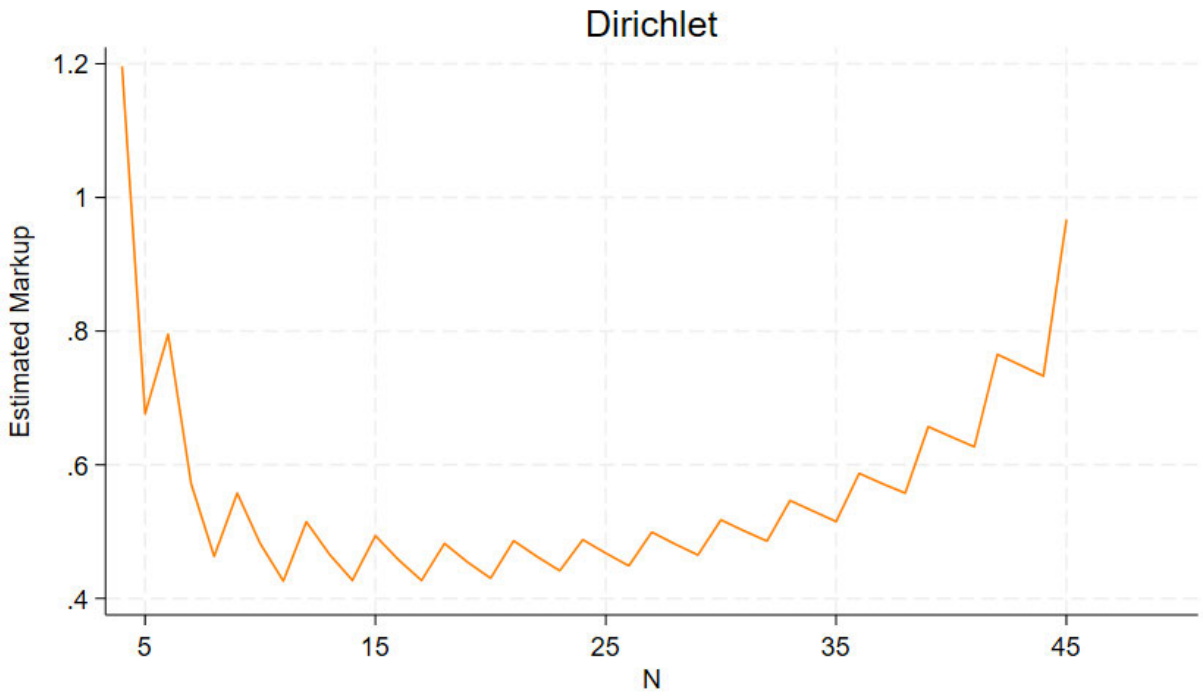

Figure 8: Output-derived markup estimates of bundles containing the first N inputs. Extraneous and substitutable inputs are added to the observed bundle in alternating order.

6.3.3 | Implications of Misspecification Bias

The implication of this finding for empirical work is stark. The production approach to markup estimation is undermined not only by revenue data but also by a fundamental, mechanical flaw in the near-unavoidable misspecification of production functions. The level of the estimated markup is not an estimate of market power, but an arbitrary value that we effectively *choose* via the choice of variable input.

This perfectly answers the second of the quandaries posed in Section 3.2 regarding the persistent gap between the markup estimates of Traina (2018) and De Loecker et al. (2020). Traina attempts to correct De Loecker et al.'s estimated markup series by broadening the scope of the defined bundle to include sales, general, and administrative expenses. The result is that Traina's markups sit strictly below De Loecker et al.'s for the entire sample spanning 1950-2016. This is not a correction reflecting improved methodological validity, but the predictable outcome of a deterministic process: broader bundles mechanically produce lower markup estimates, regardless of the underlying market.

When combined with the revenue bias explored in Section 5, the conclusion is inescapable: the standard implementation of the production approach, reliant on revenue data and arbitrarily defined, over-included input bundles, cannot reliably recover the level of true markups. We will confirm this with our empirical evidence in Section 8, before attempting to determine what estimated trends actually recover.

7 | Data

The previous two sections have illustrated the biases and potential pitfalls of inferring market power from revenue-derived markup estimates. Ultimately, recovery of the markup is unlikely in the absence of output data. Even with output data, our observed bundle must be perfectly composed of substitutable inputs, or the markup level will be biased by misspecification. Similarly, recovery of markup trends relies heavily on factors we cannot observe or control playing in our favor. The true output elasticity must either be constant, and we know this and forgo estimation, or the direction of the nonlinear bias on our elasticity estimates happens to coincide with the markup trend, thereby granting us a lucky approximation.

To validate these findings, we now move onto empirical work, where we estimate New Zealand (NZ) markups from 2003-2022. These estimates will serve both as a comparison to existing markup estimates and NZ market power metrics, and as a verification of our hypotheses on misspecification and revenue bias.

We use data from the Longitudinal Business Database (LBD), provided by Stats NZ. The database includes firm level data on New Zealand firms in 39 industries, covering the period from 2001-2022[16] with over 4 million observations. The data in the LBD is compiled from a comprehensive variety of sources, and includes information from surveys, GST, tax filings, and employment data, among others. Notably, all 'economically significant[17]' firms in New Zealand are included in the dataset. The criteria for economic significance are quite low, and so the data paints a full picture of the economy, not being biased towards large and publicly traded enterprises.

Within the LBD itself there is also a variety of user-created datasets which transform the raw data of the LBD into more useful forms for research. One such dataset, which we rely on heavily, is the 2023 iteration of 'pent_prod_IDI', created by Richard Fabling and David C. Maré (2015). This dataset was created for the purpose of estimating production functions, and contains many of the necessary variables (Y, L, K, etc.) constructed from the raw LBD data. We summarize these, along with our defined M and bundle of expenditure — analogous to the 'cost of goods sold' used by De Loecker et al. (2020) — in Table 4. The methods used to construct these variables

[16] While our sample begins in 2001, we only estimate markups from 2003 onwards. This is because our estimation strategy requires both an investment proxy — the difference in capital over two periods — as well as several lagged variables, including the lag of investment. The first period with observable investment is 2002, and its lag only begins in 2003.

[17] A firm is deemed economically significant when it meets any of a number of conditions, including having annual sales or GST expense greater than $30,000, more than three paid employees, or a new GST registration, among others.

go beyond what would be obvious, so we shall draw on the explanations given in Fabling & Maré (2015) and Fabling & Maré (2019), to detail them below.

Table 4 — Dataset Summary Statistics

Variable	Mean	Std. Dev.	Median	5th Pct.	95th Pct.	Observations
Gross Output (Y)	717450.1	5,450,526	165,040	15,337	2,328,010	3,163,983
Intermediate Input Consumption (M)	612180.5	7,699,176	51,499	1,123	1,871,362	2,874,969
Capital Services (K)	89760.3	676,343	22,781	1,770	313,780	3,163,983
Full Time Equivalent Employment (L)	2.8	15.4	0.1	0.0	10.7	3,163,983
Total Variable Expenses	960814.3	8,765,778	178,717	15,300	3,020,111	3,163,983

Table 4: Summary statistics for the dataset used in estimating our markup series. 'Total Variable Expenses' refers to the combined bundle of intermediate input consumption and all other IR10 categories of expense: bad debts, depreciation and amortization, insurance, interest expense, professional and consulting fees, rates, rental and lease payments, repairs and maintenance, research and development, salaries and wages, contractor and subcontractor payments, and 'other' expenses.

Starting with our output variable, we have that output is equal to the sum of total income and change in stocks:

$$Y_{it} = Gross\ output = sales + other\ income + stock\ change$$

This is, unsurprisingly, a measure of income, rather than quantity produced, and so we will have to deal with the price bias issues discussed previously.

More interesting are our definitions of capital and investment. Typically, K_{it} represents the level of capital. Investment, which is meant to capture a firm's decision to expand output via investments in productive capability, is therefore defined as:

$$I_t = K_t - K_{t-1} + \delta K_{t-1}$$

meaning that investment is equal to the difference in capital stock from year to year, plus depreciation. For an intuitive grasp of depreciation having a positive coefficient, consider that simply maintaining a level of capital stock requires an investment in repairing/repurchasing depreciated capital.

Because the LBD covers small, non-public firms, we do not have the luxury of taking publicly reported assets as a measure of capital stock. Instead, Fabling & Maré (2015) take a measure of the annual flow of capital services, defined as the sum of depreciation, rental and leasing costs, and borrowing costs:

$$K_{it} = Value\ of\ capital\ services = depreciation + rental\ and\ leasing\ costs + borrowing\ costs$$

This is a nonstandard measure of capital and is in fact close to the typical measure of investment — capturing capital flows directly — than a measure of capital stock. It would be natural to take this as our investment variable, but doing so would leave us without a measure of capital. We opt to keep capital services as our K_{it}, and define investment as:

$$I_t = K_t - K_{t-1}$$

This definition of investment is not lacking in flaws; hence we primarily depend on the Levinsohn and Petrin (2003) method of estimation, which does not require an investment proxy.

Labor is constructed from employee level data and comes in the form of the monthly average number of full-time equivalent (FTE) employees at the firm:

$$L_{it} = Average\ monthly\ FTE\ employment$$

The LBD provides access to IR10 filing data, which includes reported expenditure on a number of items encompassing most variable expenses. We combine[18] this expense data with 'pent_prod_IDI' to construct our measure of intermediate input, as well as various bundles which will be used in testing the effect of misspecification bias in an empirical context. We define intermediate usage as:

$$M_{it} = Purchases - goods\ for\ resale$$

This approximates the purchases which are used by a business, rather than resold. It is worth noting that the IDI productivity dataset contains a different measure of intermediate consumption which we avoid using. This is defined as the sum of purchases and total expenses, excluding employment, depreciation, interest, debt write-offs, road-user charges, goods for resale, and rental and leasing rates[19]:

$$\begin{aligned} M_{it} &= Intermediate\ consumption \\ &= Purchases + total\ expenses - salaries - wages - bad\ debts - depreciation - interest \\ &- road - user\ charges - goods\ for\ resale - rental\ and\ leasing\ rates \end{aligned}$$

We opt out of using this definition primarily because it is a pre-constructed bundle. Much like the cost of goods sold measure used by De Loecker et al. (2020), this definition of M_{it} aggregates categories of expense arbitrarily, being made up in large part by depreciation, repair costs, fees, research and development, miscellaneous expenses categorized as 'other', and remuneration of executive and associated persons. The latter of these is the most egregious, making up over 30% of total expenses on average, and having no connection whatsoever to the process of physical production; these are exactly the additions to an observed bundle which are most likely to magnify the effect of misspecification bias.

[18] We merge the datasets on year and firm identifier. Enterprise number is the identifier in the IR10 data, while 'pent' (permanent enterprise number) is the equivalent in 'pent_prod_IDI'.

[19] There are also adjustments made to account for misallocated costs (research and development (R&D) labor costs being put into R&D expense rather than salaries & wages, for example), as well as some inconsistent handling of gains/losses on asset sales. See Fabling & Maré (2019) for more details.

It is worth reiterating that, like our output measure, this is not a measure of physical input but of expenditure. It is possible to create 'deflated' measures of input and output, which are meant to approximate the physical units, but these essentially rely on dividing expenditure or revenue by a common price. This would imply that all firms in an industry are homogenous in the prices they set and receive, so this is really an assumption of minimal market power. Such an assumption defeats the point of estimating markups to begin with, so we do not use deflated measures.

Finally, we note that a small proportion of firms in the dataset have ratios of $\frac{Y_{it}}{M_{it}}$ that are extremely high, even exceeding 10,000. These observations are either the result of measurement error[20], or represent extreme examples of high revenue firms in low-input industries. If estimated elasticities are in normal ranges — i.e., greater than 0.1 — the estimated markups for these outliers will inevitably be extremely high. And when they are included in the analysis, the average estimate markup is pushed to implausible levels, ranging from 40-50 depending on estimation method. In our primary analysis, we therefore drop observations where the ratio exceeds 1000, this represents roughly 12,000 observations of the initial sample.

In doing this, we are perhaps dropping from the sample some highly successful insurance, finance, etc., firms which do not rely on purchased inputs; however, this does not do much harm to the veracity of our estimates. If these firms do not use inputs in the traditional sense, then the estimated elasticity of intermediate inputs does not represent anything relevant to the calculation of their markups. With their production functions diverging dramatically from those of the typical firm, they are best left out of the sample.

8 | Empirical Results

This section presents our empirical results. We construct a markup series with data from the Longitudinal Business Database (LBD), using our empirical estimates to validate our prior theoretical and simulated explorations of revenue distortions and misspecification bias. Section 8.1 directly confirms the mechanical channel created by misspecification bias, showing that empirical markup estimates decrease with the size of the defined input bundle. Then, in Section 8.2, we present our markup series, evidencing the validity of our expression of the revenue-derived estimate, and showing that estimated markups co-move with economic cycles rather than capturing changes in competition.

[20] This is not at all unlikely. Many firms in the sample report negative expenses, individual expenses which do not sum to their reported total, all expenditure in the 'other' category, and other such incongruencies. When firms in machinery and input heavy industries report revenue in the tens of millions and purchases in the thousands, measurement error is thus a reasonable explanation.

8.1 | Misspecification Bias in Empirical Markups

We now provide direct empirical evidence for the mechanical relationship between bundle size and markup estimates predicted by our theoretical framework. Following our simulation design, we construct input bundles of varying size using detailed expense categories from the LBD, testing whether real-world data exhibits the same deterministic patterns.

We make use of twelve distinct expense categories from the IR10 data: bad debts, depreciation and amortization, insurance, interest expense, professional and consulting fees, rates, rental and lease payments, repairs and maintenance, research and development, salaries and wages, contractor and subcontractor payments, and 'other' expenses. Purchases — our measure of intermediate input — is then treated as an additional category, creating a comprehensive set of thirteen bundle components.

Table 5 displays summary statistics for these expenditure categories, providing context for their relative importance in firms' cost structures.

Table 5 — Summary of Expenditure Categories[21]

Added to	Description	Mean	Median	5th Pct.	95th Pct.	Observations
Bundle 1	Intermediate Inputs	618675	51514	1124	1865787	2,745,954
Bundle 2	Bad Debts	937	0	0	746	2,745,954
Bundle 3	Depreciation	23881	5493	0	71552	2,745,954
Bundle 4	Insurance	6309	2017	0	21617	2,745,954
Bundle 5	Interest	16965	1003	0	63795	2,745,954
Bundle 6	Rates	2703	0	0	11620	2,745,954
Bundle 7	Rent	33528	3733	0	127257	2,745,954
Bundle 8	Repair	11819	1300	0	41351	2,745,954
Bundle 9	Research and development	646	0	0	0	2,745,954
Bundle 10	Subcontractor payment	27451	0	0	67881	2,745,954
Bundle 11	Misc.	109250	23860	2892	370058	2,745,954
Bundle 12	Other	36257	6431	0	152880	2,745,954

[21] Observation counts for this sample are slightly lower than the count for intermediate inputs in table 4. This is because we drop all observations where there is discrepancy between the sum of each expenditure category (i.e., total expenses) and the reported total expenses.

Bundle 13	Salaries and wages	145609	19885	0	543161	2,745,954

Table 5: Summary statistics for IR10 expenditure categories. "Added to" denotes the ordered bundle they first appear in. "Misc." refers to specific subcategories of expenditure that are present in some years but not others. For example, the pre-2012 IR10 form requires reporting of vehicular expenses but not associated persons remuneration. This is inverted in the post-2012 form. The removal of this category does not change the relationship between markups and bundle size.

We construct thirteen progressively larger bundles. The first bundle contains only intermediate inputs, with each subsequent bundle adding one expense category. The final bundle represents total expenditure across all categories, analogous to the broad aggregate used by Traina (2018) in deriving their estimated markup that hovers consistently close to 1.1. Given that our first input — intermediate inputs — is itself an arbitrary aggregation of various purchases, all our constructed bundles will be over-included. Consequently, we expect that the estimated markup will progressively decrease with bundle size.

As with our simulation, the ordering of category addition affects the specific path taken between the estimates on the first and last bundle, but not the overall relationship. When we randomize category sequences, we consistently observe the same downward trend in markup estimates, confirming that bundle size — not composition — drives the results. Furthermore, the composition of the first bundle is not entirely consequential, as the available disaggregated categories that may satisfy perfect substitutability — e.g., bad debts, interest expense — simultaneously fail to be freely variable — preventing cost-minimization — and are not relevant to the majority of firms. When we take these inputs as the first bundle, we thus find absurdly high markup estimates, akin to the simulated estimates on the first few bundles of the Dirichlet specification in Figure 7.

Figure 9 plots the estimated elasticities of each bundle, which are obtained from the specification outlined in appendix A2 — i.e., a regression of revenue on the bundle in addition to capital, labor, and the productivity control. The relationship between bundle size and estimate is not obvious from this depiction, given the fluctuations of the slope, but this is simply the effect of ordering. Bundles 11, 12, and 13 represent the additions of salaries and wages, contractor payments, and 'other' expenses respectively. These are some of the largest expenses, and also the most impactful in terms of production — consider the effect of interest expense on production as opposed to contractor payments — hence, it is inevitable that their impact on the estimated elasticity is large. In other permutations of the ordering, where they appear earlier in the process, their effect on the estimated elasticity is larger.

Figure 9 — Estimated Elasticities of Bundles

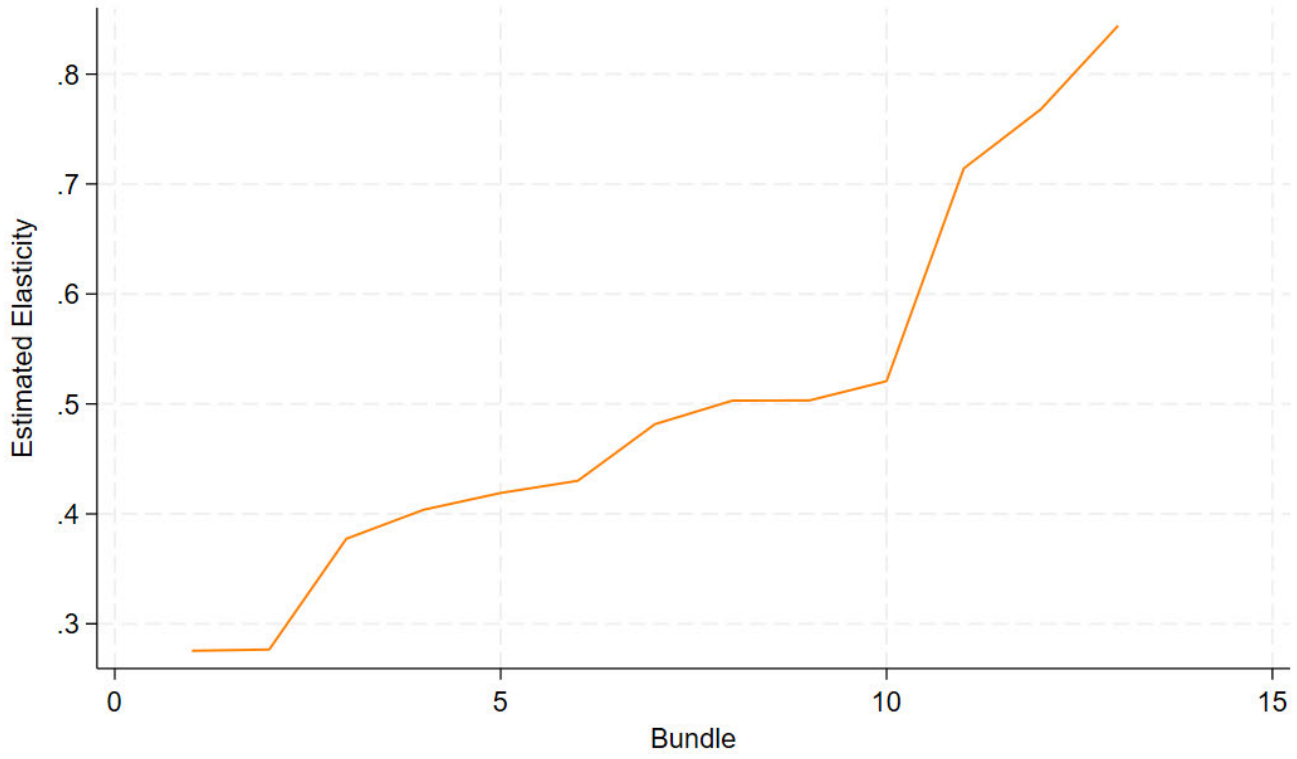

Figure 9: Empirical elasticity estimates of bundles containing the first N inputs (denoted by "Bundle"). Bundle 1 contains only our measure of intermediate input, while bundle 14 contains all variable expenditure.

Figure 10 provides the crucial evidence: estimated markups systematically decrease as bundle size increases. The empirical data perfectly replicates the mechanical relationship predicted by our theoretical framework and demonstrated in our simulations. This arises regardless of ordering, is not sensitive to the productivity control method, and, as we explore in Section 8.2.4, holds across every estimated industry.

Figure 10 — Estimated Markups of Bundles

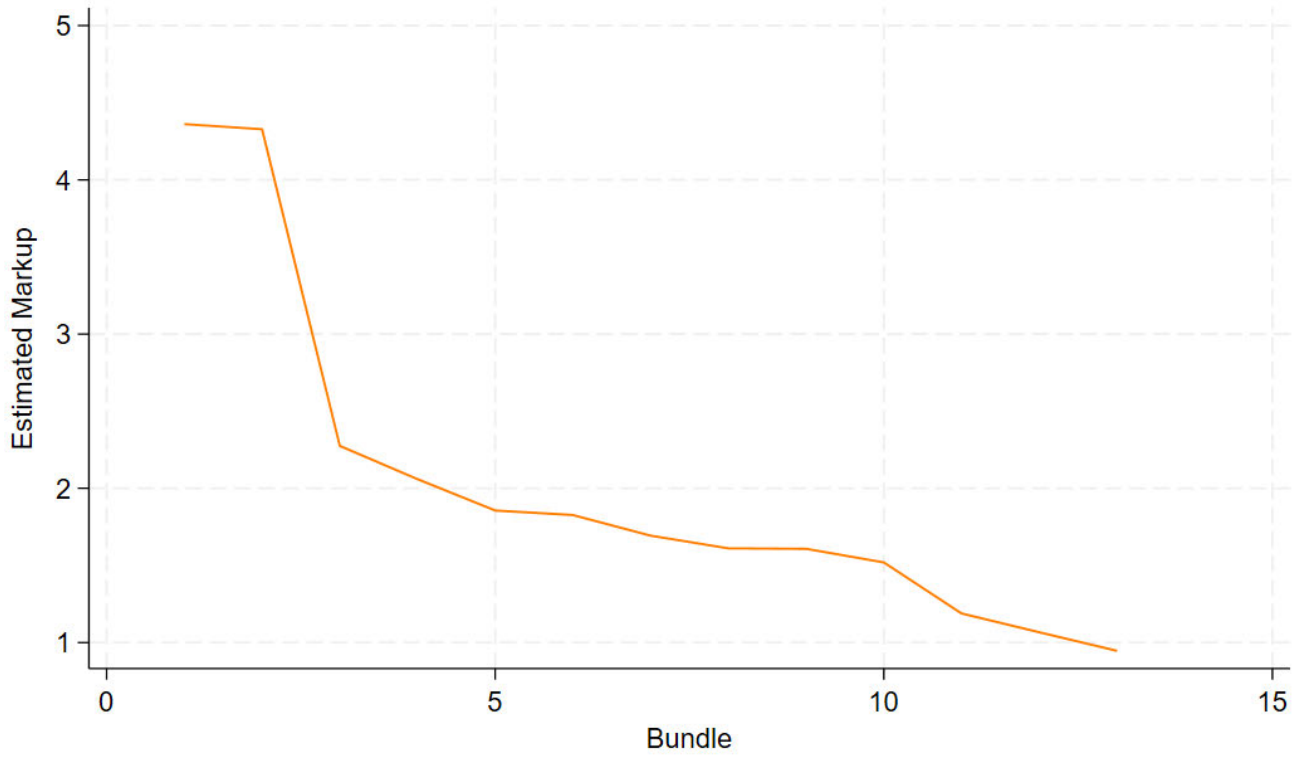

Figure 10: Empirical markup estimates of bundles containing the first N inputs (denoted by "Bundle"). Bundle 1 contains only our measure of intermediate input, while bundle 14 contains all variable expenditure.

The magnitude of this effect is economically striking. Markup estimates range from 4.35 to 0.95 — the difference between apparent hyper-competition and severe market power. This variation arises not from true economic differences but from arbitrary methodological choices about input aggregation.

To reiterate the conclusion of Section 6, this means that estimated markups are effectively arbitrary. The level of empirical markup series — such as that of De Loecker et al. (2020) — reflects their choice of 'variable input', rather than market power. Had De Loecker et al. observed a narrower input bundle, they would have found a higher markup. Likewise — and as evidenced by Traina (2018)'s series — had they taken a larger, more aggregated bundle, then they would have found a much lower markup, flipping the corresponding conclusion about competition in the US economy on its head. Simply by varying the number of inputs in our observed bundle, we can find any markup, and thus come to any conclusion we want.

We thus conclude that, in the absence of perfectly disaggregated data that observes inputs individually, the production approach to markup estimation does not recover the true markup levels.

8.2 | Estimated Markup Series

In this section, we present our estimated markup series, using it to validate our prior theoretical derivations of bias. In Section 8.2.1 we examine elasticity estimates and their structural relationship with revenue shares, confirming the empirical validity of our revenue-derived elasticity estimate expression. Section 8.2.2 then analyzes markup levels and trends, comparing our estimates to those of De Loecker et al. (2020). In Section 8.3.3, we make further comparisons to alternative revenue-data-based competitive metrics, providing evidence that they, alongside markups, are primarily driven by cyclical economic effects. Finally, Section 8.2.4 investigates cross-industry markup variation, revealing how the researcher's choice of variable input affects the industry dispersion of markup estimates.

8.2.1 | Elasticity Estimates and Revenue Bias

We now present our time-series elasticity and markup estimates. We estimate separate elasticities via the Levinsohn and Petrin method for each industry/year group, covering 39 identified industries[22] over the period 2003-2022. Markups are then calculated for each firm as:

$$\mu_{it} = \theta_{jt} \frac{Y_{it}}{M_{it}}$$

Where Y_{it} is our observed revenue variable, M_{it} is our observed expenditure on intermediate inputs, and θ_{jt} is the estimated industry/year elasticity of a specified bundle.

Two specifications of the revenue function are estimated, differing in the contents of their observed bundle. In the first specification, we estimate and use the elasticity of the intermediate input (M_{it}) alone, implementing controls for all other expense categories. This is the approach suggested by theory — a single, flexible input used for markup estimation. The second specification uses the sum of intermediate input and all reported expenses. This is bundle 13 from the previous section and covers the opposite extreme of the possible observed bundles.

We find the estimated elasticities of both specifications to be relatively stable over the 20 periods, consistent with the estimates of De Loecker et al. (2020). Estimates on our full bundle, ranging from 0.79-0.87, bear a strong resemblance to those of De Loecker et al., which range from 0.85-0.92. This may be because our 'full bundle' is analogous to the comprehensive 'cost of goods sold' they use. However, given we have found estimated elasticities to exhibit diminishing returns to bundle size, it is also possible that all estimates on large bundles fall within this range.

[22] These are rather broad industry classifications derived from NZSIOC, and are the only ones available in the productivity dataset.

The estimate on our single input varies more, ranging from 0.41-0.54, but this is to be expected given that it makes up much less of a firm's overall expenditure and is therefore more volatile.

Interestingly, the variation present in the elasticity estimates has a strong inverse relationship with the ratio of revenue to bundle expenditure. The correlation between θ_{jt} and $\frac{P_{it}Q_{it}}{P_{it}^{M}M_{it}}$ is -0.35 for the single input, and a striking -0.52 for the full bundle. This is observable in Figure 11, which plots both on the same axis under the full bundle specification.

Figure 11 — Estimated elasticities and inverted revenue share

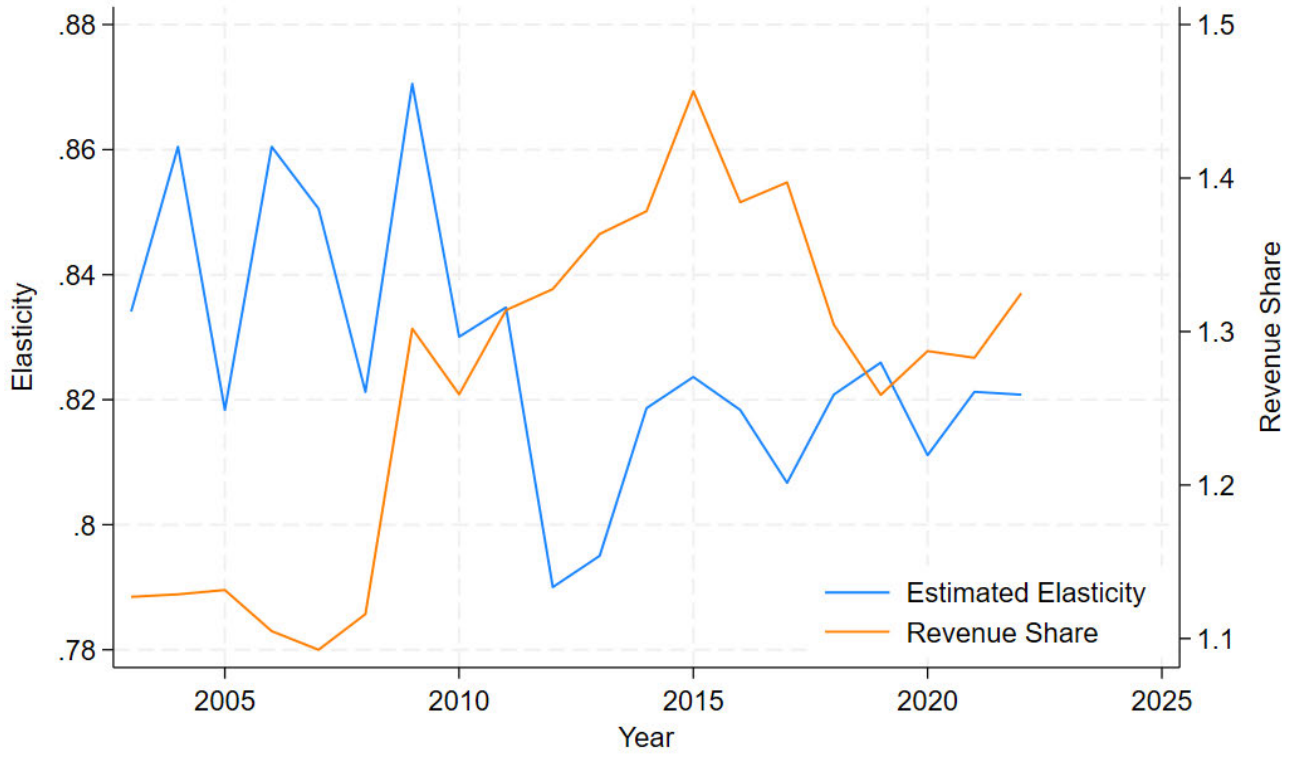

Figure 11: Elasticity estimates and revenue shares under the full bundle specification.

If our estimate were equivalent to the true output elasticity and trended over significant periods, with firms being sensitive to those trends, this would be economically justifiable. If a firm sees that their inputs are more productive, they may expand output and decrease price such that an increase in θ_{jt} is balanced by a decrease in $\frac{P_{it}Q_{it}}{P_{it}^{M}M_{it}}$, with the effect on the markup being ambiguous. This is exactly what we observed in Figure 2 of Section 5.4. In our case, however, it does not seem a plausible explanation. In the absence of sweeping technological revolutions, the inherent productivity of a given input is unlikely to vary significantly from year to year. This is even more true of our full bundle, which includes everything from wages to insurance and depreciation. Whatever

variation does occur is not guaranteed to be perceived by firms, much less provoke the expansion or contraction of output in response.

The strong inverse relationship we observe is more plausibly explained by the structural relationship between markups and revenue elasticity estimates derived in Section 5.3. That is:

$$plim\ \hat{\theta}_t^r = E\left[\frac{\theta_t}{\mu_{it}}\right] + \frac{cov\left(\frac{\theta_t}{\mu_{it}}, m_{it,\perp}^2\right)}{var(m_{it,\perp})} + \frac{cov(m_{it,\perp}, d_{it,\perp})}{var(m_{it,\perp})}$$

Given the definition of the markup, we can re-express this as:

$$plim\ \hat{\theta}_t^r = E\left[\frac{P_{it}^M M_{it}}{P_{it}Q_{it}}\right] + \frac{cov\left(\frac{P_{it}^M M_{it}}{P_{it}Q_{it}}, m_{it,\perp}^2\right)}{var(m_{it,\perp})} + \frac{cov(m_{it,\perp}, d_{it,\perp})}{var(m_{it,\perp})}$$

Hence, the revenue elasticity estimate is mechanically related to $\frac{P_{it}Q_{it}}{P_{it}^M M_{it}}$, and unless $E\left[\frac{P_{it}^M M_{it}}{P_{it}Q_{it}}\right]$ is dominated by an increase in the other terms, this relationship will be negative.

This empirical pattern provides strong evidence of the validity of our derived expression for the revenue elasticity estimate. The inverse relationship we observe between estimated elasticities and revenue shares directly mirrors the mechanical connection derived in Section 5.3, where $\hat{\theta}_t^r$ is structurally linked to $\frac{P_{it}^M M_{it}}{P_{it}Q_{it}}$. This structural relationship creates an automatic variance-flattening mechanism that dampens markup fluctuations — a pattern clearly evident in De Loecker et al.'s (2020) own estimates, where their benchmark series shows conspicuously muted volatility (higher lows and lower highs) compared to their constant elasticity series where the estimate elasticity is replaced by a constant 0.85.

The critical implication is that revenue data fundamentally prevents recovery of true elasticity levels and trends. This explains the first puzzle posed in Section 3.2: why De Loecker et al.'s estimated elasticities fail to adjust downward to offset the diminishing productive relevance of cost of goods sold over time. Were recovery of elasticity trends not hindered by revenue bias, we would observe elasticity estimates declining in tandem with changing input productivity, neutralizing the trend in revenue shares pointed out by Traina (2018). Instead, the revenue bias ensures that revenue share movements dominate, producing the upward trend observed by De Loecker et al.

8.2.2 | Markup Estimates and Trends

We now move on to our markup estimates. We consider two weighting schemes for calculating the overall markup each year; the first is a simple average, and the second is revenue-weighted:

$$\mu_t = \sum_i \rho_{it}\mu_{it}$$

where ρ_{it} is the ratio of firm i's revenue to the total revenue of the sample in year t. These are both plotted in Figure 12.

Figure 12 — Estimated markups

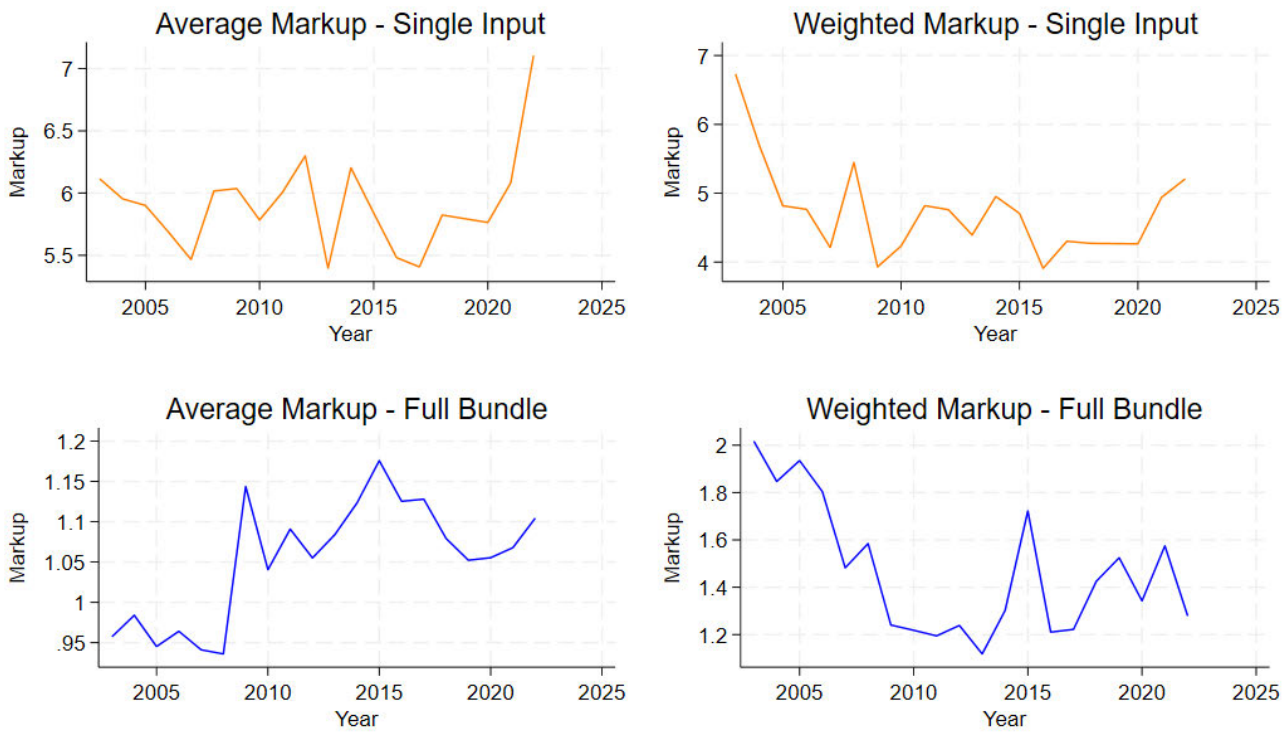

Figure 12: Estimated markup series under the weighted and unweighted single input and full bundle specifications. Evolution over 2003-2022.

Consistent with our previous findings, we observe that the markup estimates under the single-input specification are much higher than those of the full-bundle specification, and beyond what is economically reasonable. Despite this, both series are quite similar in shape, with misspecification bias having little impact on trend. This is to be expected. Unless misspecification bias itself — i.e., the gap between the true and observed bundles — evolves over time, it should have no effect on the time variation of estimated markups.

Additionally, and consistent with De Loecker et al. (2020), we find that weighting by revenue has a significant influence on the estimated trend. This is evident from the differences between our simple average and revenue-weighted series. The revenue-weighted markup is generally higher in level and trends in a stronger manner, implying that the markups of the larger firms in the sample are higher and more sensitive to the conditions determining markups.

Direct comparison between our revenue-weighted, full-bundle series, with that of De Loecker et al. reveals significant differences. De Loecker et al.'s series remains stable around 1.45-1.48 for the entirety of the 2000-2010 period, before increasing year-on-year to 1.61 in 2016[23]. Our estimates, on the other hand, exhibit a strong downward trend from 2-1.2 over 2003-2010, spiking to 1.7 only in 2015 and then returning to 1.2 in 2016.

It is remarkable that our markup series takes values both above and below De Loecker et al.'s series for the same periods. Given the apparent similarities in our elasticity estimates, this can only be attributed to differences in the inverted revenue share ($\frac{P_{it}Q_{it}}{P_{it}^{M}M_{it}}$). Were our observed bundles dissimilar, we would expect a consistent gap between our estimated markups. What we observe instead suggests that our observed bundle is of a comparable size, but is simply more volatile, either due to its composition, the relative instability of the NZ economy, or other unknown differences in our markup series.

8.2.3 | Competition Metrics and Economic Trends

While the difference in volatility between our estimates and those of De Loecker et al. is puzzling, the movement of our estimates does not seem to be spurious. As shown in Figure 13, there is a clear connection between the average estimated markup and the general trend of the New Zealand economy (GDP growth), with both declining over 2003-2008, recovering between 2009-2015, then dipping again over 2018-2020.

[23] On the whole, their markup series increases steadily from 1980-2016. Recall that Traina (2018) proposed that this was due to the diminishing importance of the physical-production oriented 'cost of goods sold' bundle, relative to other expenses — namely SG&A — which are not included in cost of goods sold. We are now equipped to understand that this is caused by the use of revenue data, which prevents recovery of the true output elasticity trend. Were elasticities identified correctly, we would perhaps see the estimated elasticity of COGS decrease over time, neutralizing the upward trend.

Figure 13 — Estimated Markups and NZ GDP Growth

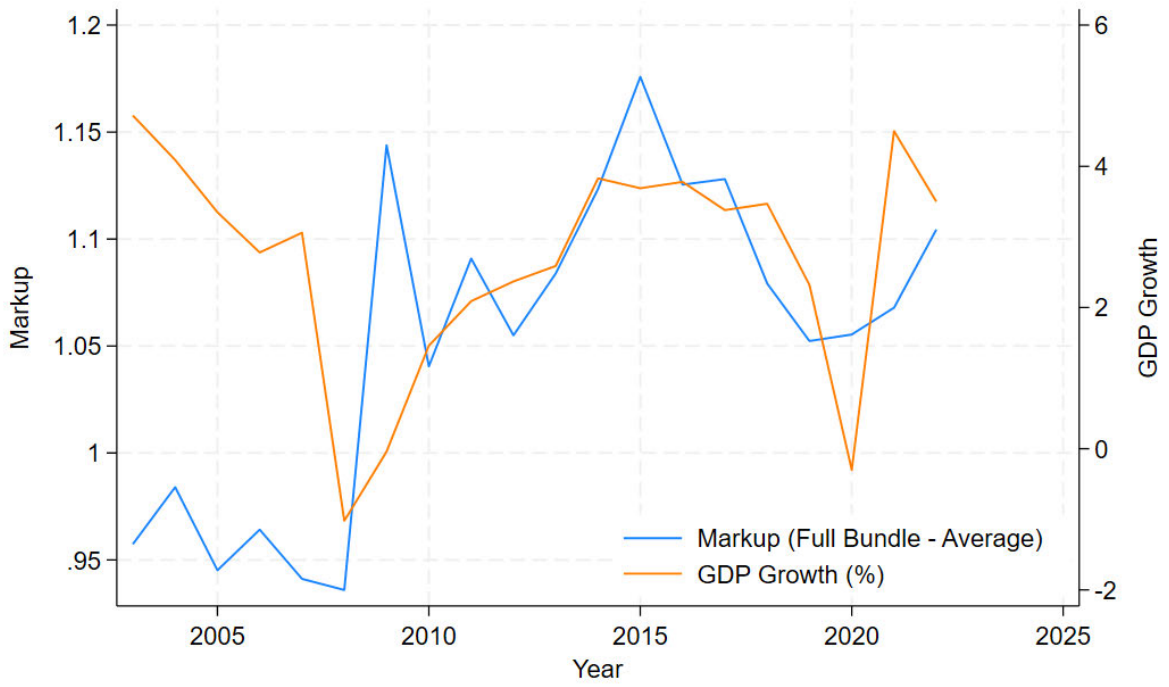

Figure 13: Estimated markups and New Zealand GDP growth (World Bank, 2025).

While the co-movement with GDP growth suggests our series is not mere noise, our estimated markups seem to capture general economic trends, which are not necessarily trends in market power.

Rather than being informative about competition, the relationship between our estimated markup and economic trends is purely mechanical, following inevitably from the relative constancy of our estimated elasticity. With variation in $\frac{P_{it}Q_{it}}{P_{it}^{M}M_{it}}$ dominating variation in θ_{jt}, our markup estimate is essentially a scaled series of the ratio of revenue to expenditure. Economic conditions induce change in this ratio — firm revenue decreasing in recessions and increasing in expansions — thereby tying our estimated markup to trends of the general economy. This would only be acceptable if we assume competition between firms to be closely related to cyclical economic trends — an assumption that would also render the construction of markups unnecessary. If, however, there is movement in competition which is independent of general economic conditions, our estimates seem unable to capture it, and are therefore insufficient as a measure of competition, at least in the short term.

Other direct measures of competition, which similarly rely on accounting-based measures of revenue and cost, also exhibit this pattern. Figure 14 plots NZ GDP growth alongside the NZ price-cost margin series constructed by Fabling and Maré (2019). We see that the series closely mirrors economic conditions, with a pre-2010 decline and post-2010 recovery.

Figure 14 — Price-Cost Margin and NZ GDP Growth

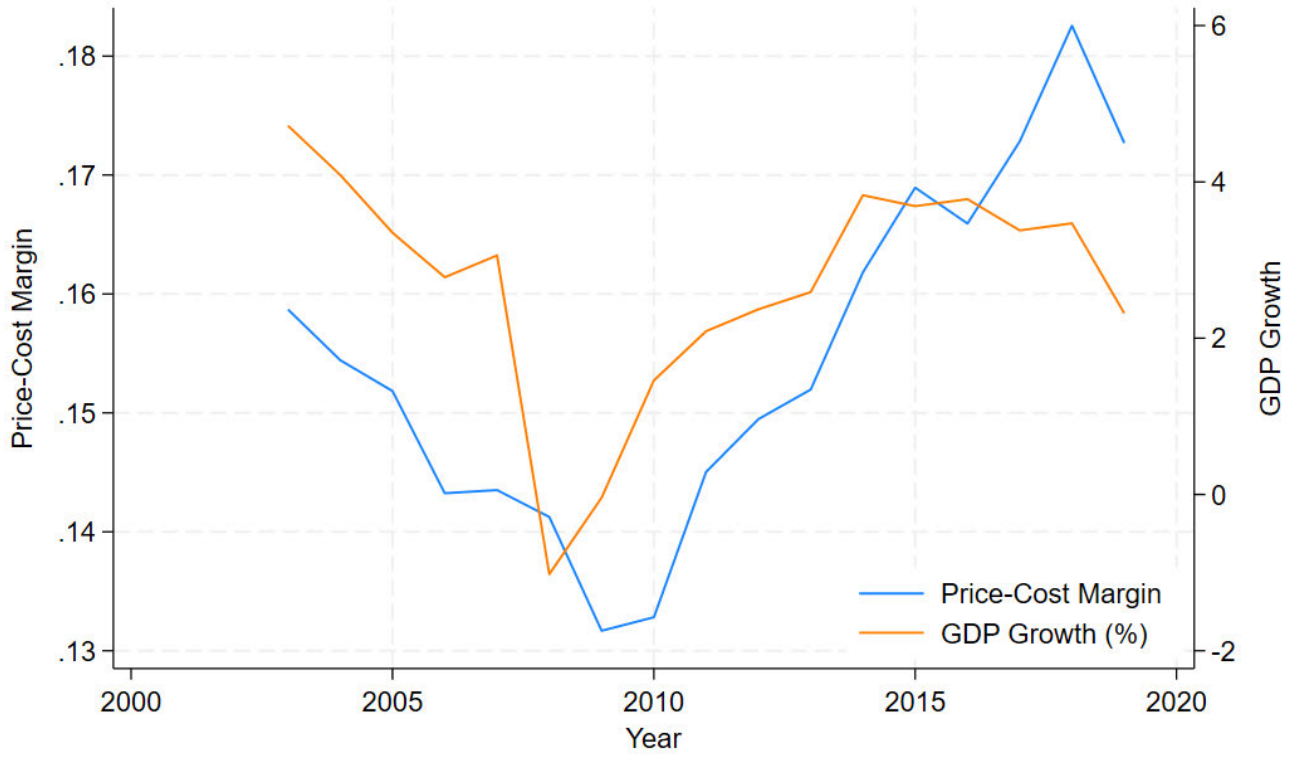

Figure 14: Fabling and Maré (2019) price-cost margin and New Zealand GDP growth (World Bank, 2025).

In addition, Figure 15 (below) shows a strong resemblance between our markup estimates and the Fabling and Maré PCM series. Despite using somewhat different samples,[24] the driving force behind both measures is the relative magnitudes of revenue and cost, hence, their co-movement is expected. The average price cost margin, as defined by Fabling and Maré (2019), is:

$$\overline{PCM_{jt}} = \frac{1}{N_{jt}} \sum_{i=1}^{N_{jt}} \max\left\{\frac{Y_{ijt} - C_{ijt}}{Y_{ijt}}, -1\right\}$$

where Y_{ijt} is revenue and C_{ijt} is total variable cost. This is very similar to the revenue share term which dominates our markup. If we were to express the revenue share in similar terms, it would simply be:

$$revenue\ share = \frac{Y_{it}}{C_{it}}$$

[24] The Fabling and Maré PCM series is derived from the same productivity dataset we use for our markup estimation, however we end up dropping roughly half of the original sample.

Both terms will increase as revenue grows relative to cost and decrease as cost grows relative to revenue. PCM (which is unweighted) and the unweighted markup series will therefore move closely with each other, being driven by cyclical economic trends.

Figure 15 — Price-Cost Margin and Markups

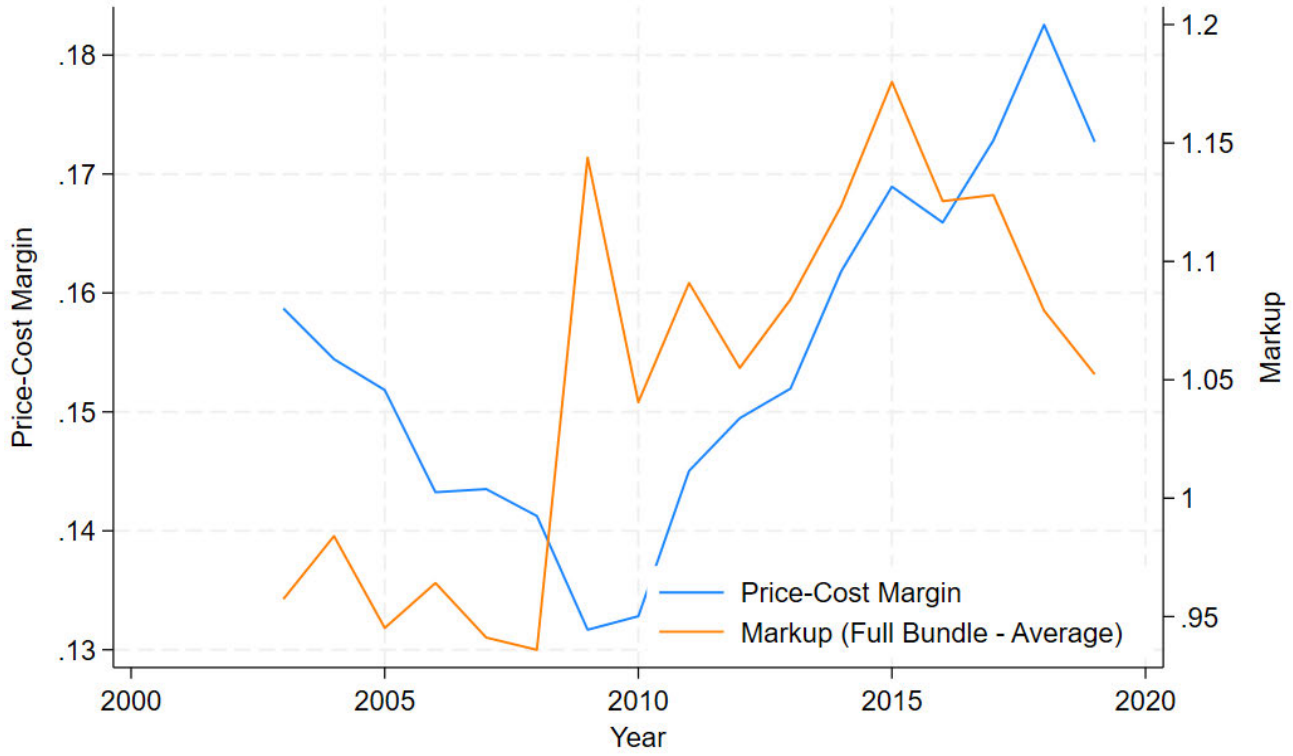

Figure 15: Estimated markups and Fabling and Maré (2019) price-cost margin.

Interestingly, the same applies to profit elasticity, which we discussed in Section 4.2. Figures 16 and 17 plot Fabling and Maré (2019)'s standard and fixed effects profit elasticity series against NZ GDP growth. Again, both series seem to move closely with GDP, indicating their sensitivity to cyclical trends.

Figure 16 — Profit Elasticity and NZ GDP Growth (No Fixed Effects)

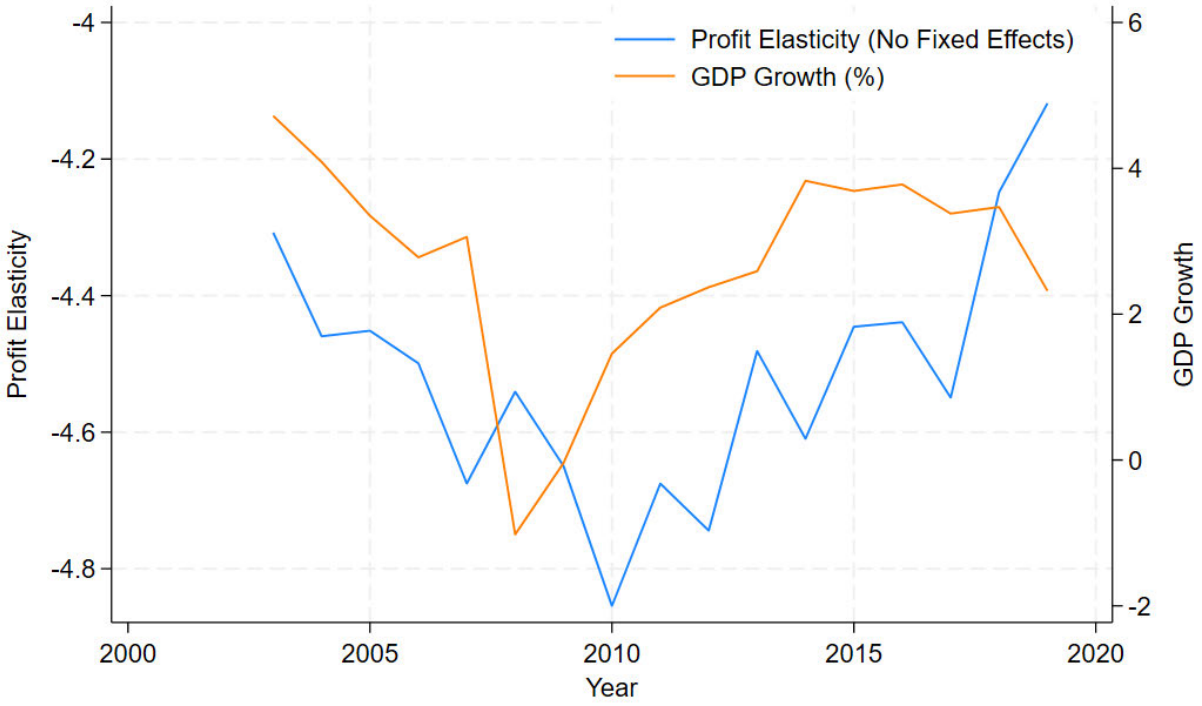

Figure 16: Fabling and Maré (2019) profit elasticity (no fixed effects) and New Zealand GDP growth (World Bank, 2025).

Figure 17 — Profit Elasticity and NZ GDP Growth (Fixed Effects)

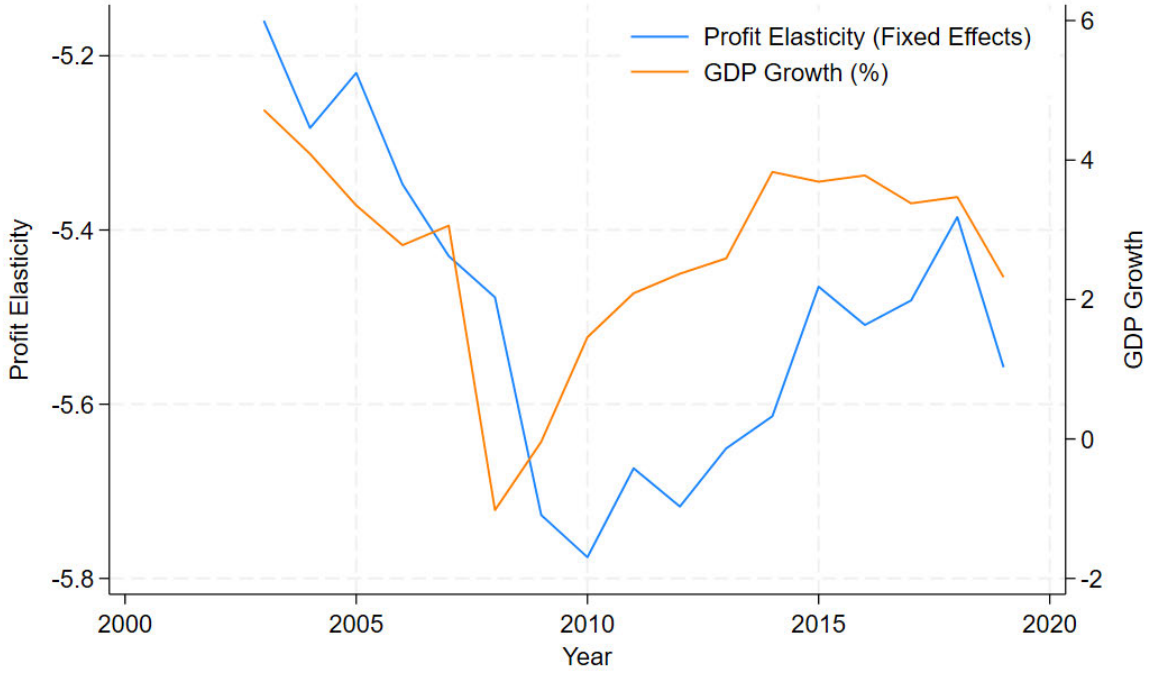

Figure 17: Fabling and Maré (2019) profit elasticity (fixed effects) and New Zealand GDP growth (World Bank, 2025).

Unlike price-cost margins and revenue shares, profit elasticity, rather than being a simple expression of revenue and cost terms, attempts to directly estimate the reward for efficiency with a regression of profit on marginal cost. In practice, this becomes a regression of profit on the ratio of cost to revenue:

$$\ln(Y_{it} - C_{it}) = \alpha + \beta \ln\left(\frac{C_{it}}{Y_{it}}\right)$$

While the revenue and cost terms are still present here, estimated profit elasticity is not moved directly by them. If revenue drops proportionally for all firms while costs stay the same, markups and price-cost margins will drop, but the within-industry relationship between profit and costs will stay the same.

Although there is no *mechanical* link binding profit elasticity estimates to economic trends, we have already described a plausible channel through which a relationship between them could arise. Recall that we previously criticized profit elasticity for being effectively blind to market power, as it will always indicate strong competition so long as firms with low cost-to-revenue ratios are making high profit. For profit elasticity to function as a measure of competition, we are relying on the assumption that this ratio is driven by efficiency — i.e., more competitive firms which produce with lower costs — rather than firms using market power to increase their markups. But if this is not the case, then negative profit elasticity estimates may simply be the product of high markup firms — which will naturally have low cost-to-revenue ratios — making more profit than low markup firms, which is almost true by definition.

This can explain movements in profit elasticity. If the profit of high market power firms is less sensitive to economic conditions than low market power firms, economic trends will drive profit elasticity. In recessions, the profit of low market power firms will fall relative to those with high market power, strengthening the negative relationship between profit and the cost-revenue ratio. When the economy recovers, so will the profits of low market power firms, and so the estimated competitive trend will reverse accordingly.

8.2.4 | Industry Markup Comparison

The preceding subsection has shown that variation in markup estimates is driven primarily by variation in revenue shares, causing time series markup estimates to mirror cyclical economic trends. A similar principle applies to the variation of markups across industries. Unlike time-variation, which moves with economic conditions, the industry-dispersion of revenue shares will depend on industry characteristics — including true markups — and the researcher's choice of variable input.

The industry breakdown of our estimated markup (Tables F1 and F2 in Appendix F) highlights the extent of the estimated markup's dependence on mechanical processes, particularly those related to the choice of variable

input. As Table 6 shows, the highest-markup industries under the unweighted[25] single-input specification — where only purchases are considered — are also those that depend the least on physical production processes which consume purchased inputs. Of the industries with the top five highest markups, scientific and tech services, as well as financial and insurance services, intuitively use few physical inputs. On the other hand, the dairy, mining, and forestry industries, despite having tangible output, rely primarily on labor and capital, rather than purchasable inputs.

The low-markup industries reinforce this pattern. Four of the bottom five are centered around reselling, namely: supermarkets, vehicle and vehicle parts retail, other retail, and wholesale trade. Purchases, which are then resold, will naturally make up the majority of expenditure for firms in these industries. This mechanically compresses their revenue shares and — as evidenced by the perfect overlap between the industries with the lowest revenue shares and the industries with the lowest markups — results in very low estimated markups.

Table 6 — Highest, Lowest, and Median Industry Markups (Single Input)

Rank	Industry	Description	Markup	Elasticity	Revenue Share
1	MN11	Professional, scientific & tech. serv.	16.47	0.46	33.79
2	AA13	Dairy Cattle Farming	15.58	0.22	70.63
3	KK13	Auxiliary finance & insurance serv.	15.44	0.41	38.70
4	BB11	Mining	14.91	0.41	37.74
5	AA21	Forestry and logging	13.72	0.48	28.98
20	LL11	Rental & hiring serv.	7.09	0.35	19.93
35	CC5	Petrochemical product manufacturing	2.44	0.49	4.93
36	FF11	Wholesale trade	1.68	0.34	4.86
37	GH13	Other store-based & non-store retailing	1.49	0.37	3.98
38	GH11	Motor vehicle/parts & fuel retailing	0.72	0.27	2.56
39	GH12	Supermarket, grocery & spec. food retailing	0.42	0.26	1.61

Table 6: Industry average markups, estimated elasticities, and revenue shares for the single-input specification within the merged set of single-input and full-bundle elasticity outputs.

The low markups of high-purchase firms may be partially attributable to the underestimation of output elasticities caused by bias from revenue data. Intuitively, we would expect the reliance of these firms on

[25] We focus on the unweighted estimates in this section to highlight and better explore mechanical distortions. These distortions are obscured, but not removed or solved by the simple process of weighting.

purchases to be reflected in their 'productive' process and a correspondingly high elasticity. Instead, we find the average estimated elasticity of the bottom four industries to be 0.31, below the sample average of 0.39.

While elasticities may thus be a factor, as was the case for the time series estimates, their effect on the industry-dispersion of markups seems to be overshadowed by the dominance of revenue shares. Comparing the bottom and median-markup industries, we find that the elasticity of the lowest-markup industry is 0.7x that of the median industry, while their revenue share is 7x larger. For the lowest markup to equal the median, we would need to raise the bottom industry's elasticity from 0.26 to 4.4 — a seventeen-fold increase which is economically unjustifiable.

The researcher's choice of variable input is thus extremely consequential, independent of misspecification or revenue bias. There is significant variation in the types of inputs used by different industries, and any specific input (or narrowly defined category of inputs) is unlikely to be equally relevant to all of them. Just as some industries don't rely on purchases or intermediate inputs, others may scarcely use labor, pay little rent, have small marketing budgets, and so on. The obvious course is to cast a broad net with large input bundles; however, as we have seen, misspecification bias renders this approach equivalent to decreasing the markup arbitrarily.

This is demonstrated (again) in Table 7, where we see the estimated markups of all industries drop precipitously. More notably, the top end of the markup rankings sees significant change, with dairy cattle farming, financial services, and scientific services — high markup industries under the previous specification — dropping to the 13th, 22nd, and 29th places respectively. This shift is caused by the entrance of these industries' relevant expenses into the defined variable input under the full-bundle specification. That their fall is so dramatic demonstrates that their high markups under the single-input specification were caused mechanically by the narrowness of the observed bundle, rather than reflecting the true markups of these industries.

Table 7 — Highest, Lowest, and Median Industry Markups (Full Bundle)

Rank	Industry	Description	Markup	Elasticity	Revenue Share
1	AA21	Forestry and logging	1.5	1.2	1.2
2	AA31	Fishing and aquaculture	1.4	1.0	1.5
3	AA11	Horticulture and fruit growing	1.3	1.1	1.1
4	BB11	Mining	1.2	1.0	1.3
5	EE11	Building construction	1.2	0.9	1.4
20	KK13	Auxiliary finance & insurance serv.	1.1	0.9	1.2
35	GH21	Accommodation & food serv.	0.8	0.9	1.0
36	GH13	Other store-based & non-store retailing	0.5	0.8	0.6
37	FF11	Wholesale trade	0.3	0.6	0.5
38	GH11	Motor vehicle/parts & fuel retailing	0.2	0.4	0.4
39	GH12	Supermarket, grocery & spec. food retailing	0.2	0.5	0.4

Table 7: Industry average markups, estimated elasticities, and revenue shares for the full-bundle specification within the merged set of single-input and full-bundle elasticity outputs.

In contrast to the shift at the top end, the bottom five remains almost unchanged between the two specifications. The bottom four consists of exactly the same firms, with the rank exchange between wholesale trade and 'other' retail being the only difference. This is explained entirely by misspecification bias. For these purchase-reliant industries, the single-input specification already fully captured their relevant inputs. The full-bundle's definitional expansion merely adds superfluous expenditures which further inflates observed expenditures while contributing only noise to the elasticity estimates.

9 | Discussion

Our results indicate three things. Firstly, that the level of an estimated markup will be largely determined by the size of the observed bundle. Secondly, that when revenue data is used, estimated elasticities are mechanically and inversely related to the inverted revenue share. And thirdly, that in the short run, markup estimates are dominated by the inverted revenue share, which itself primarily reflects cyclical economic trends.

Each of these is uniquely damaging to the feasibility of markup estimation. Issues with the use of revenue data were raised prior by Bond et al. (2021), and our results further evolve this line of critique. Not only will the level of revenue-derived elasticity estimates be merely tangentially related to the true output elasticity, but the structural

form of the revenue elasticity estimate also precludes the identification of its trends — governed as it is by an inverse relationship with the inverted revenue share, and covariance terms which scale nonlinearly with market power and the true output elasticity.

The preclusion of trend identification in particular works unfavorably alongside the short-term dominance of economic trends on markup estimates — an issue common to all markup series covering limited periods. Even when output data is used, and output elasticities are correctly recovered, over the short term, markup estimates will be dominated by variation in the inverted revenue share. A markup series covering only a decade or two, therefore, will primarily reflect short term, cyclical economic trends. If there is a long-term rise in market power, for example, but our series only covers a recessionary period, all we will observe is a decline in markups.

Such an observation would not be incorrect. In a recession, we would indeed expect the markup, as defined, to fall; but this is simply not useful. There is no lack of general economic indicators. Even if output elasticities are constant in the short term, creating the perfect correlation between true and estimated markups proposed by De Ridder et al. (2024), we will not have learned much more than we could glean from generic reports on the health of the economy.

It is only over long periods that we will be able to distinguish cyclical trends in the economy from long-term trends in market power. This places a burden of adequate data collection on the researcher as a baseline requirement — one which we ourselves have not fulfilled. However, even if the span of our data were sufficient, over the long-run, output elasticities are more likely to move. This makes assumptions of constant elasticity tenuous, and so recovery of long-run markup trends will depend on recovery of movements in the output elasticity. It is exactly these movements which revenue data is ill-suited to identify.

Finally, while not directly impeding the identification of markup trends, misspecification bias, and the quandary of properly defining the ‘true’ bundle of inputs, renders estimated markups all but meaningless. With the level of the estimated markup being mechanically determined by bundle size — in effect, by the researcher’s choice of bundle — we cannot reasonably make any inference about market power or competition. In a similar vein, without a consistent definition of the observed bundle, markup estimates cannot be compared across studies, and any such attempt is bound to lead to confusion. This is demonstrated perfectly by the comparison between De Loecker et al. (2020) and Traina (2018). Crucially, these issues arise regardless of whether we have access to output data, posing a fundamental complication to markup estimation.

Given the breadth and severity of these issues, we conclude that the production approach to markup estimation, in empirical settings that face data constraints, is flawed, and does not recover the level or trend of estimated markups. Estimates are instead a composite of various biases, each alone sufficient to invalidate markups as a

reliable measure of competition. For markups to remain a viable metric in competition analysis, future research must develop empirical strategies that directly address and circumvent these biases.

References

Ackerberg, D. A., Caves, K., & Frazer, G. (2015). Identification properties of recent production function estimators. *Econometrica, 83*(6), 2411–2451. https://www.jstor.org/stable/43866416

Basu, D. (2023). *The Yule-Frisch-Waugh-Lovell Theorem for linear instrumental variables estimation* [Working paper]. SSRN. https://doi.org/10.2139/ssrn.4514656

Berry, S., Gaynor, M., & Scott Morton, F. (2019). Do increasing markups matter? Lessons from empirical industrial organization. *Journal of Economic Perspectives, 33*(3), 44–68. https://doi.org/10.1257/jep.33.3.44

Bond, S., Hashemi, A., Kaplan, G., & Zoch, P. (2021). Some unpleasant markup arithmetic: Production function elasticities and their estimation from production data. *Journal of Monetary Economics, 121*, 1–14. https://doi.org/10.1016/j.jmoneco.2021.05.004

Boone, J., van Ours, J. C., & van der Wiel, H. (2007). *How (not) to measure competition* (TILEC Discussion Paper No. 2007-014). TILEC. https://doi.org/10.2139/ssrn.985270

De Loecker, J., Eeckhout, J., & Unger, G. (2020). The rise of market power and the macroeconomic implications. *The Quarterly Journal of Economics, 135*(2), 561–644. https://doi.org/10.1093/qje/qjz041

Demsetz, H. (1973). Industry structure, market rivalry, and public policy. *Journal of Law and Economics, 16*(1), 1–9. https://doi.org/10.1086/466752

De Ridder, M., Grassi, B., & Morzenti, G. (2024). *The hitchhiker's guide to markup estimation* (Discussion Paper No. 2210). Centre for Macroeconomics (CFM). https://www.lse.ac.uk/CFM/assets/pdf/CFM-Discussion-Papers-2022/CFMDP2022-10-Paper3.pdf

Fabling, R., & Maré, D. C. (2015). *Production function estimation using New Zealand's Longitudinal Business Database* (Motu Working Paper No. 15-15). Motu Economic and Public Policy Research. https://motu-www.motu.org.nz/wpapers/15_15.pdf

Fabling, R., & Maré, D. C. (2019). *Competition and productivity: Do commonly used metrics suggest a relationship?* (Motu Working Paper No. 19-16). Motu Economic and Public Policy Research. https://motu-www.motu.org.nz/wpapers/19_16.pdf

Fabling, R., & Maré, D. C. (2019). *Improved productivity measurement in New Zealand's Longitudinal Business Database* (Motu Working Paper No. 19-03). Motu Economic and Public Policy Research. https://motu-www.motu.org.nz/wpapers/19_03.pdf

Ganapati, S. (2024). *The modern wholesaler: Global sourcing, domestic distribution, and scale economies* (NBER Working Paper No. 32036). National Bureau of Economic Research. https://doi.org/10.3386/w32036

Gandhi, A., Navarro, S., & Rivers, D. A. (2020). On the identification of gross output production functions. *Journal of Political Economy, 128*(8). https://doi.org/10.1086/707736

Hall, R. E. (1988). The relation between price and marginal cost in U.S. industry. *Journal of Political Economy, 96*(5), 921–947. https://doi.org/10.1086/261570

Hashemi, A., Kirov, I., & Traina, J. (2022). The production approach to markup estimation often measures input distortions. *Economics Letters, 217*, 110673. https://doi.org/10.1016/j.econlet.2022.110673

Levinsohn, J., & Petrin, A. (2003). Estimating production functions using inputs to control for unobservables. *The Review of Economic Studies, 70*(2), 317–341. https://doi.org/10.1111/1467-937X.00246

Marschak, J., & Andrews, W. H., Jr. (1944). Random simultaneous equations and the theory of production. *Econometrica, 12*(3/4), 143–205. https://doi.org/10.2307/1905432

Olley, G. S., & Pakes, A. (1996). The dynamics of productivity in the telecommunications equipment industry. *Econometrica, 64*(6), 1263–1297. https://doi.org/10.2307/2171831

Traina, J. (2018). *Is aggregate market power increasing? Production trends using financial statements* [Working paper]. SSRN. https://doi.org/10.2139/ssrn.3120849

Van Dijcke, D. (2023). *On the non-identification of revenue production functions* (Bank of England Working Paper No. 1015). Bank of England. https://doi.org/10.2139/ssrn.4410709

World Bank Group. (2025). *World Development Indicators: GDP growth (annual %) – New Zealand*. https://data.worldbank.org/indicator/NY.GDP.MKTP.KD.ZG?locations=NZ

Appendix

Appendix A1 — Olley and Pakes (1996) Productivity Control Method

The OP method relies on a formalization of the investment process which is characterized by a series of assumptions. For consistency with later discussion, we shall refer to the assumptions as formulated in Ackerberg et al., (2015), which is a more modern overview of the method.

The first assumption consists of two parts. The first describes the information which firms possess and can use to inform their decisions. Crucially, this includes the persistent productivity shock. Productivity affects output, which is observed by firms, and so even if firms cannot quantify or articulate their productivity, they will understand it to some extent. Firm information in the current period, therefore, will include the past and present levels of productivity in addition to information on capital, labor, input, etc.

The second part states that conditional on the information set, which includes productivity, any remaining output shock is transitory. This stems from the dependence of output on the factors of production. In theory, if we have knowledge of the persistent productivity shock, the levels of the factors, and their output elasticities, we would be able to perfectly predict output. Deviation from this predicted level would therefore reflect unpredictable circumstance outside of the production function[26], and would not be used in the decision-making process.

Assumption 1 — Firm information: In the current period t, the firm has an information set I_{it}, which includes the current and all past levels of productivity $\{w_{it}\}_{\tau=0}^{t}$. Output shocks e_{it} are transitory, satisfying $\mathbb{E}[e_{it}|\, I_{it}] = 0$

The second assumption describes a process of capital accumulation. The current level of capital will be determined entirely by the previous level of capital, and an investment that takes place in the prior period. Crucially, this means that k_{it} is determined in the past, in period $t-1$. In each period, firms will make an investment decision based on the current information, with the effect of this investment — which is an increase in capital — only coming into effect in the next period (this also necessitates that capital cannot change in the current period).

Assumption 2 — Capital accumulation: Firms accumulate capital according to the function

$$k_{it} = z(k_{it-1}, \iota_{it-1})$$

where ι_{it} is firm investment.

The third assumption specifies the manner in which firms can predict future productivity. Expected productivity incentivizes investment in the present, hence, this assumption characterizes the resource allocation process.

Assumption 3 — Evolution of productivity: Firm productivity evolves according to a first order Markov process, with associated probability distribution:

$$p(w_{it+1}|I_{it}) = p(w_{it+1}|w_{it})$$

This distribution is known to firms and is stochastically increasing in w_{it}.

[26] Potential sources of deviation would be all things that impact production without being predictable or related to inherent productivity or the factors of production. These could be things like labor strikes, abnormal weather, or reporting/measurement errors.

There are several elements to this assumption. Firstly, a first order Markov process is simply one where the probabilistic distribution of future outcomes depends only on the current state. Thus, in this context, the distribution of future productivity depends only on current productivity. The previous levels of productivity do not matter. Next, the distribution is stochastically increasing in w_{it}, meaning that higher levels of future productivity become more likely when current productivity is high (and lower levels become less likely). This reflects that firms with high productivity should be highly productive in the future, with large changes to the status-quo being unlikely. Firms understand this, and can therefore predict, to some degree, future productivity using current information. The equality between the two expressions simply makes this explicit. Future productivity is based on current productivity, which is within the firm's current information set, hence the two expressions are equivalent.

The next assumptions, 4 and 5, are perhaps the most important, as they allow us to model investment, and therefore productivity. These state that firm investment is a function of capital and productivity, and that the function is strictly increasing.[27] Firms with higher levels of capital and productivity will therefore invest more, creating a link between observable investment and unobservable productivity.

Assumption 4 — Firm investment: Firm investment is given by the function

$$\iota_{it} = f(k_{it}, w_{it})$$

Assumption 5 — Strict monotonicity: $\iota_{it} = f(k_{it}, w_{it})$ is strictly increasing in k_{it} and w_{it}.

With these assumptions in place, we can proceed to derive the form of the control. From the monotonicity assumption, and with capital being fixed in each period we can invert the investment function such that we have a function for productivity:

$$w_{it} = f^{-1}(k_{it}, \iota_{it})$$

While we do not know the exact form of this function, if we have capital and investment data, we can estimate it. Substituting this into the production function we get:

$$q_{it} = \beta_0 + \theta_t^v v_{it} + \theta_t^k k_{it} + \theta_t^l l_{it} + f^{-1}(k_{it}, \iota_{it}) + e_{it}$$

With the form of $f^{-1}(k_{it}, \iota_{it})$ being unknown, non- or semi-parametric methods would be the go-to estimation strategies, but there is a problem: k_{it} appears twice in this production function — once as factor of production, and once as an indicator of productivity (w_{it}), and it is difficult to separate these effects. To reuse an example from Ackerberg (2006), assuming a linear specification of $f^{-1}(k_{it}, \iota_{it})$, we would be trying to estimate something like:

$$q_{it} = \beta_0 + \theta_t^v v_{it} + \theta_t^k k_{it} + \theta_t^l l_{it} + \beta_{1,t} k_{it} + \beta_{2,t} \iota_{it} + e_{it}$$

Clearly, k_{it} and k_{it} are perfectly collinear in this instance, and we will not be able to separately estimate their coefficients or distinguish θ_t^k from $\beta_{1,t}$

For now, however, we do not need to distinguish these effects. For the purposes of markup estimation, we only require an estimate of the output elasticity of the input (θ_t^v). Controlling for productivity while losing the effect of

[27] That the function is strictly increasing in w_{it} mostly follows from assumption 3. Future productivity increases with current productivity, and so if future productivity incentivizes investment, investment will also increase with current productivity. Investment increasing with capital is a bit less founded, but we can consider that firms with more capital require a higher maintenance investment, and that large firms will have more resources to invest.

capital will suffice. Hence, we proceed by bundling the intercept (β_0), investment, and both capital terms into the non-parametric term $\Phi_t(k_{it}, \iota_{it})$. The function we estimate is then:

$$q_{it} = \theta_t^v v_{it} + \theta_t^l l_{it} + \Phi_t(k_{it}, \iota_{it}) + e_{it}$$

This can be done in many ways. The standard approach in the literature is to use a GMM procedure based on assumption 1:

$$\mathbb{E}[e_{it} | \, I_{it}] = \mathbb{E}[q_{it} - \theta_t^v v_{it} - \theta_t^l l_{it} - \Phi_t(k_{it}, \iota_{it}) | \, I_{it}] = 0$$

As Ackerberg (2015) points out, however, if Φ_t is assumed to have a polynomial form, then this can be as simple as running an OLS regression of q_{it} on v_{it} and Φ_t. In the case of a third-order polynomial we have:

$$q_{it} = \theta_t^v v_{it} + \theta_t^l l_{it} + \beta_{1,t} k_{it} + \beta_{2,t} \iota + \beta_{3,t} k_{it}^2 + \iota_{it}^2 + \beta_{4.t} k_{it} \iota_{it} + \beta_{5,t} k_{it}^3 + \beta_6 \iota_{it}^3 + \beta_7 k_{it} \iota_{it}^2 + \beta_8 \iota_{it} k_{it}^2$$

Regardless of which approach is used, we will get the estimate, $\hat{\theta}_t^v$, allowing us to construct the markup measure.

Appendix A2 — Levinsohn and Petrin (2003) Productivity Control Method

The Levinsohn and Petrin approach (LP) follows the Olley and Pakes method, the only difference being that, where Olley and Pakes use investment, to avoid data loss and 'lumpiness' Levinsohn and Petrin use intermediate inputs, which we denote with m_{it}. These are inputs which are purchased and used up in production and are often chosen as the considered variable input (v_{it}) in production function and markup estimation.

This leads to an alternate expression of assumptions 4 and 5 where investment is replaced by intermediate input usage. That is, intermediate inputs are a strictly increasing function of productivity and capital. This not only ameliorates the data issues that arise when using investment but are also weaker assumptions in general. Given that productivity is an output multiplier, and that the effect of capital is generally to optimize and facilitate the greater use of inputs, it directly follows that intermediate input use will increase with capital and productivity.

Assumption 4(b) — Firm intermediate input use: Firms' intermediate input use is given by the function:

$$m_{it} = f_t(k_{it}, w_{it})$$

Assumption 5(b) — Strict monotonicity: $m_{it} = f_t(k_{it}, w_{it})$ is strictly increasing in k_{it} and w_{it}.

Under these alternate assumptions, the productivity function becomes:

$$w_{it} = f^{-1}(k_{it}, m_{it})$$

and our production function becomes:

$$q_{it} = \theta_t^v v_{it} + \theta_t^k k_{it} + \theta_t^l l_{it} + f^{-1}(k_{it}, m_{it}) + e_{it}$$

Unlike the Olley and Pakes method, we cannot condense this function to obtain $\hat{\theta}_t^v$ while discarding the other effects. This is because m_{it} is a common choice for the considered variable input or is included within chosen input bundle, resulting in the production function specification:

$$q_{it} = \theta_t^m m_{it} + \theta_t^k k_{it} + \theta_t^l l_{it} + f^{-1}(k_{it}, m_{it}) + e_{it}$$

This is often imposed by the data itself. Rather than reporting expenditure on individual inputs, firms generally report expenditure by category. In available datasets, these categories are often further combined into aggregate variables such as 'total variable cost'. The best-case scenario is that our dataset will contain smaller bundles like

'purchases'. This is essentially intermediate input use, and so in almost all cases, our options for v_{it} will be such that m_{it} is equivalent to or included within it.

This issue cannot be circumvented by considering labor for the markup calculation (rather than 'variable input'), as this would violate cost-minimization. To minimize the use of an input for some level of output, that input must be frictionlessly adjustable — i.e., the use of that input can be changed at any time without any consequence beyond the immediate effect on output. Intermediate input use perfectly fulfills these criteria, whereas labor almost perfectly does not. Labor can be subject to unions, labor laws, contractual obligations, and personal relationships which can restrict or at least disincentivize the casual dismissal of workers. Additionally, firms may want to keep skilled or tenured workers even when it is not cost-optimal in the present situation. Keeping these workers may prove better in the long run, and there are costs associated with training new hires and losing talent to competition.

Hence, we will generally need to consider m_{it} — or a bundle which includes it — as our variable input, and must therefore recover its elasticity. The method to do so is a two-stage version of the estimation strategy set out in Olley & Pakes (1996).[28]

In the first stage, we estimate the production function with the control term, a non-parametric function which includes the intercept, capital terms, and intermediate input:

$$q_{it} = \theta^{l} l_{it} + \Phi_{\mathrm{t}}(k_{it}, m_{it}) + e_{it}$$

Assuming that $\Phi_{\mathrm{t}}(k_{it}, m_{it})$ is a third-order polynomial, this is equivalent to estimating the function:

$$q_{it} = \theta_t^{l} l_{it} + \beta_{1,t} k_{it} + \beta_{2,t} m_{it} + \beta_{3,t} k_{it}^2 + m_{it}^2 + \beta_{4.t} k_{it} m_{it} + \beta_{5,t} k_{it}^3 + \beta_6 m_{it}^3 + \beta_7 k_{it} m_{it}^2 + \beta_8 m_{it} k_{it}^2$$

This gives us $\hat{\theta}_t^{l}$ and $\hat{\Phi}_{\mathrm{t}}$; the former being the proportion of output attributable to labor, and the latter being the proportion attributable to capital, productivity, and intermediate input.

The second stage allows us to recover the effects of capital and intermediate input. Recall that firms understand future productivity to be a probabilistic distribution based on current productivity — this is the basis for their expectations and investment decisions. Of course, the productivity predictions made by firms will not be perfect, and so we can decompose current productivity into the portion predicted by firms, and the deviation of current productivity from that prediction:

$$w_{it} = \mathbb{E}[w_{it} | w_{it-1}] + \xi_{it} = g(w_{it-1}) + \xi_{it}$$

where $g(w_{it-1})$ is essentially the firm prediction function, giving the productivity expected in period t for each productivity observed in period $t-1$. Now, consider that, if we subtracted the intercept and the output elasticities of capital and intermediate input from Φ_t, we would be left with the isolated productivity effect:

$$w_{it} = \Phi_{\mathrm{t}}(k_{it}, m_{it}) - \beta_0 - \theta_t^{k} k_{it} - \theta_t^{m} m_{it}$$

and therefore:

$$w_{it-1} = \Phi_{\mathrm{t}-1}(k_{it-1}, m_{it-1}) - \beta_0 - \theta_t^{k} k_{it-1} - \theta_t^{m} m_{it-1}$$

Hence, we can formulate an expression for w_{it} using only the past levels of the productive factors:

[28] We ignored this previously as, when using the investment control, we can recover the elasticity of m_{it} without a second stage.

$$w_{it} = g(\Phi_{t-1}(k_{it-1}, m_{it-1}) - \beta_0 - \theta^k k_{it-1} - \theta^m m_{it-1}) + \xi_{it}$$

Substituting this into the production function we have:

$$q_{it} = \theta^l l_{it} + \theta^m m_{it} + \theta^k k_{it} + g(\Phi_{t-1}(k_{it-1}, m_{it-1}) - \beta_0 - \theta^k k_{it-1} - \theta^m m_{it-1}) + \xi_{it} + e_{it}$$

This gets us very close to unbiased output elasticity estimates. The intuition here is that, after controlling for unobserved productivity with our $g(.)$ function, the remaining variation in output is caused by the inherent productivity of the factors, and the unexpected productivity shock ξ_{it}. To the extent that the factors were chosen in the previous period — k satisfies this inherently with assumption 2, and we can instrument m_{it} with m_{it-1} — they will be completely uncorrelated with the unanticipated shock ξ_{it}. Hence, we have effectively controlled for simultaneity and omitted variable bias.

The only remaining step is estimation, and for this we use the moment condition:

$$\mathbb{E}[\xi_{it} + e_{it} | \, I_{it}] =$$

$$\mathbb{E}[q_{it} - \theta^l l_{it} - \theta^m m_{it} - \theta_t^k k_{it} - g(\Phi_{t-1}(k_{it-1}, m_{it-1}) - \beta_0 - \theta^k k_{it-1} - \theta^m m_{it-1}) | \, I_{it-1}] = 0$$

which follows from $\mathbb{E}[\xi_{it} | \, I_{it-1}] = 0$ and $\mathbb{E}[e_{it} | \, I_{it-1}] = 0$. We can then 'plug in' the previously estimated $\hat{\Phi}_{t-1}$ and $\hat{\theta}^l$, leaving us with only θ^k and θ^m to estimate. Using this moment condition as the identifying restriction in a GMM procedure, and including the relevant instruments, we will get an unbiased estimate of $\hat{\theta}^m$.

Appendix B — De Ridder et al. (2021) Derivation of Output-Based IV-GMM Estimator

Under the assumption that productivity is I.I.D, we will satisfy the moment condition:

$$\mathbb{E}[\hat{w}_{it} v_{it-1}] = 0.$$

Next, our estimated production function is:

$$y_{it} = \hat{\theta}^v v_{it} + \hat{w}_{it}$$

Rearranging the production function for $\hat{w}_{it}$ gives:

$$\hat{w}_{it} = y_{it} - \hat{\theta}^v v_{it}$$

We can then substitute in the production function for y_{it}:

$$\hat{w}_{it} = \theta^v v_{it} + w_{it} - \hat{\theta}^v v_{it} = (\theta^v - \hat{\theta}^v) v_{it} + w_{it}$$

This gives us an expression for $\hat{w}_{it}$ which we can put into the moment condition:

$$\mathbb{E}\left[\left((\theta^v - \hat{\theta}^v) v_{it} + w_{it}\right) v_{it-1}\right]$$

$$= \mathbb{E}\left[(\theta^v - \hat{\theta}^v) v_{it} v_{it-1} + w_{it} v_{it-1}\right]$$

$$= (\theta^v - \hat{\theta}^v) \mathbb{E}[v_{it} v_{it-1}] + \mathbb{E}[w_{it} v_{it-1}] = 0$$

This yields the IV-GMM estimator referred to in the main text. We can then further simplify:

$$(\theta^v - \hat{\theta}^v) \mathbb{E}[v_{it} v_{it-1}] = -\mathbb{E}[w_{it} v_{it-1}]$$

$$\theta^v - \hat{\theta}^v = -\frac{\mathbb{E}[w_{it} v_{it-1}]}{\mathbb{E}[v_{it} v_{it-1}]}$$

$$\hat{\theta}^v = \theta^v + \frac{\mathbb{E}[w_{it} v_{it-1}]}{\mathbb{E}[v_{it} v_{it-1}]}$$

Appendix C — Levels of Simulated Markups

We show here that the true level of markups in our simulation is given by:

$$\mu_{it} = 1 + \Omega_{it}$$

Recall that our marginal revenue function is:

$$MR_{it} = T\Omega_{it} Q_{it}^{-\frac{1}{\eta_{it}}} - \frac{1}{\eta_{it}} T\Omega_{it} Q_{it}^{-\frac{1}{\eta_{it}}-1} Q_{it}$$

This is equivalent to:

$$MR_{it} = (1 - \frac{1}{\eta_{it}}) T\Omega_{it} Q_{it}^{-\frac{1}{\eta_{it}}}$$

We can then substitute our definition of price into this expression. Price is given by:

$$P_{it} = T\Omega_{it} Q_{it}^{-\frac{1}{\eta_{it}}}, \quad \eta_{it} = 1 + \frac{1}{\Omega_{it}}$$

Hence, marginal revenue can be written as:

$$MR_{it} = \left(1 - \frac{1}{\eta_{it}}\right) P_{it}$$

Substituting in our definition of η_{it}:

$$MR_{it} = \left(1 - \frac{1}{1 + \frac{1}{\Omega_{it}}}\right) P_{it} = \left(1 - \frac{\Omega_{it}}{1 + \Omega_{it}}\right) P_{it} = \left(\frac{1}{1 + \Omega_{it}}\right) P_{it}$$

Under profit maximization, marginal revenue is equal to marginal cost. Hence, in our solved equilibrium model, we have:

$$MC_{it} = \left(\frac{1}{1 + \Omega_{it}}\right) P_{it}$$

And so:

$$\frac{MC_{it}}{P_{it}} = \frac{1}{1 + \Omega_{it}}$$

And finally:

$$\frac{P_{it}}{MC_{it}} = \mu_{it} = 1 + \Omega_{it}$$

The average markup in our simulation is therefore directly determined by the average level of market power.

Appendix D — Numerical Solver

We start by guessing two values for Q_{it} such that $f(Q_{it})$ takes opposite signs at these points. The value at which the function is negative is denoted Q_{it}^{min}, and the value at which it is positive is Q_{it}^{max}. We then take the average of these values:

$$Q_{it}^{mid} = \frac{Q_{it}^{max} + Q_{it}^{min}}{2}$$

We then evaluate the FOC at Q_{it}^{mid}, and narrow down our next pair of guessed values based on the result. If $f\left(Q_{it}^{mid}\right) > 0$, we replace Q_{it}^{max} with Q_{it}^{mid} in the next iteration. And if $f\left(Q_{it}^{mid}\right) < 0$, we replace Q_{it}^{min} with Q_{it}^{mid}. Over many iterations of this process, the gap between the maximum and minimum guesses will become miniscule, and $f\left(Q_{it}^{mid}\right)$ will move arbitrarily close to 0. The value of Q_{it}^{mid} at that point will be taken as the firm's profit-maximizing output, and, after repeating this for each firm, we can use the sum of the solutions to complete the outer loop.

Appendix E — Derivation of Revenue Elasticity Estimate

We have the revenue function:

$$r_{it} = \frac{\omega_{it}}{\mu_{it}} + \frac{(1-\theta_t)}{\mu_{it}} k_{it} + \frac{\theta_t}{\mu_{it}} m_{it} + t + \ln(\Omega_{it}) + \frac{e_{it}}{\mu_{it}}$$

Which we shall write more compactly as:

$$r_{it} = a_{it} m_{it} + b_{it} k_{it} + d_{it} + g_{it}$$

where $a_{it} = \frac{\theta_t}{\mu_{it}}$, $b_{it} = \frac{(1-\theta_t)}{\mu_{it}}$, $d_{it} = \ln(\Omega_{it}) + \frac{e_{it}}{\mu_{it}}$, and $g_{it} = t + \frac{\omega_{it}}{\mu_{it}}$. By the Frisch-Waugh-Lovell theorem,[29] the estimate of a single variable in a multivariate regression of dependent variable y_{it} on independent variable x_{it} and controls can be expressed as:

$$\hat{\beta}_n = \frac{\sum(x_{it,\perp} * y_{it,\perp})}{\sum\left(x_{it,\perp}\right)^2}$$

where $x_{it,\perp}$ denotes the residuals of the regression of x_{it} on the other independent variables, and $y_{it,\perp}$ the residuals from the regression of y_{it} on all independent variables other than x_{it}. We can therefore express the estimated revenue elasticity as:

$$plim\ \hat{\theta}_t = \frac{E[m_{it,\perp} * r_{it,\perp}]}{E[m_{it,\perp}^2]}$$

[29] See Basu (2023) for an overview.

Where $\perp$ denote the residuals after regressing on our controls for capital and productivity (k_{it} and ω_{it}). Given the residualized revenue function:

$$r_{it,\perp} = a_{it}m_{it,\perp} + b_{it}k_{it,\perp} + d_{it,\perp} + g_{it,\perp}$$

We can rewrite the estimate as:

$$plim\ \hat{\theta}_t = \frac{E[m_{it,\perp}(a_{it}m_{it,\perp} + b_{it}k_{it,\perp} + d_{it,\perp})]}{E[m_{it,\perp}^2]} = \frac{E\left[a_{it}m_{it,\perp}^2\right] + E\left[m_{it,\perp}b_{it}k_{it,\perp}\right] + E\left[m_{it,\perp}d_{it,\perp}\right] + E\left[m_{it,\perp}g_{it,\perp}\right]}{E[m_{it,\perp}^2]}$$

As t is a constant and we have included controls for capital and productivity $E\left[m_{it,\perp}b_{it}k_{it,\perp}\right] = 0$ and $E\left[m_{it,\perp}g_{it,\perp}\right] = 0$, hence:

$$plim\ \hat{\theta}_t = \frac{E\left[a_{it}m_{it,\perp}^2\right] + E[m_{it,\perp}d_{it,\perp}]}{E[m_{it,\perp}^2]} = \frac{E\left[a_{it}m_{it,\perp}^2\right]}{E[m_{it,\perp}^2]} + \frac{E\left[m_{it,\perp}d_{it,\perp}\right]}{E[m_{it,\perp}^2]}$$

By the definition of covariance, we have:

$$cov\left(a_{it}, m_{it,\perp}^2\right) = E\left[a_{it}m_{it,\perp}^2\right] - E[a_{it}]E\left[m_{it,\perp}^2\right]$$

which we can rearrange to make:

$$\frac{E\left[a_{it}m_{it,\perp}^2\right]}{E[m_{it,\perp}^2]} = E[a_{it}] + \frac{cov\left(a_{it}, m_{it,\perp}^2\right)}{E\left[m_{it,\perp}^2\right]} = E[a_{it}] + \frac{cov\left(a_{it}, m_{it,\perp}^2\right)}{var(m_{it,\perp})}$$

with $E\left[m_{it,\perp}^2\right] = var(m_{it,\perp})$ following from the fact that the residual terms are mean-zero by construction so long as we do not omit constants from our regression. Finally, then, we can express the revenue elasticity estimate as:

$$plim\ \hat{\theta}_t = E[a_{it}] + \frac{cov\left(a_{it}, m_{it,\perp}^2\right)}{var(m_{it,\perp})} + \frac{cov\left(m_{it,\perp}, d_{it,\perp}\right)}{var(m_{it,\perp})}$$

Or, in terms of the true elasticity and markup:

$$plim\ \hat{\theta}_t = E\left[\frac{\theta_t}{\mu_{it}}\right] + \frac{cov\left(\frac{\theta_t}{\mu_{it}}, m_{it,\perp}^2\right)}{var(m_{it,\perp})} + \frac{cov\left(m_{it,\perp}, d_{it,\perp}\right)}{var(m_{it,\perp})}$$

We can break this expression down as follows: Because θ_t is constant within years, $cov\left(\frac{\theta_t}{\mu_{it}}, m_{it,\perp}^2\right)$ will be determined by $cov\left(\frac{1}{\mu_{it}}, m_{it,\perp}^2\right)$. The relationship between this covariance and trends in market power will depend on the structure of demand. In our model, $cov\left(\frac{1}{\mu_{it}}, m_{it,\perp}^2\right)$ increases with market power[30]. At low levels of market power, the covariance is negative, while at high levels, it is positive.

The movement of second covariance term, $cov\left(m_{it,\perp}, d_{it,\perp}\right)$, is less clear; $\ln(\Omega_{it})$ is obviously increasing in market power, whereas $\frac{e_{it}}{\mu_{it}}$ is decreasing. In general, $cov\left(m_{it,\perp}, d_{it,\perp}\right)$ will initially be very large, before quickly converging

[30] Intuitively, this arises from a negative relationship between markups and intermediate usage. When market power rises, monopolists decrease output and increase price. When average market power is low, this cannot be observed, but when it is high, we will see that $cov\left(\mu_{it}, m_{it,\perp}^2\right)$ is negative, and conversely, that $cov\left(\frac{1}{\mu_{it}}, m_{it,\perp}^2\right)$ is positive.

to 0 as market power increases. The sum, $cov\left(\frac{\theta_t}{\mu_{it}}, m_{it,\perp}^2\right) + cov(m_{it,\perp}, d_{it,\perp})$ will therefore start at a high level, decrease until $cov(m_{it,\perp}, d_{it,\perp}) = 0$ and then increase steadily.

Finally, $E\left[\frac{\theta_t}{\mu_{it}}\right]$ strictly decreases with markups. The expression is thus a nonlinear function of market power with multiple turning points.

Appendix F — Industry Averages of Elasticities, Expenditure on Observed Variable Input, Revenue Shares, and Markups

Table F1 – Industry Markups: Single Input Specification

Industry	Description	Elasticity	Expenditure	Revenue Share	Markup
AA11	Horticulture and fruit growing	0.43	167582	23.9	10.2
AA12	Sheep, Beef cattle and Grain Farming	0.37	107475	22.4	8.2
AA13	Dairy Cattle Farming	0.22	93297	70.6	15.6
AA14	Poultry, deer, and other livestock farming	0.44	172240	21.5	9.6
AA21	Forestry and logging	0.48	409483	29.0	13.7
AA31	Fishing and aquaculture	0.39	208646	31.5	12.2
AA32	Agric, forest, fish support services, and hunting	0.34	212042	33.3	11.3
BB11	Mining	0.41	1307739	37.7	14.9
CC1	Food & beverage manufacturing	0.51	2273916	5.3	2.7
CC21	Textile, leather, cloth & footwear manufacturing	0.40	748921	12.7	5.2
CC3	Wood & paper product manufacturing	0.43	1179139	8.4	3.7
CC41	Printing	0.46	474655	6.8	3.1
CC5	Petrochemical product manufacturing	0.49	2063288	4.9	2.4
CC61	Non-metallic mineral product manufacturing	0.46	1189080	7.5	3.5
CC7	Metal & metal product manufacturing	0.41	1047739	12.8	5.4
CC81	Transport equipment manufacturing	0.45	634874	10.0	4.6
CC82	Machinery & other equipment manufacturing	0.40	731755	12.7	5.2
CC91	Furniture & other manufacturing	0.45	344346	7.1	3.2
DD1	Electricity, gas & water	0.34	738380	24.8	7.7
EE11	Building construction	0.47	502615	14.6	7.1
EE12	Heavy & civil engineering construction	0.37	1056130	17.5	6.4
EE13	Construction services	0.39	231243	12.9	5.0
FF11	Wholesale trade	0.34	2255809	4.9	1.7
GH11	Motor vehicle/parts & fuel retailing	0.27	3829341	2.6	0.7
GH12	Supermarket, grocery & spec. food retailing	0.26	2062964	1.6	0.4
GH13	Other store-based & non-store retailing	0.37	680639	4.0	1.5
GH21	Accommodation & food serv.	0.43	218527	6.8	2.9
II11	Road transport	0.36	469081	30.5	10.6
II12	Rail, water, air & other transport	0.37	536204	30.8	11.3
II13	Post, courier support & warehouse serv.	0.44	1449321	31.5	13.7
JJ11	Information media serv.	0.38	354349	27.1	10.3

JJ12	Telecom., internet & library serv.	0.34	745132	17.9	5.9
KK13	Auxiliary finance & insurance serv.	0.41	882666	38.7	15.4
KK1_	Finance, insurance & real estate	0.44	1250447	24.8	10.6
LL11	Rental & hiring serv.	0.35	298603	19.9	7.1
MN11	Professional, scientific & tech. serv.	0.46	272449	33.8	16.5
MN21	Administrative & support serv.	0.35	407773	28.8	10.1
RS11	Arts & recreation serv.	0.32	81055	26.1	8.4
RS21	Other serv.	0.32	156310	10.1	3.3

Table F1: Average estimated elasticities, expenditures on defined variable input, revenue shares, and markups for all industries in the productivity dataset under the single input specification.

Table F2 – Industry Markups: Full Bundle Specification

Industry	Description	Elasticity	Expenditure	Revenue Share	Markup
AA11	Horticulture and fruit growing	1.13	506675	1.11	1.25
AA12	Sheep, Beef cattle and Grain Farming	1.11	359617	1.11	1.19
AA13	Dairy Cattle Farming	0.95	753500	1.27	1.16
AA14	Poultry, deer, and other livestock farming	1.15	452163	1.06	1.20
AA21	Forestry and logging	1.23	828315	1.19	1.49
AA31	Fishing and aquaculture	0.97	524049	1.46	1.42
AA32	Agric, forest, fish support services, and hunting	0.96	559729	1.27	1.20
BB11	Mining	0.98	3468272	1.26	1.24
CC1	Food & beverage manufacturing	0.89	3531340	1.12	0.98
CC21	Textile, leather, cloth & footwear manufacturing	0.86	1221661	1.24	1.07
CC3	Wood & paper product manufacturing	0.88	1945875	1.20	1.06
CC41	Printing	0.87	1049799	1.10	0.94
CC5	Petrochemical product manufacturing	0.91	3398720	1.01	0.91
CC61	Non-metallic mineral product manufacturing	0.90	2108193	1.16	1.04
CC7	Metal & metal product manufacturing	0.84	1788218	1.23	1.02
CC81	Transport equipment manufacturing	0.82	1115417	1.26	1.04
CC82	Machinery & other equipment manufacturing	0.83	1327943	1.23	1.01
CC91	Furniture & other manufacturing	0.88	657351	1.24	1.08
DD1	Electricity, gas & water	0.83	1793690	1.20	0.99
EE11	Building construction	0.87	827309	1.39	1.21
EE12	Heavy & civil engineering construction	0.90	2206300	1.29	1.16
EE13	Construction services	0.83	498440	1.39	1.14
FF11	Wholesale trade	0.65	2934210	0.54	0.35
GH11	Motor vehicle/parts & fuel retailing	0.44	4496361	0.44	0.19
GH12	Supermarket, grocery & spec. food retailing	0.50	2526713	0.37	0.18
GH13	Other store-based & non-store retailing	0.75	1048359	0.59	0.45
GH21	Accommodation & food serv.	0.86	560398	0.97	0.84
II11	Road transport	0.87	1223885	1.37	1.21
II12	Rail, water, air & other transport	1.00	1211555	1.05	1.05
II13	Post, courier support & warehouse serv.	0.83	1922151	1.33	1.08
JJ11	Information media serv.	0.91	702128	1.22	1.11

JJ12	Telecom., internet & library serv.	0.89	1520661	0.99	0.88
KK13	Auxiliary finance & insurance serv.	0.89	1413165	1.18	1.05
KK1_	Finance, insurance & real estate	0.95	2272632	1.21	1.11
LL11	Rental & hiring serv.	0.92	840559	1.03	0.94
MN11	Professional, scientific & tech. serv.	0.77	634177	1.29	0.99
MN21	Administrative & support serv.	0.80	660900	1.31	1.05
RS11	Arts & recreation serv.	0.90	348632	1.21	1.09
RS21	Other serv.	0.75	358774	1.14	0.85

Table F2: Average estimated elasticities, expenditures on defined variable input, revenue shares, and markups for all industries in the productivity dataset under the full bundle specification.