THE UNIVERSITY OF CHICAGO

LABOR MARKET POWER: THEORY AND EVIDENCE

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE SOCIAL SCIENCES
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

KENNETH C. GRIFFIN DEPARTMENT OF ECONOMICS

BY
IVAN KIROV

CHICAGO, ILLINOIS
DECEMBER 2021

Everything I have managed to do with my short time at the University is due entirely to the support and encouragement of my family: Rebecca, Alex, and Leo. You are my world.

Table of Contents

List of Figures

List of Tables

ACKNOWLEDGMENTS

These papers are co-written with James Traina and a part of a long-running research collaboration with him. I thank Erik Hurst, Chad Syverson, Ali Hortaçsu, Brent Neiman, and Luigi Zingales for their comments and support. I also thank Pablo Ottonello and Tom Winberry for their blessing in working with the data, and Zach Flynn for helpful econometric discussions.

I would not be where I am today without my first-year study group (Joshua Shea & Yulia Zhestkova). I also want to thank all participants in Rob Shimer and Joe Vavra's second-year seminar in 2017-2018: your input was valuable in shaping this paper. Thanks also to participants in the UChicago IO Lunch, Capital Theory, and Applied Macro Theory working groups.

Finally, thanks to my parents, my university professors, my graduate instructors, and all of my teachers throughout the years. This is my 22nd year of education, and all of you have helped shape me and my scholarship.

Corresponding email address: ikirov@uchicago.edu

ABSTRACT

This thesis investigates the theoretical definition, econometric estimation, and empirical realities of labor-market power. Its main chapter focuses on US manufacturing, finding that manufacturing workers received $1.10 cents of each marginal unit of production in 1973 but only 48 cents in 2014. I find that labor-market power is strongly correlated with adoption of ICT and automation technology. My estimates imply that labor-market power is significantly more important than markup power in US manufacturing. In the second chapter, I show that when current production-based ”ratio estimators” of market power return values different from 1, they imply either model misspecification or input frictions (such as labor market power or labor adjustment costs). I argue that the latter interpretation is more plausible in practice, and that researchers and policymakers should largely reinterpret existing work using these estimators as measuring input frictions. Finally, in the third chapter I propose an estimation method that produces unbiased and consistent estimates of labor-market and markup power by flexibly modeling markups as a specified function of observables and fixed effects. My modified two-step estimator is simple in concept and implementation, requiring less onerous assumptions than popular proxy production estimators.

In sum, this thesis argues that labor-market power is an important feature of the US labor market, and provides theoretical support and estimation techniques to study it further.

Chapter 1

RISING LABOR MARKET POWER AND CHANGING TECHNOLOGY IN US MANUFACTURING

1.1 Introduction

Understanding the size and origins of market power is among the most critical tasks of economics. As Adam Smith put it, "people of the same trade seldom meet together, even for merriment and diversion, but the conversation ends in a conspiracy against the public, or in some contrivance to raise prices." In modern terms, market power leads to inefficiencies and misallocation. The specific nature of these distortions depends on the source of market power. In the output market, markup power increases prices and reduces output. In the labor market, labor-market power (or markdown power[1]) reduces wages and employment. In the past 20 years, real wages in the US have stagnated as productivity has grown. Inflation has remained low despite steady growth and low unemployment rates. At the same time, the labor share of income in the US fell from 64 percent in the mid-1980s to 58 percent today (Karabarbounis and Neiman (2013)). In manufacturing, the labor share declined from 62 percent in 1967 to 41 percent in 2012 (Kehrig and Vincent (2021)). These trends suggest that monopsony power may be significant in US labor markets: if the structure of labor markets has changed so that firms do not pass productivity gains onto workers in the form of higher wages, the link between unemployment, output growth, and inflation may be weakened.

How important is this mechanism? How much has labor-market power increased from 1972-2014? What are the sources and implications of this increase? These are the questions we seek to answer in this paper. We focus on US manufacturing sector for two reasons: first,

1. Also called monopsony power after Robinson (1933).

outputs and inputs are more readily measurable in manufacturing, and we have access to a near-universal dataset of US manufacturing plants. Second, increasing automation and changing technology has particularly affected the manufacturing sector (as evidenced by the larger decline in the manufacturing labor share). To estimate labor-market wedges, we extend methods in production-function estimation to overcome identification challenges that made studying market power in input markets difficult. Our method allows us to identify market power wedges - ratios of marginal benefit to marginal cost - for both output and labor inputs. We apply our method to manufacturing microdata from the US Census to recover labor-market wedges at the establishment-year level.

We find that labor-market power in manufacturing is high and has risen since the mid-1970s: the average manufacturing production worker was paid 110% of their marginal revenue product in 1973 but only 48% in 2014. By contrast, white-collar managers have always been paid their marginal revenue products[2]. To make our definition of labor-market power more precise, we define a 'labor-market wedge' as the ratio of the marginal revenue product of labor to the wage $\left(\frac{MRPL}{W}\right)$ - analagous to the markup wedge $\left(\frac{P}{MC}\right)$. We find that the aggregate manufacturing labor-market wedge was approximately unity (implying no labor-market power) from 1973 to 1990. Then the wedge rose rapidly, with inflection points in 1992 and 2000. Overall, our estimated labor-market wedge rose from 0.9 in 1973 to 2.1 in 2014.

Moreover, we find that changing technology is an important source of labor-market power. In particular, automation and information-technology (ICT) adoption increases the labor-market wedge at the plant level. Higher labor-market wedges are also associated with more managerial employees and a deeper capital stock. By contrast, deunionization is uncorrelated with levels of or changes in labor-market wedges. At the plant level, we find that labor-market power reduces employment more than wages. Labor-market wedges are also strongly

2. Here 'managers' are those defined by the Census Bureau as 'nonproduction workers.'

negatively correlated with establishment-level labor shares: rising labor-market power can account for about two-thirds of the decline in the manufacturing labor share in the US.

This paper links literature on automation and technological change with the macroeconomic market-power literature. Autor and Dorn (2013) find that technological change polarizes labor markets by reducing the demand for intermediate-skilled, routine jobs. Acemoglu and Restrepo (2020) show that robot penetration reduces wages and employment in local labor markets, while Acemoglu and Restrepo (2021) find that automation technology displaces routine workers. Meanwhile, a body of literature at the intersection of industrial organization and macroeconomics focuses on the evolution of markups in the US economy. Broadly speaking, this literature argues markups are rising and driving down labor's share of income as firms with monopolistic power decrease output, increase prices, and reduce wages (Barkai (2020), De Loecker et al. (2020), and Hall (2018)). Other work calibrates macroeconomic models to these estimates to explore their welfare implications (Edmond et al. (2021), Eggertsson et al. (2018), and Aghion et al. (2020)). Autor et al. (2020) emphasizes the role of right-productivity-tail 'superstar' firms in driving technological change, suggesting that higher markups are a reward for innovation. This paper instead emphasizes that automation and technological change alter the structure of labor markets, reducing workers' bargaining power even if they are not employed in routine tasks. We connect the automation and market-power literatures by focusing on the impact of technology on monopsony wedges, and by finding direct evidence of the connection between markdowns and technology-driven automation.

We find that the markdown wedge is significantly larger than the markup wedge, and has grown much faster. Inspired by the decline of the labor share (Elsby et al. (2013) and Karabarbounis and Neiman (2013); Kehrig and Vincent (2021) documents the underlying establishment-level evolution) and by rising concentration[3] (Barkai (2020) and Autor et

3. Concentration within industries has also risen in recent decades: the sales share of the largest four firms

al. (2020)), and falling firm and labor dynamism (Decker et al. (2016)), the market-power literature has focused on the markup (output-market) wedge (De Loecker et al. (2020) and Hall (2018)). But as Syverson (2019) argues, markups are prices over costs ($\mu = \frac{P}{MC}$), so growth in markups is approximately inflation less cost growth. Inflation has been unusually low and cost growth has been steady, suggesting low markup growth[4]. By contrast, labor-market power is labor productivity over wages $\left(\frac{MRPL}{W}\right)$, so growth in labor-market power is approximately labor productivity growth less wage growth. Labor productivity growth has been steady while wage growth has been stagnant. These back-of-the-envelope approximations imply that rising labor-market power may be a better candidate to explain the decline of the labor share. Rather than raising profits by increasing output prices, perhaps firms pay workers less than their marginal value due to inelastic residual labor-supply curves[5]. Indeed, we find that in labor-market wedges are more important than markups in an aggregate 'composite wedge' (markups $\times$ labor-market wedges).

We also contribute to the literature on monopsony power (Lamadon et al. (2020), Berger et al. (2019), Azar et al. (2020), and Benmelech et al. (2018); see Card et al. (2018) for an overview). We identify and structural estimate labor-market wedges using production microdata, providing an alternative method to the concentration proxies and natural experiments in the monopsony literature. Concentration is positively correlated with market power in Cournot oligopoly, but is negatively correlated or uncorrelated with market power in other conduct models; indeed, we offer evidence that concentration is uncorrelated with monopsony power on average. Our novel estimation methodology takes into account the identification challenges of mainstream prior approaches, thereby improving on Morlacco

in an average industry increased from 26 percent in 1997 to 33 percent in 2012. Rossi-Hansberg et al. (2021) and Rinz (2018) show that concentration actually falls in local markets, suggesting that the rise in aggregate concentration is driven by the increasing expansion of national firms.

4. See Traina (2018), Edmond et al. (2021), and Karabarbounis and Neiman (2019), and Flynn et al. (2019) for evidence against rising markups.

5. Arising from local labor-market imperfections, search frictions, or other labor-market rigidities.

(2020) and Hershbein et al. (n.d.). We estimate static market-power wedges, putting our paper in the 'new classical' tradition of Manning (2021).

Finally, our paper contributes to the econometric methods literature on measuring market power. Dobbelaere and Mairesse (2013) and De Loecker et al. (2020) use insights from Hall (1988) to show how researchers can use cost minimization conditions with estimated input elasticities to recover market power wedges. However, Gandhi et al. (2020) shows that the production estimators of Olley and Pakes (1996), Levinsohn and Petrin (2003), and Wooldridge (2009), and Ackerberg et al. (2015) do not identify the flexible input elasticities that are necessary to estimate market power. In the context of imperfect competition, Flynn et al. (2019) shows that ignoring the identification problem typically results in higher market power estimates in both level and dispersion. Moreover, Bond et al. (2020) show that using revenue data to estimate output elasticities results in inconsistent estimates of market-power wedges: in particular, it implies that markups should be unity in the absence of specification error. We overcome the revenue-estimation concerns of Bond et al. (2020) and the nonidentification challenges of Gandhi et al. (2020) by proposing a new estimation technique which combines a parametric materials demand function with a proxy production structure. Our method is consistent in the presence of revenue data. It allows for estimation of an arbitrary number of flexible input elasticities, and consequently an arbitrary number of market power wedges. By separating production and nonproduction workers, we address the critique in Bond et al. (2020) that some employees may impact demand.

The paper proceeds as follows. Section 1.2 outlines our theoretical framework for identifying labor-market power in production data. Section 1.3 explains our estimation methodology. Section 1.4 outlines our Census manufacturing microdata. We present our empirical results in Section 1.5. Section 1.6 concludes.

1.2 A Model of Labor Market Power

To analyze labor market wedges, we first outline a micro-founded model of firm behavior and an estimation strategy to recover the key parameters of this model. Firms in our model choose their labor inputs based on their residual labor supply curves: they are local monopsonists. These firms internalize the effect of increasing labor inputs on their inframarginal labor expenditures, and therefore reduce hiring and wages relative to a competitive benchmark. On the other hand, firms buy intermediate inputs in a competitive market. If the output market were also competitive, firms would equate average and marginal products of intermediates. If a firm uses fewer intermediates than the competitive benchmark, we infer a markup wedge. Such a wedge reduces utilization of all inputs symmetrically. Labor market wedges, however, reduce labor utilization only. Comparing the ratio of average to marginal products of labor and intermediates allows us to recover these monopsony wedges. This double-ratio estimator is in the tradition of Hall (2018), De Loecker et al. (2020), and Dobbelaere and Mairesse (2013).

To implement our ratio estimator, we must specify and estimate a production function. We develop a new approach to estimating production functions in the context of revenue data (the details are in our companion paper Kirov and Traina (2021a)). Our approach combines a two-stage regression in the spirit of Gandhi et al. (2020) with a proxy variable approach to control for unobserved markups. In this latter sense the model is similar to Olley and Pakes (1996), Levinsohn and Petrin (2003), Wooldridge (2009), and Ackerberg et al. (2015), but proxies for markups directly rather than for productivity. We use a translog production function to allow for significant flexibility in the mapping between input use and productivity and in the nature of technological progress. Through an informational assumption on revenue productivity, our model corrects for unobserved prices. Thus it ameliorates the revenue issue described in Bond et al. (2020) and Klette and Griliches (1996): we recover physical output elasticities even in the presence of revenue data.

Our approach imposes minimal structure on supply and is agnostic about demand. It can therefore recover labor and product market power in many settings.

1.2.1 Conceptual Framework: Market Power Wedges from Cost Minimization

The supply side of the economy consists of a mass of firms which use labor L, intermediates M, capital K, and nonproduction labor N to produce output Q. Labor is supplied by households according to firm-specific residual labor supply curves. Firms must pay a single wage W for undifferentiated labor input L. Intermediates are inelastically supplied in perfectly competitive markets at price C. Capital takes one period to build, so firms decide their capital stocks at time $t-1$. Nonproduction labor stands in for management, and is also decided on at $t-1$[6]. Markets in both K and N are competitive, with respective equilibrium prices R and S. Each firm's production function is given by $F(M_t, L_t, N_t, K_t)$[7]. Firms sell output by choosing price and quantity $\{P, Q\}$ on their residual demand curves.

Consider the cost-minimization problem of an individual firm. Imperfect competition in the labor market implies that wages are a function of labor input choice:

$$\min_{M_t, L_t, K_t} C_t M_t + W_t(L_t)L_t + S_t N_{t-1} + R_t K_{t-1} \quad (1.1)$$
$$s.t. \quad F(M_t, L_t, N_t, K_t) = Q_t$$

The Envelope Theorem implies that the Lagrange multiplier of (1.1) is $\lambda_t = \frac{\partial \mathcal{L}}{\partial Q_t} = MC_t$.

6. We make this abstraction because we view managers as a form of capital stock: they take time to train and deploy.

7. K_t and N_t are chosen in $t-1$ due to the time-to-build assumption. In estimation, we allow $F(\cdot)$ to vary across estimation groups.

The firm's first-order conditions for cost minimization in (1.1) are:

$$[M_t] \quad C_t - \lambda_t \frac{\partial F(\cdot)}{\partial M_t} = 0 \Rightarrow \frac{\partial F(\cdot)}{\partial M_t} = \frac{1}{\lambda_t C_t} \tag{1.2}$$

$$[L_t] \quad W_t + \frac{\partial W_t}{\partial L_t} L_t - \lambda_t \frac{\partial F(\cdot)}{\partial L_t} = 0 \tag{1.3}$$

Using the definition of λ_t as marginal cost and multiplying the intermediates FOC by $P_t \frac{M_t}{Q_t}$ yields:

$$\theta^M P_t = \mu_t \frac{C_t M_t}{Q_t} \quad \Rightarrow \quad \mu_t = \theta^M \left(\frac{P_t Q_t}{C_t M_t} \right) \tag{1.4}$$

Where $\theta^M \equiv \frac{\partial F(\cdot)}{\partial M_t} \frac{M}{Q}$ is the output elasticity of materials; and we define the markup as the ratio of price to marignal cost: $\mu_t \equiv \frac{P}{MC}$. (1.4) is the ratio estimator popularized by Hall (1988) and De Loecker and Warzynski (2012). It expresses the markup as a wedge between a firm's marginal value of intermediates θ^M and its average spending on intermediates as a proportion of revenue $\frac{C_t M_t}{P_t Q_t}$.

Now turn to the labor FOC. Substituting μ_t into (1.3), dividing by W_t, and simplifying yields:

$$\mu_t \left(1 + \frac{\partial W_t}{\partial L_t} \frac{L_t}{W_t} \right) = \frac{P_t}{W_t} \frac{Q_t}{L_t} \frac{L_t}{Q_t} \frac{\partial F(\cdot)}{\partial L_t} \Rightarrow \mu_t (1 + \nu_t) = \left(\frac{P_t Q_t}{W_t L_t} \right) \theta^L \tag{1.5}$$

Where $\nu_t \equiv \frac{\partial W_t}{\partial L_t} \frac{L_t}{W_t} = \left(\epsilon^{LS} \right)^{-1}$ is the elasticity of the firm's wage with respect to labor input (equivalently, the inverse of the firm's residual labor supply elasticity) and θ_L is the output elasticity of labor. It is important to emphasize that $\left(\epsilon^{LS} \right)^{-1}$ is the inverse elasticity of the firm's conduct-adjusted residual labor supply, in the sense that the firm-level residual labor supply curve also incorporates competitors' conduct. This is distinct from the macroeconomic inverse labor-supply elasticity, which is a household object and does not consider firm conduct. We define the labor market wedge as $\delta_t \equiv (1 + \nu_t)$. From (1.5), this wedge can be written as:

$$\delta_t \equiv (\nu_t + 1) = \left(\frac{P_tQ_t}{W_tL_t}\right)\frac{\theta^L}{\mu_t} = \left(\frac{P_tQ_t}{W_tL_t}\right)\frac{C_tM_t}{P_tQ_t}\cdot\frac{\theta^L}{\theta^M}$$
$$= \left(\frac{C_tM_t}{W_tL_t}\right)\frac{\theta^L}{\theta^M} \tag{1.6}$$

We infer a larger labor market wedge when a firm spends more on intermediates than labor relative to their respective values in production. Equations (1.5) and (1.6) give some intuition on the nature of labor market power. In the absence of any market power, $\mu_t = 1$ and $\nu_t = 0$ (that is, the firm's labor supply curve is infinitely elastic), which implies that $\delta_t = 1$. Then (1.5) implies that the firm hires until its labor expenditure share equals labor's value in production θ_L. Market power in the output market adds a wedge μ_t between these objects. If the firm also has labor market power ($\nu_t > 0$), it faces a tradeoff in its wage decision: paying higher wages attracts more workers but requires paying inframarginal workers more. This tradeoff adds an additional wedge $(1+\nu_t)$ between labor expenditure and θ_L. Hence the firm uses less labor than a firm hiring in a competitive labor market. Note that the markup term μ_t is endogenous[8], in that a firm with more or less output market power will choose a different $\{P, Q\}$ combination, and therefore a different $\{W, L\}$ combination. In other words, higher markups do not mechanically imply lower markdowns.

Equation (1.6) makes this point explicitly, defining labor market wedges as the ratio of markups from labor $\left(\theta_L \cdot \frac{P_tQ_t}{W_tL_t}\right)$ to markups from intermediates $\left(\theta_M \cdot \frac{P_tQ_t}{C_tM_t}\right)$. Intuitively, output market power induces the firm to reduce all of its inputs proportionally, while labor market power induces the firm to reduce only its labor input. Equation (1.6) backs out the labor market wedge from ratio of these two objects.

8. This firm chooses its price P_t to maximize profits subject to its output demand curve, because it has output market power. The markup term μ_t reflects the firm's optimal price choice.

Revenue and Quantity Elasticities

Our formulas for μ and δ require the θ^j to be physical output elasticities rather than revenue elasticities: that is, they must describe marginal increases in physical production Q rather than marginal increases in revenues $P \cdot Q$. Revenue elasticities conflate changes in demand (and therefore price) with changes in physical productivity. Bond et al. (2020) describe two problems with this approach: first, market power allows firms to change Q *specifically in order* to change P. Second, some inputs may determine demand (workers in advertising, for instance). We address the second point by only using production workers as our measure of L, and incorporating nonproduction workers as a separate productive input.

The first point implies that using revenue elasticities in the formulae for μ and δ generates inconsistent estimates of market power wedges (a point also emphasized im Klette and Griliches (1996)). In particular, the FOC (1.4) implies that $\mu \equiv 1$ with revenue elasticities. Our companion paper Kirov and Traina (2021b) shows that, if one can only estimate the revenue elasticity of labor inputs θ^L_R, then the appropriate expression for the labor-market wedge δ is $\theta^L_R \frac{PQ}{WL}$ - the expression for markups computed using the labor input L. However, we do not estimate revenue elasticities in this paper. Instead, in what follows we develop a method to estimate quantity elasticities from revenue data. The full details of this method are available in our companion paper Kirov and Traina (2021a).

1.2.2 Interpreting Market Power Wedges

If the labor market is perfectly competitive, firms face perfectly elastic labor supply and the inverse labor-supply elasticity $\nu_t = 0$, which implies $\delta = 1$. If marginal revenue products of labor are greater than wages, then $\delta > 1$. A higher δ implies a higher wedge between wages and marginal revenue products, and therefore greater labor market frictions[9] If the output market is perfectly competitive, firms sell at marginal cost and $\mu_t = 1$. Impefect competition

9. Such frictions could also be due to adjustment costs or other structural characteristics of labor markets.

raises prices above marginal costs and implies $\mu_t > 1$. Models with increasing returns to scale or dynamic monopsony may find $\mu_t < 1$ or $\delta_t < 1$. A wide number of monopsonistic or oligopsonistic environments would induce a labor market wedge in this model: for instance, labor search frictions or implicit contracting. In particular, static labor search frictions would be isomorphic to our labor-market wedges, though their magnitude would differ in dynamic models. Implicit contracting would skew labor-market wedges lower - perhaps even to less than unity - at the start of a contract, and higher near the end of a contract.

Labor-market wedges δ and markups μ also relate closely to the labor share and profit share. Using the definition of markdowns (1.6) in (1.5) and rearranging yields:

$$\frac{WL}{PQ} = \frac{\theta^L}{\mu\delta}$$

So that markups and markdowns both decrease the labor share. If the production function is Cobb-Douglas with labor exponent α, $\theta^L = \alpha$ and the labor share equals labor's marginal product in the absence of market power wedges. Similarly, an accounting identity for each firm implies that all revenues go to profits π or payments to factors. Inserting the definition of μ and δ yields:

$$WL + CM + RK + SN + \pi = PQ$$
$$\frac{WL}{PQ} + \frac{CM}{PQ} + \frac{RK}{PQ} + \frac{SN}{PQ} + \frac{\pi}{PQ} = 1$$
$$\frac{\theta^L}{\mu\delta} + \frac{\theta^M}{\mu} + \frac{\theta^N}{\mu} + \frac{\theta^K}{\mu} + \frac{\pi}{PQ} = 1$$

An increase in markups decreases the share of revenues paid to all factors, while an increase in labor-market wedges decreases only the share paid to labor. In either case, the share of profits in revenues rises, since the RHS is unchanged.

1.3 Estimating Labor Market Power

In this section, we discuss estimation of our labor-market power parameter δ and our market-power wedge μ. We propose a new method for estimating production functions with revenue data, and then describe our estimation procedure.

1.3.1 Estimating Output Elasticities

To estimate markdown and markup wedges, we require output elasticites and so must estimate firms' production functions. This is made more difficult by the fact that we only have data on firm revenues PQ rather than output quantities Q. This section outlines our estimation strategy. We propose a new method of estimating production functions with revenue data which combines a two-step strategy akin to Gandhi et al. (2020) with a proxy assumption on markups in the spirit of Olley and Pakes (1996). This section outlines the model; the full details are in our companion paper Kirov and Traina (2021a).

As in the model in the previous section, firms produce output using intermediates M, production labor L, nonproduction labor N, and capital K. Intermediates can be thought of as inputs which the firm transforms into output: steel, partially finished goods, and other such inputs (e.g., inputs into value added). Production labor actively works to create output, while nonproduction labor can be thought of as management or administrative staff. Capital consists of machines and buildings used in production.

Inputs create output through a production function with Hicks-neutral productivity e^{ω_t}. Productivity is a log-additive combination of forseeable[10] and idiosyncratic terms, so that $\omega_t \equiv a_t + \epsilon_t$ with $E[\epsilon_t|\mathcal{I}_t] = 0$. In other words, firms cannot forsee the productivity shock ϵ, but expect it to be zero. The log production function of a firm is (denoting logs in

10. The term a_t is observable to the firm (but not to the econometrician) before choosing inputs, while ϵ_t is unknown to the firm at t.

lower-case):

$$q_t = f(m_t, \ell_t, n_t, k_t) + a_t + \epsilon_t \tag{1.7}$$

As in the previous section, each firm decides on capital and managerial stocks prior to production. By contrast, intermediate materials and labor are flexible (static) inputs: last periods' choice of these inputs does not affect this period's choice. This rules out, for instance, adjustment costs on labor. The timing of production is: (1) the firm chooses K_t and N_t to be used in production in t; (2) a_t is realized; (3) the firm chooses M_t and L_t; (4) ϵ_t is realized. The realization of ϵ, along with the firm's choice of inputs, determines markups and labor-market wedges. Since expectations about productivity partially determine inputs, market power wedges depend on both productivity and firm conduct.

Estimating (3.1) by OLS is subject to simultaneity bias: firms with higher productivity draws a_t will want to produce more and therefore use more inputs, so regressing output on inputs confounds output elasticities with realizations of the productivity process (Marschak and Andrews (1944)). Moreover, we do not observe q_t but only revenues $p_t \cdot q_t$. To estimate output elasticities, we need to both (i) separate productivity from input use; and (ii) separate demand-driven price variations from markup-driven price variations.

To do so, we combine the markup first-order condition (1.4) from the previous section with the production function (3.1). The first-order condition (1.4) gives us a parametric relationship between intermediates demand M_t, materials productivity, and market power. Conditional on market power, this pins down intermediates use in terms of parameters. This is important because intermediates suffer from simultaneity bias with productivity: M is flexible, and so an econometrician cannot easily separate movements in M from movements in A. But if we were able to control for markups, (1.4) allows us to control for intermediates. This, along with our Markovian timing assumptions on the productivity process, will allow us to separate TFP from input elasticties and prices from quantities.

Define revenues $R \equiv P \cdot Q$ and rewrite the FOC (1.4) in logs as (again using lower-case letters to denote logs) [11]:

$$\log(\mu) = \log(\theta^M) + r - cm + b - \epsilon$$
$$cm - r = \log(\theta^M) - \log(\mu) + b - \epsilon \tag{1.8}$$

Where the constant $b \equiv \frac{E[\exp(\epsilon)]}{\exp(\epsilon)}$. We call (1.8) the *share regression*: it defines elasticities and markups in terms of the observable share of intermediates expenditures to revenues. The share regression uses firm optimization to address the revenue problem: the LHS of (1.8) is in revenue terms, while the RHS is in quantity terms. The markup term μ connects the two. The output elasticity θ^M depends only on (quantity) inputs[12], so estimating (1.8) requires us to specify sources of variation in μ. We discuss this at greater length below. Our share regression is akin to that in Gandhi et al. (2020) but explicitly in terms of the markup. In this sense, our estimator proxies directly for markups rather than for productivity.

To address the simultaneity problem and separate input elasticities from TFP, we substitute our share regression into the revenue production function. We make an assumption on the evolution of productivity:

Assumption 1: Prices p and productivity ω jointly follow a Markov process with mean-zero shocks η. Defining $p + \omega \equiv \nu$, this implies $\nu_t = g(\nu_{t-1}) + \eta$ with $E[\eta] = 0$.

Assumption 1 says that prices and physical productivity *together* follow a first-order Markov

11. The first line follows from inserting TFP A into the derivation of (1.4): $\mu = \frac{E[A]}{A} \cdot \theta^M \left(\frac{R}{cM}\right)$, and the second from noting that $\frac{E[A]}{A} = \frac{\exp(\omega)E[\exp \epsilon]}{\exp \omega \exp \epsilon} = \frac{E[\exp(\epsilon)]}{\exp(\epsilon)}$.

12. In particular, θ^M does not depend on TFP A from our assumption of Hicks neutrality.

process: it is an assumption on the evolution of TFPR. Write logged firm revenues as:

$$
\begin{aligned}
r_t &= p_t + q_t \\
&= f_t + p_t + \omega_t + \epsilon_t \\
&= f_t + \nu_t + \epsilon_t \\
&= f_t + g(\nu_{t-1}) + \eta_t + \epsilon_t
\end{aligned} \tag{1.9}
$$

Note that the share equation (1.8) implies that $\nu_t = cm_t - f_t - \log(\theta^M) + \log(\mu_t) - b$[13]. Inserting this into (1.9) yields:

$$
r_t = f_t + g(cm_{t-1} - f_{t-1} - \log(\theta^M) + \log(\mu_{t-1}) - b) + \eta_t + \epsilon_t \tag{1.10}
$$

Equation (3.6) identifies the output elasticities in $f(\cdot)$ so long as we can control (proxy) for markups μ. We can obtain an unbiased estimate of the productivity error $\hat{\epsilon}_t$ from regressing the intermediates cost share of revenues on inputs and determinants of markups μ (that is, (1.8)). In our empirical specification, we use plant-year fixed effects as controls for markups, though other researchers applying our methodology may be flexible (for details, see our companion paper Kirov and Traina (2021a)).

Intuitively, our empirical model works as follows: residual TFPR $p \cdot \epsilon$ is a state variable which follows a stochastic process. This is a composite technological and demand process. Firms observe the evolution of this process and adjust output decisions to maximize revenues given its state. But this observation is imperfect: firms make guesses about as-yet-unobserved productivity states. We infer the TFPR state from firm intermediates decisions and an optimization assumption. Then we identify output elasticities from variation in inputs which is orthogonal to variation in innovations to the productivity process η. Identification of

13. To see this, write $\log(\mu_t) = \log(\theta^M) + r_t - cm_t + b - \epsilon_t = \log(\theta^M) + (f_t + \nu_t + \epsilon_t) - cm_t + b - \epsilon_t$ and simplify.

physical output elasticities thus comes from information and timing assumptions on the production process. Finally, we construct markups and labor-market wedges (markdowns) from these output elasticities and observed input choices using the ratio formulas in the previous section.

Comparison to Proxy Estimation

Our estimator is related to the proxy production-function estimation model commonly used in the literature (for example, Olley and Pakes (1996), Levinsohn and Petrin (2003), Ackerberg et al. (2015), Gandhi et al. (2020)). We modify the proxy structure to account for the fact that we only have revenue data, and relax a key assumption in these models. In particular, we impose a Markov timing assumption on TFPR rather than physical productivity a. Proxy models add an assumption that intermediate demand is a monotonic function of other inputs and productivity. Such a monotonicity assumption allows these models to invert productivity a as a function of observed inputs. We do not require a monotonicity assumption since we directly control for markups μ. This allows us to relax the necessary implication that productivity has no other determinants besides observed inputs. This *scalar unobservable* assumption is a requirement in such models because TFP must be inverted as a function of inputs. It is one of the more stylized assumptions in the proxy-model literature, and relaxing it is therefore valuable.

In the production-function estimation literature, the monotonicity assumption is written as $m_t = \mathbb{M}(n_t, k_t, \ell_t, a_t)$. Substituting this proxy function and the evolution of a_t into the quantity production function (3.1) yields:

$$\begin{aligned} q_t &= f(m_t, \ell_t, n_t, k_t) + g(a_{t-1}) + \eta_t + \epsilon_t \\ &= f(m_t, \ell_t, n_t, k_t) + g(\mathbb{M}^{-1}(n_{t-1}, k_{t-1}, \ell_{t-1}, m_{t-1})) + \eta_t + \epsilon_t \end{aligned}$$

In principle, identification of the output elasticities could come from the moment conditions $E[(n_{t-1}, k_{t-1}, \ell_{t-1}, m_{t-1})^T \cdot \eta_t] = 0$. The idea is that lagged inputs affect current input choice through the productivity process, but are plausibly exogenous since they were chosen before a_t.

But this model cannot identify the output elasticity of any flexible inputs (Gandhi et al. (2020)). Because both flexible inputs and productivity a_t evolve period-by-period, and $g(\cdot)$ and $\mathbb{M}(\cdot)$ are nonparametric, flexible inputs are collinear with productivity. For example, observing high m_t could mean that a_t is high or that intermediates are more productive. More formally, inserting the process for productivity into the intermediate demand function and substituting out a_{t-1} yields:

$$
\begin{aligned}
m_t &= \mathbb{M}(n_t, k_t, \ell_t, a_t) = \mathbb{M}(n_t, k_t, \ell_t, g(a_{t-1}) + \eta_t) \\
&= \mathbb{M}(n_t, k_t, \ell_t, g(\mathbb{M}^{-1}(n_{t-1}, k_{t-1}, \ell_{t-1}, m_{t-1})) + \eta_t) \\
&= \mathcal{M}(n_t, n_{t-1}, k_t, k_{t-1}, \ell_t, \ell_{t-1}, m_{t-1}, \eta_t)
\end{aligned}
$$

Since $\mathcal{M}(\cdot)$ is deterministic, lagged inputs determine m_t and so cannot instrument for it. Put another way, there is a tradeoff between the relevance and exogeneity of potential instruments for flexible inputs. A relevant instrument must covary with m_t and so enter $\mathbb{M}(\cdot)$. But if it does, then variation in that instrument is soaked up in pinning down the arbitrary functional form of $\mathbb{M}(\cdot)$. Since our expressions for μ and δ require intermediates and labor to be flexible[14], we need another source of exogenous variation to identify output elasticities.

The source of the collinearity is the full flexibility of $g(\cdot)$ and $\mathbb{M}(\cdot)$. If some candidate instrument z covaries with m, then it should enter $\mathbb{M}(\cdot)$ and any variation in z would be absorbed in estimating $\mathbb{M}(\cdot)$. But in our setting, the input demand functions are the parametric FOCs

14. Ackerberg et al. (2015) acknowledge this issue and assume that labor is not flexible. This solves the econometric issue, but we cannot use the results to estimate markups or markdowns since labor is no longer flexible.

(1.4) and (1.6). Since they are parametric, we do not need to use variation in instruments to pin down the functional form of the productivity-input demand relationship. If we had instruments z_μ and z_δ that generated variation in μ_t and δ_t and were uncorrelated with $m_t, \ell_t, m_{t-1}, \ell_{t-1}$, we could identify θ^M and θ^L[15]. We use plant-year fixed effects as instruments for μ, and wages as instruments for δ (by instrumenting ℓ). Other instruments may be appropriate in different contexts (see our companion paper Kirov and Traina (2021a)).

Moreover, the typical proxy structure in the literature does not account for the fact that we require physical output elasticities but only have revenue data. Our method incorporates this limitation and addresses it by: (1) using the share regression (1.8) to separate prices into a portion accounted for by demand and a portion accounted for by markups; (2) imposing a limitation on the joint evolution of innovations to the processes for prices and productivity.

1.3.2 Empirical Implementation

We use a translog specification for the production function $f(\cdot)$. This allows us to be flexible in describing technology and technological change since output elasticities depend partially on input quantities. We instrument production labor ℓ with wages, following Doraszelski and Jaumandreu (2013). The idea is that wages are partially independently determined from physical labor inputs. Although wages are subject to labor-market wedges δ, they also reflect labor quality and other institutional features of the labor market. We assume that the TFPR process $g(\cdot)$ is AR-1[16]. We assume constant returns to scale in physical output to reduce the number of parameters to be estimated (Flynn et al. (2019) uses the CRS assumption to identify μ). We implement our estimation technique within 6-digit NAICS estimation groups. Denote the vector of inputs as $\mathbf{J}$, and denote logs of variables in lowercase.

15. The FOCs (1.2) and (1.3) are also linear in intermediates and labor *prices* C_t, W_t. One might suggest using independent variation in wages W_t to identify θ_L. But variation in prices must be unrelated to variation in input quality, a condition which is usually unlikely to be met.

16. Alternative specifications of $g(\cdot)$ are possible; our method does not require AR-1 TFPR.

Within each estimation group, we do the following:

1. Regress the intermediates expenditure share of revenues on inputs and plant-year fixed effects. Obtain the predicted share of expenditures $\hat{s_t} = \log(\theta^M) - \log(\mu_t) + b$ and the predicted residual $\hat{\epsilon_t}$. The plant-year fixed effects proxy for determinants of markups μ_t.

2. Regress revenues r_t on inputs m_t, ℓ_t, n_t, k_t, lagged inputs $m_{t-1}, \ell_{t-1}, n_{t-1}, k_{t-1}$, and a polynomial in lagged $\hat{s}_{t-1}$ and $\hat{\epsilon}_{t-1}$. That is, $r_t = f(\theta^{\mathbf{j}} \cdot \mathbf{J_t}) + g(\hat{s}_{t-1} - f(\theta^{\mathbf{J}} \cdot \mathbf{J_{t-1}})) + \hat{\epsilon}_t + \eta_t$. Obtain the predicted residual $\hat{\eta_t}$. Our model implies that this structural error should be 0 in expectation, so we use it to form moment conditions.

3. Use the moment condition

$$\hat{\eta_t} \cdot \begin{bmatrix} 1 \\ w_{t-1} \\ n_{t-1} \\ k_{t-1} \\ w_{t-1}^2 \\ n_{t-1}^2 \\ k_{t-1}^2 \\ w_{t-1} \cdot n_{t-1} \\ w_{t-1} \cdot k_{t-1} \\ n_{t-1} \cdot k_{t-1} \\ (cm - \hat{s})_{t-1} \end{bmatrix} = 0$$

   to recover the structural error η as a function of the vector of elasticities $\theta^{\mathbf{j}}$.

4. Minimize the distance $\hat{\eta}(\theta^{\mathbf{j}})' \cdot W^{-1} \cdot \hat{\eta}(\theta^{\mathbf{j}})$ by GMM with weight W to recover the structural parameters $\theta^{\mathbf{j}}$.

The precise form of the moment conditions comes from our translog specification: n and k instrument for themselves since they are quasi-fixed, and we use wages to instrument for labor. The other moment conditions come from the squared and cross terms. The final instrument $(cm - \hat{s})_{t-1}$ instruments for the constant in the TFPR evolution $g(\cdot)$.

We use the identity matrix for our GMM weight. We initialize this nonlinear system at $f^M = 0.5, f^L = 0.1, f^N = 0.1, f^K = 0.3$, and all of the cross terms at 0[17]. We initialize the AR-1 parameter for productivity at 0.9. Our results are robust to alternative starting values for ρ and the f^j. We constrain our estimated elasticities θ^j to be greater than the average cross-sectional cost shares in the NBER-CES manufacturing database for all inputs j (see Appendix B for details). We use the `scipy.optimize` package in Python to run our estimation. We form numerical derivatives using the quasi-Newton method of Broyden, Fletcher, Goldfarb, and Shanno (BFGS). Again, we assume constant returns to scale throughout, so that the capital elasticity is pinned down by the intermediates, labor, and nonproduction-labor elasticities[18]. This reduces the size of the parameter space and improves the power of our instruments[19].

The output from our GMM estimation are markdowns, markups, output elasticities, and TFPR. Additional details on our estimation technique can be found in Appendix B.

1.3.3 Limitations of our Framework

Our estimate of the labor market wedge δ is (1) estimable at the establishment level; (2) comparable across years; (3) does not require structural estimates of demand or labor supply; and (4) compatible with market power in both input and output markets. It avoids the need

17. f^j here is the coefficient on j in the translog specification, while θ^j is j's output elasticity.

18. CRS is a good approximation to technology in US manufacturing.

19. Following Wooldridge (2009), one could implement the entire procedure in one GMM step, forming the endogenous residual $\hat{\eta}_t$ with each guess of the structural parameters. This implies an additional moment setting $\hat{\epsilon}_t = 0$ in Step (3)

to find natural experiments to identify labor market power.

However, our labor-market wedge δ could pick up structural characteristics of the labor market. Labor adjustment costs would be picked up in δ, since we assumed labor was fully flexible in our baseline specification. Long-term contracts also create a wedge between (current) MRPL and (contractually fixed) wages as workers demand a higher wage now and accept a lower wage in the future to smooth consumption. Adjustment costs and contracts create wage stickiness which the δ estimator interprets as monopsony power. We are also pooling different types of labor together (e.g. skilled and unskilled). This is problematic if labor market power differs between worker types. However, Monte Carlo evidence from Gandhi et al. (2020) suggests such bias is not large. If adjustment costs differ across inputs, we might expect expenditures on intermediates and labor to adjust at different speeds. (1.6) mechanically interprets these movements as shifts in monopsony power.

In terms of production function estimation, a simple alternative to our approach is to assume a Cobb-Douglas production function with constant returns to scale (Syverson (2011)), which implies that output elasticities equal cost shares. But (1.6) suggests that estimating output elasticities at the firm-year level would imply labor wedges equal to unity, since:

$$\delta_t = \left(\frac{C_t M_t}{W_t L_t}\right) \frac{\theta_L}{\theta_M} = \left(\frac{C_t M_t}{W_t L_t}\right) \frac{\frac{P_t Q_t}{C_t M_t}}{\frac{P_t Q_t}{W_t L_t}} = 1$$

Thus, we cannot use a Cobb-Douglas CRS simplification to our structural production-function estimation.

In summary, we embed an intermediate share regression into a modified proxy structure to identify quantity output elasticities with revenue data in the presence of market power. We use wages as instruments for labor and lagged quasi-fixed inputs as their own instruments. Our approach scales linearly with inputs, so can be used to estimate input market power for arbitrary numbers of imperfectly competitive inputs (so long as one input is competitive).

Our estimator is consistent with any functional form for the production function. The estimator we outline is identified, unbiased, consistent, flexible, and accounts for the limitations of using revenue data.

1.4 Data

We use the Annual Survey of Manufactures (ASM) and Census of Manufactures (CM) from the Census Bureau. These data contain production information for manufacturing plants (establishments) in the manufacturing sector in the United States. In particular, we use the US Census Total Factor Productivity Beta Version 2 dataset (TFPB2), which includes the raw ASM/CM data and some constructed variables. In Census years (years which end in -2 or -7), the data come from the CM and constitute the universe of manufacturing plants. In all other years, the data come from the ASM and are sampled with probability[20]. We limit the sample to plants which are observed in ASM years. We use data from 1972-2014[21].

The data include output, intermediates, labor, capital, and other characteristics at the plant level. We merge in price indices for intermediates, investment, and output at the 2-digit DJA industry level from KLEMS, and also aggregate prices from the BEA. We also merge in input and output price indices from the NBER-CES Manufacturing Database. We use these deflators to scale output and input prices and quantities to aggregate industry averages (see Appendix A for details). The idea is to account for unmeasured or mismeasured productive inputs in the ASM/CM data. For instance, the Census data may not account for all managerial labor (particularly labor at headquarters establishments). This problem may have grown over time as economies of scale increased.

We measure output as the total value of shipments less sales from inventories: that is, output

20. Large plants enter the ASM with certainty, while small plants are subject to a sampling probability.

21. These are the years for which the ASM is easily crosswalked to a set of longitudinally consistent Fort-Klimek NAICS codes.

produced rather than output sold. If a plant has sales from inventories which are greater than total value of shipments, we set output equal to total value of shipments. We scale output revenues to 2-digit DJA industries by year using the KLEMS database.

We define intermediates as materials plus energy plus fuels. The idea is to capture a broad range of intermediate inputs, so that the bundle is approximately competitive even if each component of it is not. For quantity measures, we deflate each of materials, energy, and fuels with an appropriate 3-digit deflator from the NBER-CES Manufacturing Database (materials and fuels with $PIMAT$, energy with $PIEN$).

We measure production labor as total production-worker hours at a plant, and nonproduction labor as the total number of nonproduction workers (there is no data on nonproduction worker hours in our sample). The average wage (total pay divided by total hours worked) for nonproduction workers in our sample is over $40/hour, about four times that of production workers. We therefore interpret N as managerial and professional workers. We scale production and nonproduction labor to 2-digit industry aggregates by year using the KLEMS database and the method of Foster et al. (2008). This method apportions a plant's hours to production and nonproduction labor in proportion to its expenditures on each type.

To construct implied wages, we divide each plant's total production-labor expenditures by its total production-worker hours. We do this after scaling both worker hours and payments to labor using the KLEMS database.

We construct capital using the perpetual inventory method (PIM). We initialize each plant's capital stock in the first Census year in which it appears, then iterate forwards and backwards using PIM and investment data. The CM includes measures of plant capital stocks at the beginning and end of each year, split into equipment and structures capital. But these are book values and do not correspond to the economic concept of capital in production. We convert these stocks into market values by multiplying by the ratio of current to historical cost of capital from the BEA capital tables (Kehrig (2015)). We then define total capital as the sum of equipment capital and structures. We adjust capital stocks by 3-digit industry capital utilization rates from the Fed's Plant Capacity Utilization survey. We also scale capital inputs at the 2-digit DJA industry level using the KLEMS database.

The definition of industry codes changes every 5 years in our sample. There are particularly large changes in 1987 (when SIC codes change from the 1972 to 1987 vintage) and in 1997 (when SIC codes change to NAICS codes for all establishments). There are also significant changes in manufacturing NAICS codes in 2012. To get a consistent set of industry codes, we crosswalk establishment-year observations to their corresponding Fort-Klimek 2007 NAICS code (Fort and Klimek (2016)). This crosswalk is available within the Census and assigns a consistent 2007 NAICS code to each establishment.

We drop establishments with missing, zero, or negative inputs or revenues. We also drop establishments with missing values for current or lagged intermediates expenditure share s, as well as those with missing values of lagged inputs (lagged inputs are required for estimation since they instrument for themselves or TFPR). Finally, we drop establishments which have particularly large or small output-to-intermediates ratios or materials-to-labor ratios in each year: plants with such ratios above the 99th or below the 1st percentile by year[22]. This last condition trims μ and δ outliers preemptively. Most establishments trimmed at this stage have extremely large (small) output relative to their inputs or extremely large (small) materials expenditures. They are therefore likely to have different technology from other plants. We use ASM sampling weights in constructing all aggregates from establishment-level data. We limit our sample to complete-case observations: those with an unbroken panel of observations.

We merge in external ACS and CPS data to investigate the aggregate causes and conse-

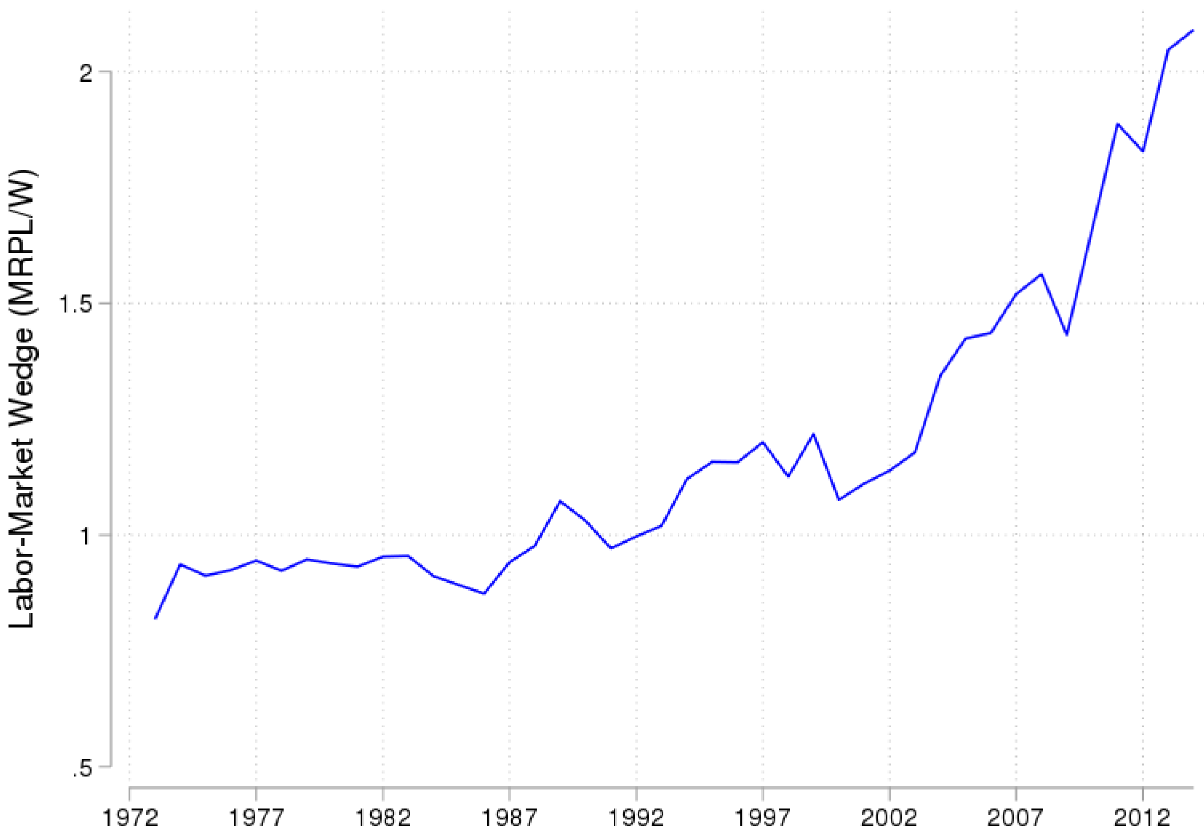
22. We trim rather than winsorize these observations.

quences of the evolution of labor-market power. We download union participation data from the CPS monthly surveys and demographic data from the ACS, including occupation, family size, location, sex, age, employment status, weeks and hours worked, industry, and income measures. We merge in task-intensity data from Autor and Dorn (2013) by occupation code. Finally, we merge this external data to our ASM/CM data by commutfing zone, industry code, and year[23]. We also merge in data on tasks by occupation and geography from Autor and Dorn (2013). We construct measures of high routine-task intensity and high task-offshorability intensity as in Autor and Dorn (2013). Appendix A has more details on the construction of our dataset.

1.5 Labor-Market Wedges and Technological Change in US Manufacturing

We present and discuss our results in this section. First, we show that the aggregate labor-market wedge has risen in the manufacturing sector, particularly after 2000. Second, we show that rises in plant-level δ decrease labor input and the labor share of value added. These wedges are mainly driven by labor demand (firm and industry fixed-effects), but are not correlated with concentration. Third, we show that increases in plant-level labor-market wedges are strongly correlated with increases in direct measures of plant-level technology. Fourth, we show that at the macro level, changes in industry-level labor-market wedges are correlated with falls in unionization and increases in noncognitive task intensities. This evidence suggests that labor-market power has risen over time, and that this rise is driven by changes in technology which has decreased workers' bargaining power relative to that of firms.

Figure 1.1: Aggregate Labor-Market Wedge



Notes: Labor market power over time: Figure shows static aggregate labor-market wedge $\delta \equiv \frac{\theta_L}{\theta_M} \cdot \left(\frac{C_t M_t}{W_t L_t}\right)$ over time for US manufacturing. Aggregation is done using production labor cost weights and ASM sampling weights for each year (see Appendix C).

1.5.1 Trends in Labor-Market Wedges

Figure 2.4 shows the aggregate labor-market wedge in the US manufacturing sector from 1973 to 2014. We obtain this series by aggregating plant-level δ by year, using production-labor expenditure weights (Edmond et al. (2021); see Appendix C for details on our choice of aggregation weights and a comparison to Baqaee and Farhi (2019)). Three points are notable from the figure. First, the aggregate wedge is very close to unity before 1990, suggesting

23. The ACS/CPS data is necessarily aggregated; e.g., it is not available at the plant level.

little aggregate labor-market power. Second, aggregate $\mathcal{D}$ begins to rise in the 1990s to approximately 1.25 in 2000. Third, there was an inflection point around 2002 after which aggregate $\mathcal{D}$ increased rapidly to just above 2. This is a large increase: manufacturing production workers produced marginal output valued approximately at their wages in 1973, but worth twice as much by 2014. The inflection point of this increase coincides both with the beginning of the China shock of the information technology (ICT)-driven boom.

We are cautious in interpreting the level of aggregate manufacturing labor-market power in Figure 2.4 entirely literally. These static estimates ignore adjustment costs, long-term contracts, training programs, search costs, or any other characteristic which makes a firm's labor choice a dynamic problem. That said, our finding that the labor market seems to have become increasingly distorted relative to the intermediates market is likely to be more robust to these criticisms. In unreported results, we show that revenue-weighted and unweighted versions of the aggregate markdown wedge $\mathcal{D}$ rise as well. We also estimated an aggregate labor-market wedge using OLS and a Cobb-Douglas production function, and found a similar time-series pattern. This is because our aggregate wedge is composed of individual wedges $\frac{\theta_L}{\theta_M} \cdot \left(\frac{C_t M_t}{W_t L_t}\right)$. Holding technology constant, our model interprets a rise in the ratio of intermediates to labor expenditure as an increase in the labor-market wedge. In Appendix E, we show the time series of the aggregate cost and elasticity ratios. Both rise, but the cost ratio rises by approximately 75% against a 14% increase in the elasticity ratio. That is, the cost ratio has risen significantly in the data. We estimate changes in elasticities which reinforce this trend, but this changing cost structure drives the aggregate results.

In Appendix E, we show that our model implies an aggregate markup wedge and an aggregate nonproduction wedge which are both approximately unity (the aggregate markup wedge falls slightly since 1987). This suggests that large increases in aggregate wedges in De Loecker et al. (2020) and others may be driven by labor-market wedges rather than markup wedges (see our companion paper Kirov and Traina (2021b) for more details). The low level and stable time series of the output and nonproduction wedges also suggest that our estimation strategy is picking up real increases in labor-market power, rather than merely reflecting misspecification[24].

Figure E.4 and Figure E.5 in Appendix E show that the skewness of the aggregate $\mathcal{D}$ has fallen while its standard deviation has risen. This is in contrast to the markup wedge: Figure E.6 and Figure E.7 in Appendix E show that markup skewness has risen significantly, as in De Loecker et al. (2020) and Kehrig and Vincent (2021). The contrast in these distributions indicate that our aggregate labor-market wedge is not simply picking up a markup wedge. Table 1.1 has pooled cross-sectional summary statistics for inputs, output, labor-market wedges, and markup wedges. δ has a higher mean but also much higher variability than mu. The output elasticities are approximately what one would expect from a gross output production function.

In sum, we find a high aggregate labor-market wedge in 2014 and a particularly large increase in this wedge post-2000. These findings are robust to a number of alternative specifications and are distinct from our findings about markups.

In what sense has the aggregate markdown risen? Figure 1.2 shows a time series of the economy-wide average marginal revenue product of labor and average wage from 1973-2014. Starting in the early 1990s, the average MRPL (red) starts rising faster than the average wage (black) (the series are normalized to 1 in 1987 due to our KLEMS scaling procedure). This process accelerates significantly in the 2000s. The labor-market wedge therefore increases because productivity rises, and not because wages fall. This suggests that technological change plays a large role in the rise of this wedge: δ rises because wage increases did not keep up with productivity after 2000. Which plants does this rise in labor-market wedges come from? Figure 1.3 decomposes the cumulative change in aggregate δ into within-plant,

24. If our results were due to misspecification, there is no reason that this would apply more to production labor than to intermediates or nonproduction labor.

Table 1.1: Input and Output Wedges: Summary Statistics (1972-2014)

	Mean	SD
δ	1.61	(4.43)
μ	1.15	(0.55)
θ^M	0.68	0.17
θ^L	0.15	0.13
θ^N	0.07	0.06
θ^K	0.11	0.09
q	0.22	1.41
m	0.23	1.45
ℓ	0.18	1.26
n	0.18	1.26
k	0.24	1.56

Source: ASM/CMF. All estimates unweighted. δ is our measure of monopsony power, while μ is our measure of output market power. θ^i are output elasticities $\forall$ inputs i. q is log revenues, m is log intermediates, ℓ is log production labor hours, n is log nonproduction workers, and k is log capital inputs. All inputs are scaled to KLEMS aggregates by 2-digit DJA industry and year.

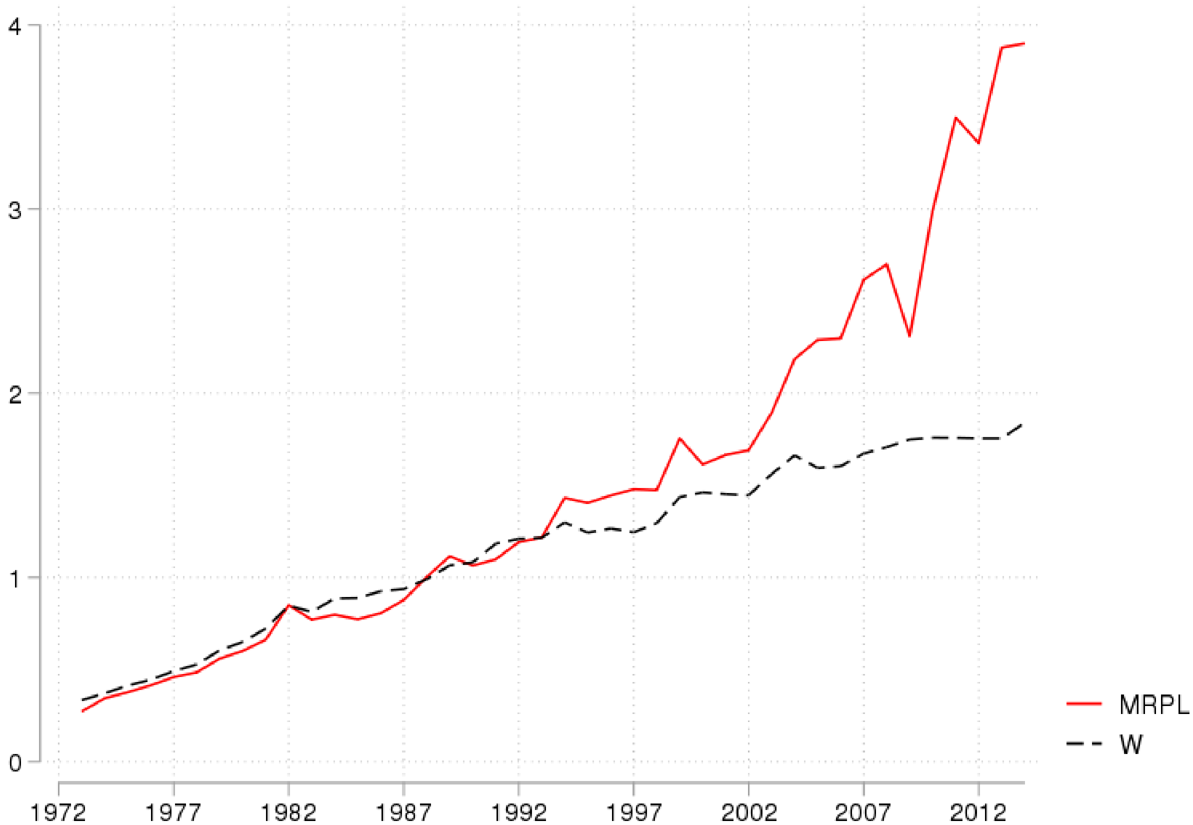
between-surviving-plant, entry, and exit components. The decomposition a Divisia version of the Neil et al. (1992) decomposition:

$$\Delta\mathcal{D}_t = \sum_i \left[\Delta\delta_{it}\frac{s_{i,t}+s_{i,t-1}}{2}\right] + \sum_{i,surv}\left[\Delta s_{it}\frac{\delta_{i,t}+\delta_{i,t-1}}{2}\right] + \sum_{i,x}\left[\Delta s_{it}\frac{\delta_{i,t}+\delta_{i,t-1}}{2}\right] + \sum_{i,e}\left[\Delta s_{it}\frac{\delta_{i,t}+\delta_{i,t-1}}{2}\right] \quad (1.11)$$

The first term represents changes in labor-market wedges holding establishment labor-expenditure shares[25] constant (within). The second represents changes in shares holding labor-market wedges constant for surviving establishments (between survivors), the third does the same for exiters, and the fourth for entrants. In each case, we take an arithmetic average of the object we hold constant as a discrete-time approximation to a continuous process.

25. We use the labor-expenditure share as our establishment-level δ weight since we use this share to aggregate up wedges in our main analysis, consistent with Edmond et al. (2021).

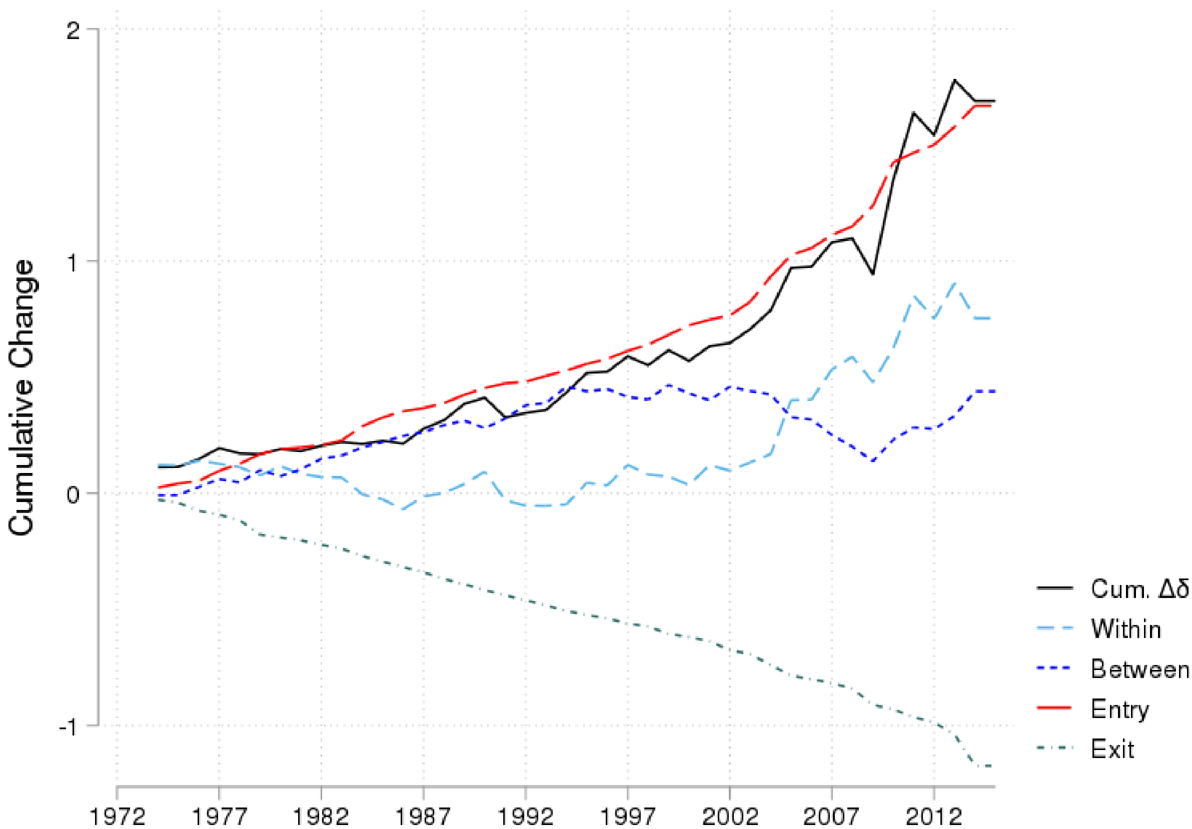
Figure 1.2: Decomposition of Labor-Market Wedge into MRPL and W



Notes: MRPL and W over time: Figure shows average marginal revenue product of labor and wage in each year. MRPL = $\delta \cdot W$ from our estimation. MRPL and W are aggregated across establishments by year using labor-expenditure weights. Both measures are scaled using KLEMS weights, which sets the mean wage to 1 in 1987.

Figure 1.3 plots the decomposition described in (1.11). Churn accounts for much of the variation in $\mathcal{D}$: the red entry and dashed-green exit lines are large in magnitude. Moreover, the cumulative entry component rises more than the cumulative exit component falls, so that net entry contributes significantly to the rise in the overall wedge. Thus new establishments enter with larger labor-market wedges, while establishments with low wedges systematically exit. After 2002, the dashed teal within component also begins to rise. These results are consistent with technological changes which decrease the share of output paid to production workers; for instance automation or telecommunications technologies. New plants adopt such technologies, but it takes time for them to grow and increase the aggregate wedge. Plants which are unable to adopt exit. Eventually, these technologies become more widespread among existing plants and increase their wedges as well. We find direct evidence for such a story below. We further investigate the evolution of labor-market power over plant life-cycles

Figure 1.3: Within-Between Decomposition of Labor-Market Wedge



Notes: Decomposition of establishment-level changes in δ: Labor-market wedge is set to 0 in 1973, then cumulative changes are computed from $\Delta\mathcal{D}_t =$
$\sum_i \left[\Delta\delta_{it}\frac{s_{i,t}+s_{i,t-1}}{2}\right] + \sum_{i,surv}\left[\Delta s_{it}\frac{\delta_{i,t}+\delta_{i,t-1}}{2}\right]\sum_{i,x}\left[\Delta s_{it}\frac{\delta_{i,t}+\delta_{i,t-1}}{2}\right]\sum_{i,e}\left[\Delta s_{it}\frac{\delta_{i,t}+\delta_{i,t-1}}{2}\right].$
The first term is within-establishment changes (dashed teal), the second between-surviving-establishment changes (dashed blue), the third is changes due to entrants (dash red), and the fourth changes due to exiters (dotted green). Figure is weighted by ASM sampling weight and labor expenditures.

by running the regression:

$$\log \delta_{ijt} = \alpha + \gamma_t + \gamma_j + \beta_1 \mathbb{1}[entrant_{ijt}] + \beta_2 \mathbb{1}[exiter_{ijt}]$$
$$+ \beta_3[young_{ijt}] + \beta_4[medium_{ijt}] + \beta_5[old_{ijt}] + \epsilon_{ijt}$$

Where *entrant* is an indicator variable for entry of plant i in industry j in year t (1 if a plant first appears in the ASM/CMF, 0 otherwise), *exiter* is an indicator variable for exit (1 if a plant last appears in the ASM/CMF, 0 otherwise), *young* is an indicator for plant age $\leq$ 5 years, *medium* is an indicator for $5 <$ plant age ≤ 10, and *old* is an indicator for plant age > 10. The regression includes year (t) and NAICS6 industry (j) fixed-effects, so that variation comes within industries and years. It is weighted by labor expenditure and ASM sampling weight. We exlcude years ending in -3 and -8: the Census rebalances its ASM panel in these years, creating excess churn due to sample rotation.

This model, analagous to that in Foster et al. (2008), tests the impact of the average plant's age on its labor-market wedge. Table 1.2 shows the results, both with and without NAICS6 fixed-effects[26] (old plants are the omitted category, so all coefficients are relative to old plants' labor-market wedges). We draw two main conclusions from Table 1.2. First, both entrants and exiters have lower labor-market wedges than survivors[27]. Second, conditional on survival, δ has an inverse-U relationship with tenure: young plants have relatively low labor-market wedges, medium-aged plants have higher wedges, and old plants lower wedges again. This pattern suggests a technology-adoption curve: survivors and medium-aged plants adopt technologies which allow them to increase δ, but it takes time to adopt. Plants which do not adopt have lower δ (exiters and old plants). This interpretation is speculative, but we find direct evidence for a link between technology adoption and δ in Section 1.5.3.

26. We always include year FEs to account for technological change.

27. The implied coefficient for entrants is $\beta_1 + \beta_3 \approx -0.1245$ (for Column 2), approximately the same as that for exiters.

Table 1.2: Labor-Market Wedges over Plant Lifecycles

	Year FEs	Year, NAICS6 FEs
	(1)	(2)
	$\log(\delta)$	$\log(\delta)$
Entry	-0.0853***	-0.0608***
	(0.0103)	(0.00459)
Exit	-0.0907***	-0.130***
	(0.0109)	(0.00487)
Young	-0.0683***	-0.0817***
	(0.00662)	(0.00318)
Medium	0.0627***	0.0317***
	(0.00739)	(0.00339)
Old	.	.
	.	.
(Constant)	-1.019***	-1.009***
	(0.00325)	(0.00153)
N	494,000	494,000
R^2	0.011	0.804
AIC	1,862,000	1,062,000

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$
Labor-expenditure- and sampling-weight-weighted regression of log labor-market wedge δ on indicators for entry (first appear in ASM/CMF), exit (last appear in ASM/CMF), and indicators for young (age $\leq$ 5), medium (5 < age $\leq$ 10, and old (age > 10). Excludes years ending in -3 and -8 due to ASM sample rotation. Includes year (Column 1) and year and NAICS6 (Column 2) fixed-effects. *Old* is the omitted category - all coefficients should be interpreted relative to that of old plants.
Source: ASM/CMF.

We turn next to some suggestive evidence on the sources of labor-market power. Is labor-market power driven more by labor demand (firms) or labor supply (workers)? To get at this question, Table 1.3 regresses $\log(\delta)$[28] on different sets of fixed-effects at the plant level. Each regression includes year fixed-effects, so the models we run are of the form:

$$\log(\delta_{cjt}) = \alpha + \gamma_c + \gamma_j + \gamma_t + \epsilon_{cjt}$$

28. We use the log of the labor-market wedge here to correct for skewness in plant-level δ.

Where c is commuting-zone (CZ), j is industry, and t is year. Each regression is weighted by labor expenditures and the ASM sampling weight. Column 1 of Table 1.3 includes only time fixed effects, and has low adjusted R^2 and high AIC: it is not a particularly predictive model. Column 2 adds industry fixed effects by including a full set of NAICS6 dummies. The R^2 of the regression increases eightyfold from 0.01 to 0.81, and the AIC falls by almost half. Thus, industry adds significant explanatory power to a simple time trend.

Columns 3 and 4 of Table 1.3 compare a specification with commuting-zone and time fixed-effects (column 3) to one with CZ, time, and NAICS6 industry (column 4). Adding NAICS6 increases the R^2 eightfold and reduces the AIC by nearly half. Again, more of the variation in δ comes from industry than from geography. Table E.1 in Appendix E compares similar regressions with plant i and firm f fixed-effects: these additional explanatory variables do not add much explanatory power beyond including NAICS6 industry. Table 1.3 shows that

Table 1.3: Fixed-Effect Regression Results

	Year FEs	Year, NAICS6 FEs	Year, CZ FEs	Year, CZ, NAICS6 FEs
	(1)	(2)	(3)	(4)
	$\log(\delta)$	$\log(\delta)$	$\log(\delta)$	$\log(\delta)$
R^2	0.01	0.81	0.111	0.815
$R^2_{adj.}$	0.01	0.81	0.110	0.814
AIC	2,343,000	1,334,000	2,277,000	1,302,000

(1) Year fixed effects; (2) Year, NAICS6 fixed effects; (3) Year, CZ fixed effects; (4) Year, CZ, NAICS6 fixed effects. Labor-expenditure- and ASM-sampling-weight- weighted regression of $\log(\delta)$ on different fixed effects. Source: ASM/CMF.

labor demand is an important contributor to the labor-market wedge δ, while labor supply is much less so. In other words, the rise in labor-market power is explained more by a plant's industry than by its commuting zone. In both levels and changes, industry-level variation is more closely correlated with variation in a plant's labor-market wedge than CZ-level variation. This suggests that changes in technology are an important contributor to increased economy-wide labor-market power: technological change at the industry level explains increased δ more strongly than institutional change at the geographic level. We provide direct evidence of this in Section 1.5.3.

In sum, then, labor-market wedges have risen steadily since 1990, and particularly since 2002. Since the late 1990s, labor-market wedges have increased greatly while markups have remained flat. Thus, labor-market power accounts for all of the increase in aggregate market power in US manufacturing in this period. The rise of labor-market power is driven both by entry of new plants and increases within existing plants, and is largely a within-industry phenomenon. This suggests that higher labor-market power is driven by changes in labor demand, and in particular by the emergence of new technologies. As plants innovate, they grow output but also reduce the share of output which goes to production workers. We investigate this directly in Section 1.5.3 below. First, we investigate the consequences of labor-market power.

1.5.2 Consequences of the Rise in Labor-Market Power

We now investigate the consequences of labor-market power. How has δ impacted wages and employment? Which plant characteristics is it correlated with? To begin with, we regress logs of wages, labor inputs, and labor expenditures on the log labor-market wedge δ at the plant level. The model for this regression is:

$$\log(L_{ijt}) = \beta_0\gamma_t + \gamma_j + \beta_1 \log(\delta_{ijt}) + \epsilon_{ijt}$$

Where L_{ijt} on the LHS is either labor inputs (production worker hours), wages, or total labor compensation. The form of the regression implies that the coefficients should be interpreted as elasticities.

Table 1.4 shows the results with various sets of fixed-effects. Columns 1-3 include only year FEs while columns 4-6 include year and NAICS6 industry FEs. In each case, a larger δ reduces labor inputs four to six times as much as it reduces wages. For instance, when including year and industry fixed effects, a 1% increase in δ decreases wages by 0.027% but decreases (production) labor inputs by 0.17%. Table 1.4 therefore shows that our labor-market wedge measures quantity restrictions on the part of firms, rather than wage reductions (of course, the two are tightly linked by the firm's residual labor supply curve). Table E.2 in Appendix E shows the same result holds when adding firm or plant fixed-effects. This is equivalent to doing the regression in *changes*, and so suggests that firms and plants with large increases in δ reduce hiring 2-6 times more than they reduce pay. This is consistent with a narrative of technological change in the labor market: more productive, capital-augmented insiders capture high pay while outsiders are not hired. What proportion of the

Table 1.4: Regression of Wages, Labor Inputs, and Labor Expenditures on Labor-Market Wedge

	Year FEs			Year, NAICS6 FEs		
	(1)	(2)	(3)	(4)	(5)	(6)
	w	ℓ	wl	w	ℓ	wl
$\log(\delta)$	-0.0168***	-0.049***	-0.134***	-0.0272***	-0.173***	-0.193***
	(0.000358)	(0.000962)	(0.00119)	(0.000639)	(0.00188)	(0.00201)
Constant	0.191***	0.145***	7.861***	0.183***	0.0475***	7.814***
	(0.00064)	(0.00172)	(0.00212)	(0.000727)	(0.00214)	(0.00229)
N	646000	646000	646000	646000	646000	646000
R^2	0.054	0.031	0.105	0.221	0.044	0.336
AIC	832,700	2,111,000	2,382,000	706,800	2,102,000	2,189,000

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$
(1)-(3) Year fixed effects; (4)-(6) Year, NAICS6 fixed effects.
Labor-expenditure- and ASM-sampling-weight- weighted regression of log wages, labor inputs, and labor expenditures on log labor-market wedge δ.
Source: ASM/CMF.

large decrease in the manufacturing labor share can be plausibly explained by the rise in the labor-market wedge? Section 1.2.2 showed that the aggregate labor share is inversely proportional to both markup and labor-market wedges: $\frac{WL}{PQ} = \frac{\theta^L}{\mu\delta}$. Figure 1.4 shows that labor shares are indeed strongly negatively correlated with labor-market wedges. Regressing the establishment-level labor share on $\log(\delta)$ with year and NAICS6 fixed-effects[29] yields

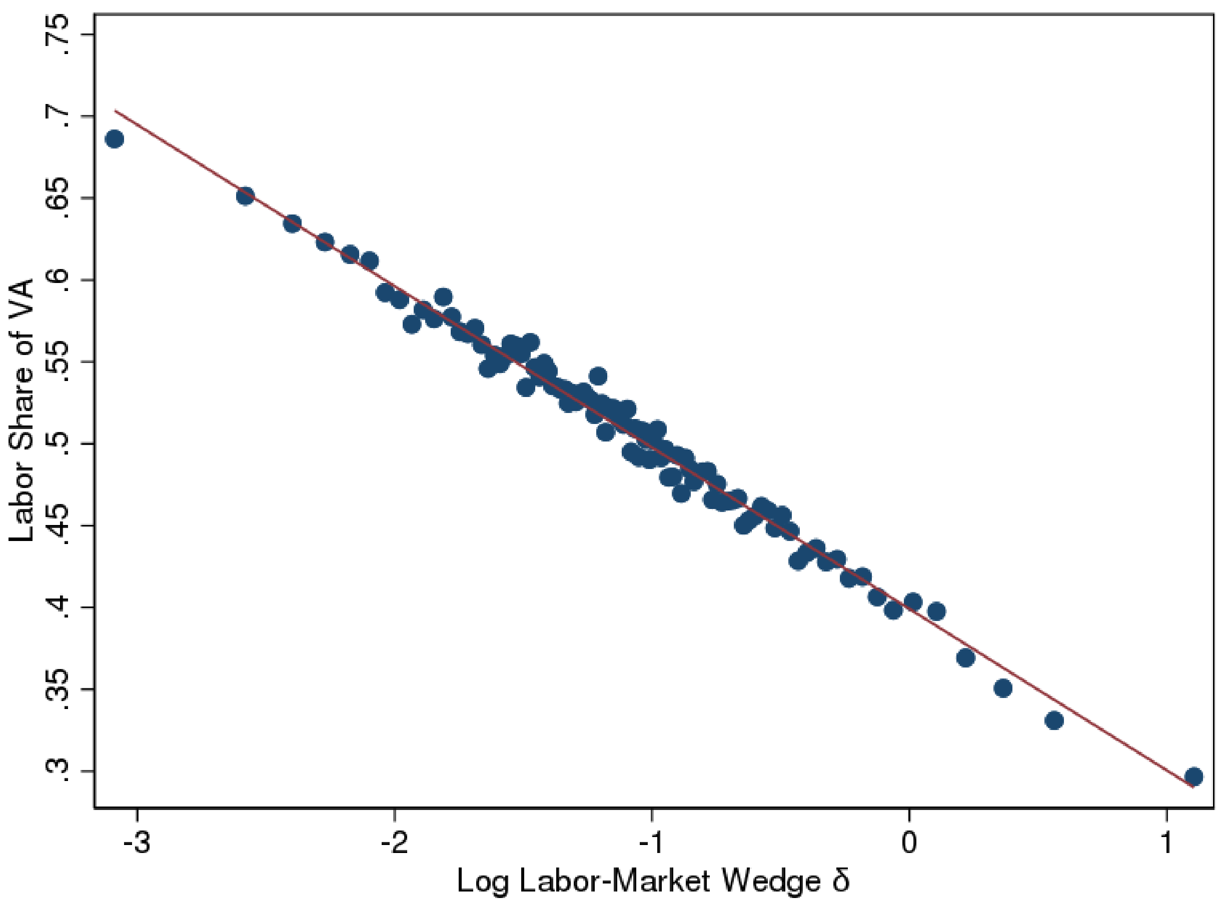
29. The results in the previous section, in particular Table 1.3, suggest that this is a reasonably complete set of fixed-effects.

a coefficient of approximately -0.09: a 1% increase in δ yields a reduction in the labor share of 0.09 percentage points[30] This is a within-year cross-sectional relationship because of the year fixed-effects. But applying this estimate to the time series in a back-of-the-envelope calculation suggests that the 144% increase in δ between 1973 and 2014 reduced the manufacturing labor share by 13%. The labor share in manufacturing fell from 62% in 1967 to 41% in 2012, so this calculation implies that two-thirds (62%) of the decline in the manufacturing labor share can be attributed to rising labor-market power. This regression ignores the reallocation of labor and markdowns across establishments over time, but it suggests that changes in labor-market power are important contributors to declines in the aggregate labor share in manufacturing. Figure 1.4 compares the cross-sectional relationship between the labor share of value added and the labor-market wedge at different levels of aggregation. The figures plot regressions of the form $LS_{ifjt} = \alpha + \gamma_t + \gamma_j + \gamma_y + \beta_{change}\left[\log(\delta_{ifjt})\right] + \epsilon_{ifjt}$, where $y \in \{i, f\}$. The left figure includes firm, NAICS6, and year fixed-effects (fjt), while the right figure includes plant, NAICS6, and year FEs (ijt). Including firm (plant) FEs gives these regressions the interpretation of *changes*: they ask, how do labor-shares change when δ changes at the firm (plant) level?

In each case, the cross-sectional relationship between labor-market wedges and labor shares continues to hold in changes: greater increases in δ lead to smaller increases (greater decreases) in the labor share. The relationship is twice as large (β_{change} is twice as large) when including firm fixed-effects (that is, controlling for firm-level characteristics). This implies that plant-level variables, such as specific technology and managerial nous, are more important in mediating the impact of labor-market wedges δ than firm-level variables such as overall technological or managerial choices. This contrasts with the industrial organization literature emphasizing firm-level optimization of multi-plant organizations (Atalay et al. (2014)). It is also another piece of supporting evidence for the importance of techno-

30. The regression model is $LS_{ijt} = \alpha + \gamma_t + \gamma_j + \beta \log(\delta_{ijt}) + \epsilon_{ijt}$, where the γ_x represent fixed-effects with respect to x.

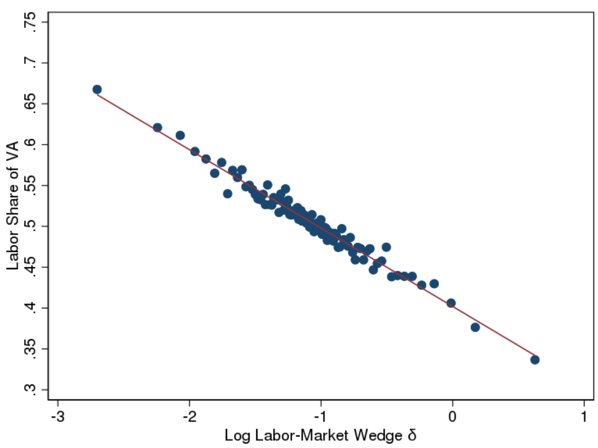
Figure 1.4: Binscatter of Labor-Market Wedge against Labor Share



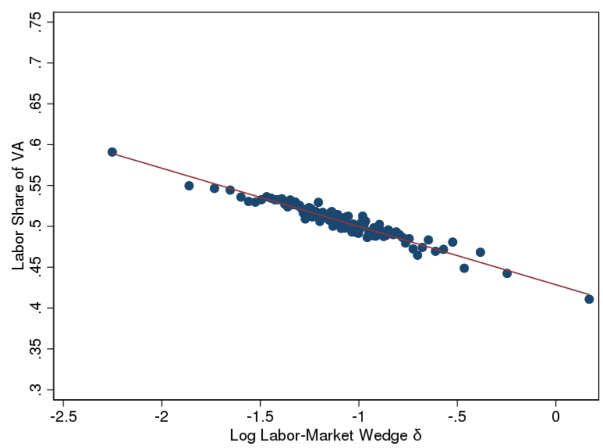
Notes: Binscatter of $\log(\delta)$ against the labor share of value added $\frac{WL}{PQ-CM}$ at the plant level with year & NAICS6 (jt) fixed effects. The figure residualizes $\log(\delta)$ and the labor share against year and NAICS6 fixed-effects, then bins it into 100 equally-sized bins and plots the mean of the labor share within each bin along with a (red) line-of-best-fit. Figure is weighted by ASM sampling weight and labor expenditures.

logical change in driving δ and its economic impacts. Figure 1.6 examines the relationship between the labor-market wedge δ and the number of workers at a plant. We plot means of $\log(\delta)$ and total employees ($\log(\texttt{tpw}+\texttt{oe})$) against each other with year and industry fixed-effects (left) and year, industry, and firm fixed-effects (right). In the cross-section (left figure), plants with larger labor-market wedges do not have significantly more or fewer total workers. Combined with the evidence in Table 1.4, this suggests that plants with high labor-market wedges reduce production employment and increase nonproduction (manage-

Figure 1.5: Binscatter of Labor-Market Wedge against Labor Share, Plant and Firm Fixed Effects



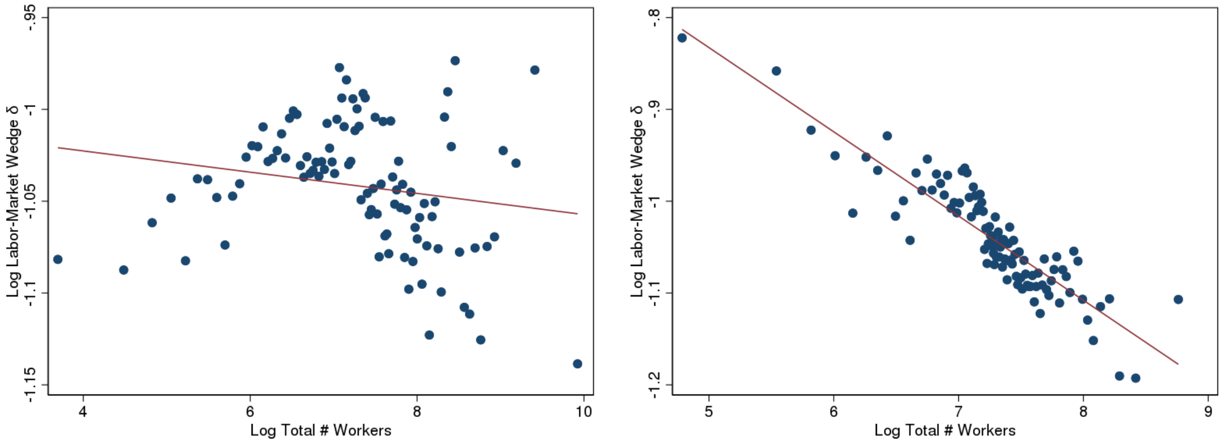
(a) Firm, NAICS6, Year FEs



(b) Plant, NAICS6, Year FEs

Notes: Binscatters of $\log(\delta)$ against the labor share of value added $\frac{WL}{PQ-CM}$ at the plant level with year, NAICS6, firm (fjt, left) and year, NAICS6, plant (ijt, right) fixed effects. Each figure residualizes $\log(\delta)$ and the labor share against fjt or ijt fixed-effects, then bins it into 100 equally-sized bins and plots the mean of the labor share within each bin along with a (red) line-of-best-fit. Figure is weighted by ASM sampling weight and labor expenditures.

Figure 1.6: Binscatter of Labor-Market Wedge Against Plant Workers



(a) Year, NAICS6 FEs **(b)** Plant, NAICS6, Year FEs

Notes: Binscatters of $\log(\delta)$ against the log total number of workers at the plant level with year, NAICS6 (jt, left) and year, NAICS6, firm (fjt, right) fixed effects. The total number of workers is defined as (production workers + nonproduction workers). Each figure residualizes $\log(\delta)$ and the total number of workers against jt or fjt fixed-effects, then bins $\log(delta)$ into 100 equally-sized bins and plots the mean of the number of workers within each bin along with a (red) line-of-best-fit. Figure is weighted by ASM sampling weight and labor expenditures.

rial) employment. This is another piece of evidence for a relationship between technological change and the labor-market wedge: large δ are associated with more intensive use of managerial technologies. Moreover, plants with large labor-market wedges subsequently reduce hiring (Figure 1.6, right panel: including firm fixed-effects makes this figure analagous to a regression in changes). Thus large and more complex[31] plants have higher labor-market wedges. These plants have a lower labor-share, more of which is explained by plant-level variation. They have more managerial employees and fewer production workers, and subsequently shrink employment. In Figure E.8 in Appendix E, we show that plants with higher revenues have larger labor-market wedges, and that rising revenues are associated with rising wedges. Figure E.10 in Appendix E shows that there is essentially no relationship between the labor-market wedge and productivity (TFPR).

31. In the sense of greater managerial intensity.

In sum, plants with higher labor-market wedges reduce hiring more than they reduce pay, while substituting managers for production workers. They also have significantly lower labor shares of value added: indeed, a rough calculation suggests that fully two-thirds of the aggregate decline in the aggregate manufacturing labor share can be attributed to increased labor-market power. Plants with higher revenues tend to have higher δ, though there is no strong relationship with productivity. Thus our labor-market wedge has significant consequences at the plant level. But what drives increases in labor market power? We turn to this question next.

1.5.3 Micro Sources of Labor-Market Power

Labor-market wedges have increased significantly in the past 40 years, and are robustly associated with a range of outcomes at the plant level. But what caused the rise in labor-market power? In this section, we find that direct measures of plant technology strongly correlate with increases in labor-market power. Concentration, by contrast, does not.

Technology

Figure 2.4 shows that the aggregate labor-market wedge had inflection points upward in the early 1990s and the early 2000s. These coincide with the ICT revolution as well as the China shock. We found that much of the rise in the wedge is driven by entrants (and incumbents post-2002), and that industry is an important determinant of plants' markdown power. This suggested a technological explanation for the labor-market wedge. In this section, we directly examine how changes in plants' technology correlate with changes in their labor-market wedges.

The ASM/CMF includes several plant-level measures of ICT-related technological change in recent years. In 2000 and 2001, plants were asked to report their new computer expenditures. In 1997, they were asked about their software expenditures. And in 1997, 2002, and each year after 2006, they were asked about their cost of purchased communications. We normalize each of these measures by dividing by total plant employment, then regress plant-level labor-market wedges on these technological ratios with fixed effects. For plants i in NAICS6 industries j in years t, we run:

$$\log(\delta_{ijt}) = \alpha + \gamma_j + \gamma_t + \beta T_{ijt} + \epsilon_{ijt} \tag{1.12}$$

Where the measure of technology $T_{ijt} \in \{\log\left(\frac{nmc}{workers}\right), \log\left(\frac{cs}{workers}\right), \log\left(\frac{cpc}{workers}\right)\}$ for ASM technological measures $nmc \equiv$ (new computer expenditures), $cs \equiv$ (software expenditures), $cpc \equiv$ (cost of purchased communications). Since the ASM variables are in nominal terms, we first deflate our measures of technological expenditures using the NBER-CES shipments price deflator at the 6-digit NAICS level. We normalize by the number of workers (production workers plus nonproduction workers) to obtain a measure of technological intensity, rather than pick up effects from plant size. The idea is to test whether plants which spend more on technology *relative to their workforce* differ systematically in δ. Our results are robust to normalizing these measures by revenues instead of number of workers, as well as to directly controlling for plant size. Table 1.5 shows the results. All three technological intensity measures are economically and statistically significant determinants of labor-market wedges[32]. The elasticity of the plant-level δ with respect to new computer expenditures per worker is 0.08, so that a 1% rise in computer expenditures implies a 0.08% rise in δ. The elasticity of the labor-market wedge with respect to software is 0.02, and with respect to communications expenditure 0.1. These are large effects, since the overall cross-sectional mean of δ is approximately 1.6. Moreover, variation in these measures of technological expenditures explain a significant portion of the variation in plant-level δ.

Nor are these results driven by outliers: Figure 1.7, Figure 1.8, and Figure 1.9 show cross-sectional binscatters of the regressions in Table 1.5. In each case, plants which spend more on

32. We cannot run a joint regression since these measures are available in the ASM/CMF in different years.

Table 1.5: Regression of Labor-Market Wedge on Direct Measures of Technology

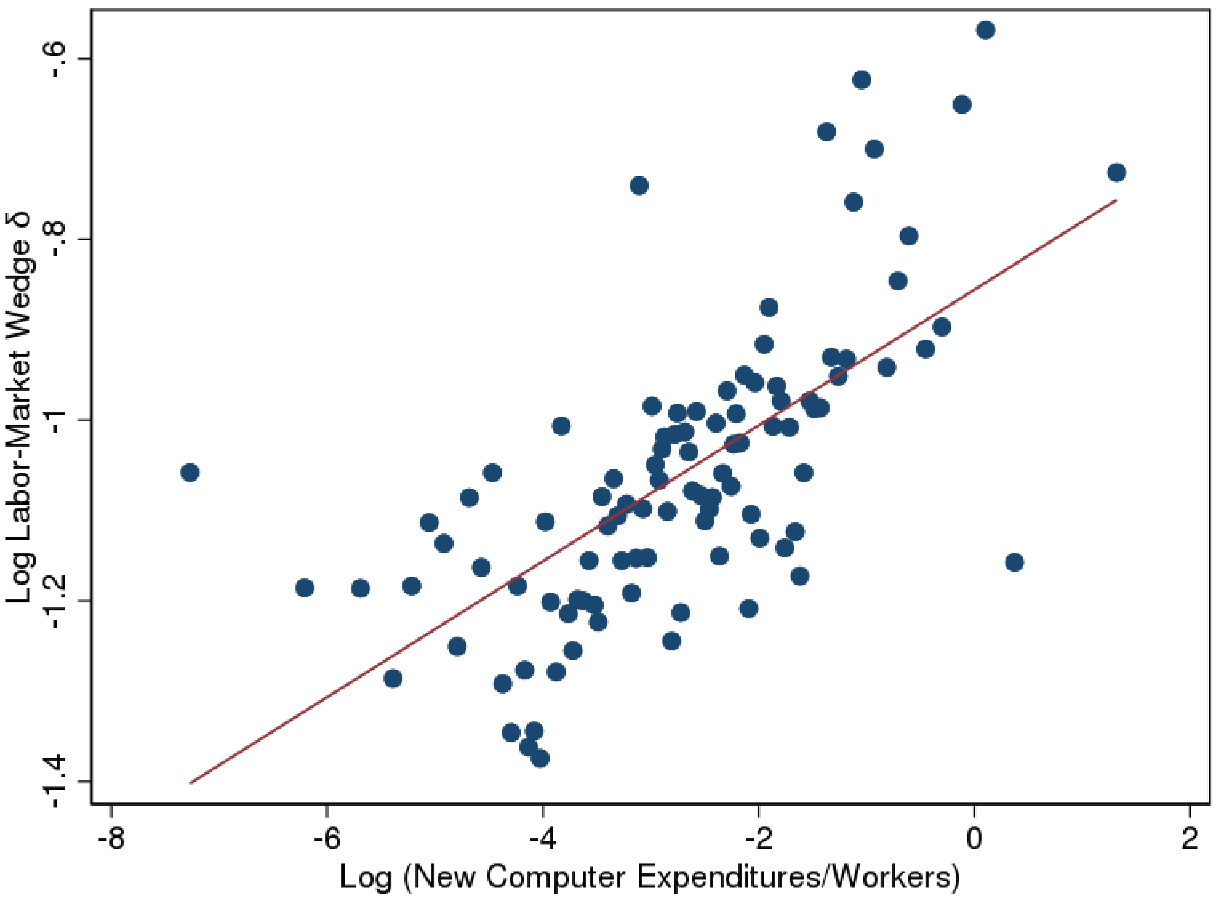
	Year, NAICS6 FEs		Year FEs
	(1)	(2)	(3)
	$\log(\delta)$	$\log(\delta)$	$\log(\delta)$
New Computer Expenditures/Worker (00-01)	0.0752***		
	(0.00419)		
Purchased Communications/Worker (97, 02, 06-14)		0.109***	
		(0.00168)	
Software Expenditures/Worker (97)			0.0209***
			(0.00551)
N	12,000	6,200	16,000
R^2	0.837	0.856	0.804
AIC	24,330	11,080	356,000

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$
Labor-expenditure- and ASM-sampling-weight- weighted regression of log labor-market wedge δ on log new computer expenditures per worker (row 1), log cost of purchased communications per worker (row 2), and log software expenditures per worker (row 3). New computer expenditures are available 2000-2001, software expenditures 1997, cost of purchased communications 1997, 2002, 2006-2014. Technological variables are deflated and normalized by (workers) $\equiv$ (production workers + other employees).
Source: ASM/CMF.

ICT technology have larger labor-market wedges. This pattern is consistent across the range of technology spending and labor-market wedges: it is not driven by a cohort of exceptional 'superstar' firms. Appendix E shows that including firm or plant fixed-effects does not change this positive association. Thus plants which buy more computers, more software, and more communications have higher labor-market wedges, and plants with larger increases in these measures have larger increases in δ (including plant or year fixed effects is equivalent to a regression in differences). Along with the suggestive findings in the previous section, this is strong evidence that the increase in the labor-market wedge is driven by technological change, and in partcular by the automation and computerization of the ICT revolution.

In Figures 1.10 - 1.12, we illustrate the effect of technological investment during the inflection years of the late 1990s. Each of these figures divides one of real per-worker computer purchases, software purchases, and communications purchases into three terciles in a base

Figure 1.7: Binscatter of Labor-Market Wedge against New Computer Expenditures Per Worker



Notes: Binscatter of $\log(\delta)$ against log new computer expenditures per worker $\log\left(\frac{nmc}{workers}\right)$ at the plant level with year & NAICS6 (jt) fixed effects. The figure residualizes $\log(\delta)$ and per-worker computer expenditures against year and NAICS6 fixed-effects, then bins each into 100 equally-sized bins and plots the mean of $\log(\delta)$ within each bin along with a (red) line-of-best-fit. Workers are defined as production workers plus nonproduction workers. New computer expenditures are deflated using a NAICS6 NBER-CES shipments price index. Figure is weighted by ASM sampling weight and labor expenditures. For years 2000-2001 due to data availability.

year, then plots the evolution of the average labor-market wedge within these (fxied) terciles in subsequent years. The idea is to see to what extent higher ICT investment in the late 1990s impacts later changes in labor-market wedges (the base year is 2000 for computer expenditures and 1997 for software and communications purchases because of data availability

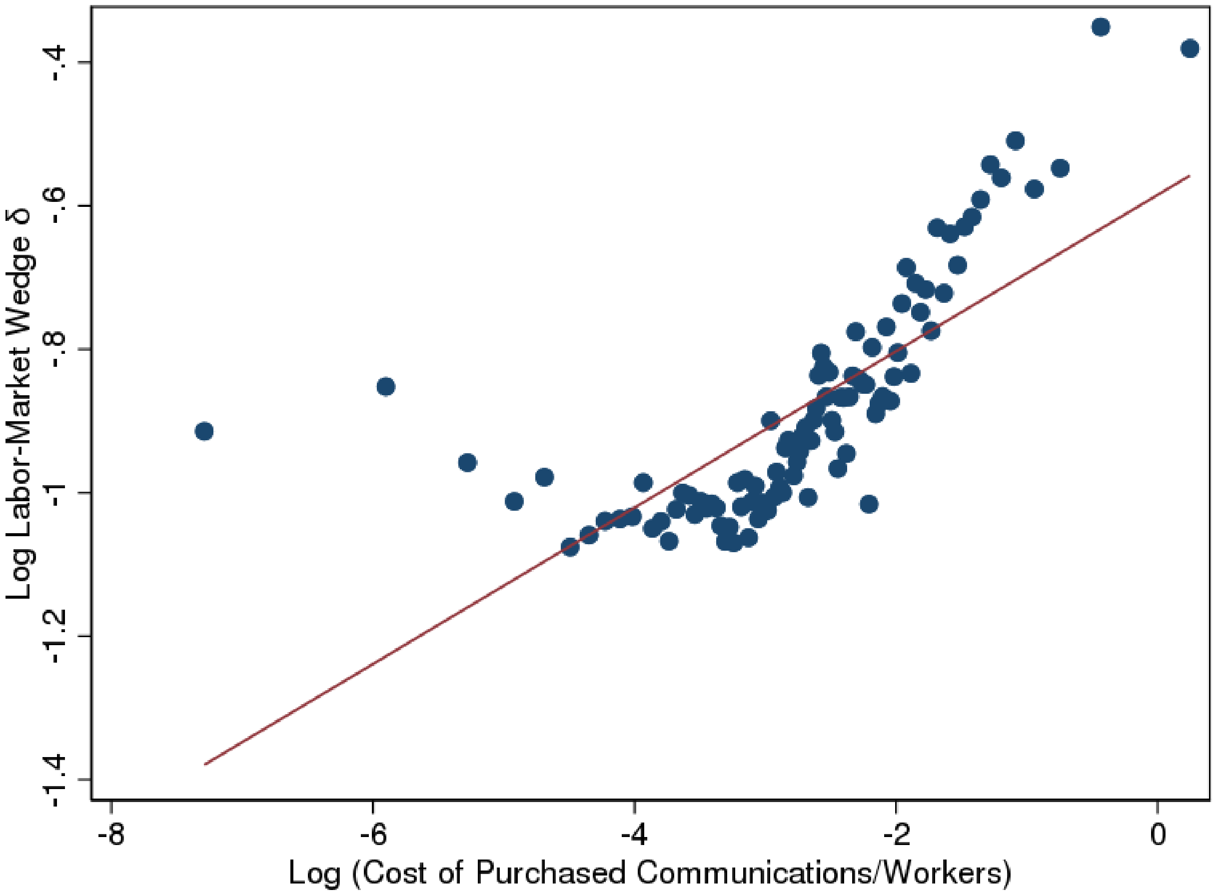
Figure 1.8: Binscatter of Labor-Market Wedge against Software Expenditures Per Worker

Notes: Binscatter of $\log(\delta)$ against log software expenditures per worker $\log\left(\frac{cs}{workers}\right)$ at the plant level with NAICS6 (j) fixed effects. The figure residualizes $\log(\delta)$ and per-worker software expenditures against NAICS6 fixed-effects, then bins each into 100 equally-sized bins and plots the mean of $\log(\delta)$ within each bin along with a (red) line-of-best-fit. Workers are defined as production workers plus nonproduction workers. Software expenditures are deflated using a NAICS6 NBER-CES shipments price index. Figure is weighted by ASM sampling weight and labor expenditures. For year 1997 due to data availability.

limitations). If technology had no impact on δ, we would expect the three lines in each graph to be roughly equivalent after the base year.

Instead, we see that ICT investment has a significant impact on plants' subsequent labor-market wedges. Figure 1.10 shows that the top two terciles (blue, red) of per-worker computer

Figure 1.9: Binscatter of Labor-Market Wedge against Communications Expenditures Per Worker



Notes: Binscatter of $\log(\delta)$ against log cost of purchased communications per worker $\log\left(\frac{cpc}{workers}\right)$ at the plant level with year & NAICS6 (jt) fixed effects. The figure residualizes $\log(\delta)$ and per-worker communications purchases against year and NAICS6 fixed-effects, then bins each into 100 equally-sized bins and plots the mean of $\log(\delta)$ within each bin along with a (red) line-of-best-fit. Workers are defined as production workers plus nonproduction workers. Communications purchases are deflated using a NAICS6 NBER-CES shipments price index. Figure is weighted by ASM sampling weight and labor expenditures. For years 1997, 2002, and 2006-2014 due to data availability.

investment saw their average post-2000 labor market wedges increase by almost 50% by 2014, while the average wedge of the lowest tercile (black dashed) increased by 20%. Moreover, the level of the aggregate wedge is higher for these terciles (1.8 in 2014) than for the lowest tercile (1.25 by 2014). Figure 1.11 shows a similar story for 1997 per-worker software expenditures: the highest tercile of plants by software expenditures (blue) increased their wedges by almost 30% to 1.8 by 2008, while the middle (red) and lower (black dashed) terciles saw smaller increases. The three terciles converge after 2008, which suggests that 1997 software expenditure probably depreciated by the late 2000s. Finally, Figure 1.12 shows that the highest two terciles of 1997 per-worker communications expenditures have higher levels of average δ until 2007, again suggesting faster depreciation of communications investment. In each of Figures 1.10 - 1.12, plants which on average invest more in technology have higher levels of and larger increases in their average labor-market wedges.

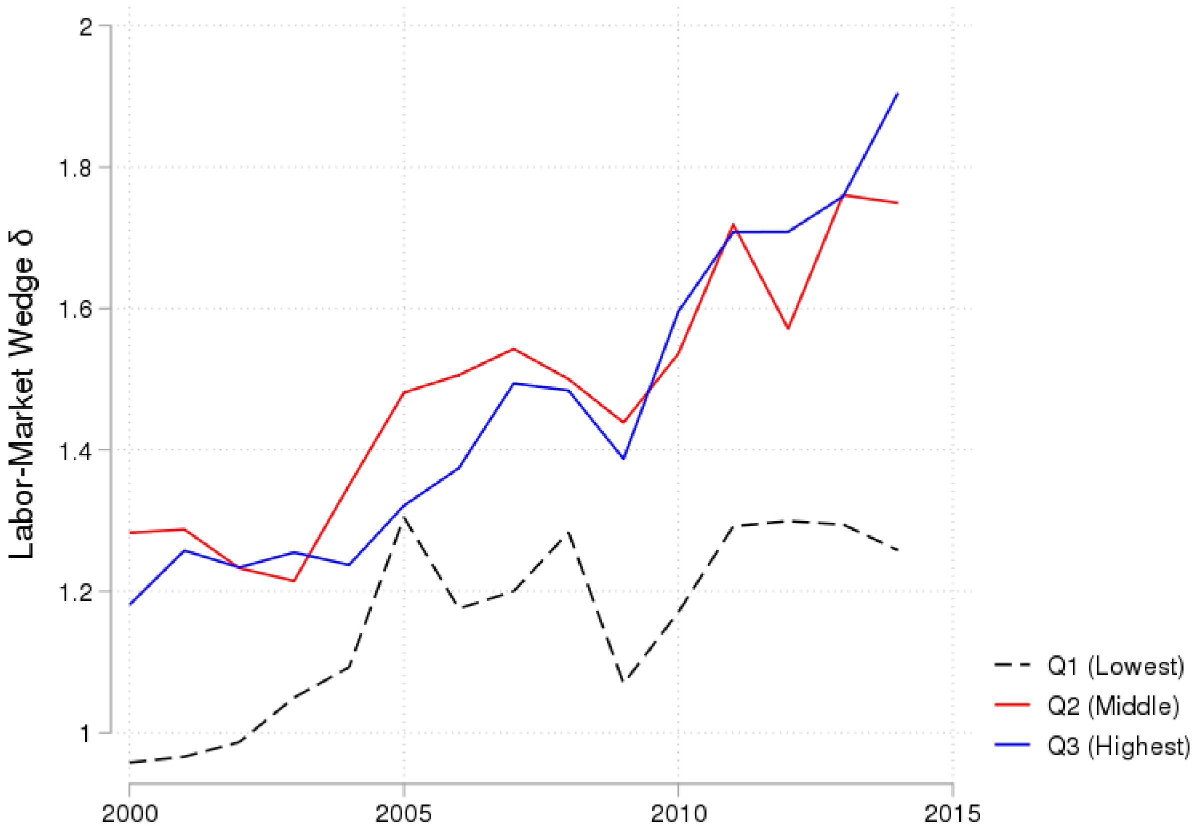
We see the results in this section as strong evidence for a technology-centered explanation of the increase in the manufacturing-wide labor-market wedge in Section 1.5.1. Direct measures of plant-level technological intensity are strongly positively correlated with plants' labor-market wedges both in the cross-section and in changes. This correlation is not driven by outliers, and holds with a range of fixed-effects. It is also robust to different measures of plant size in computing intensity, as well as to running the analysis in levels of technology spending. Finally, plants' levels of technological spending predict their positions in the distribution of labor-market wedges in subsequent years. In short, technological investment correlates strongly with labor-market wedges.

Indirect Technology: Input Intensity

In addition to the direct measures of technological intensity in the previous section, plants' labor-market wedges are positively correlated with measures of technology-correlated input intensity. In particular, we examine whether plants which more intensively use capital or nonproduction labor have higher labor-market wedges. More technologically advanced plants are likely to have more managerial inputs (Atalay et al. (2014)) and more capital.

We regress plants' labor-market wedge on their nonproduction intensity and capital intensity,

Figure 1.10: Evolution of Labor-Market Wedge by 2000 Tercile of New Computer Expenditures Per Worker



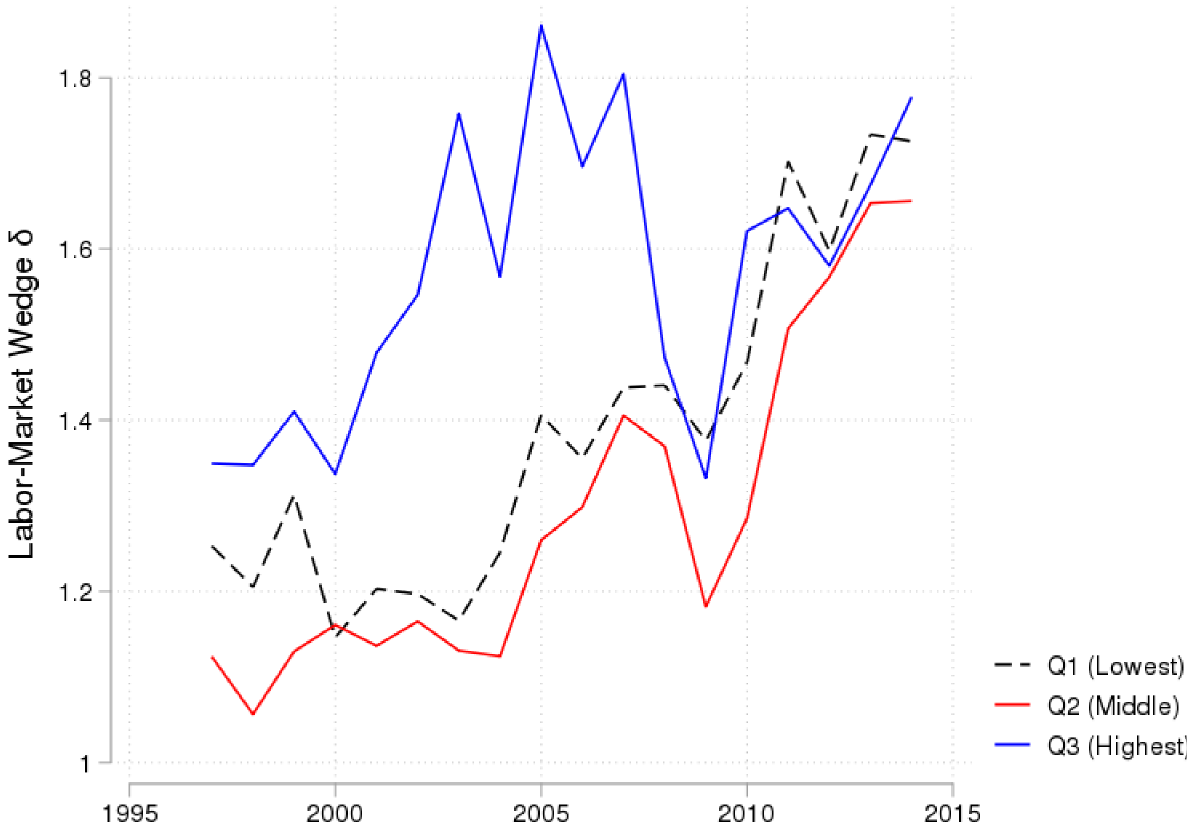
Notes: 2000-2014 evolution of labor-market wedge δ after 2000 by tercile of (real new computer expenditures) / (worker). Plants are sorted into three terciles by per-worker computer expenditures, then subsequent evolution of average δ is plotted within each 2000 tercile. Workers are defined as production workers plus nonproduction workers. New computer expenditures are deflated using a NAICS6 NBER-CES shipments price index. Terciles are weighted by ASM sampling weight and labor expenditures. Beginning year is 2000 due to data availability

both individually and jointly:

$$\log(\delta_{ijt}) = \alpha + \gamma_j + \gamma_t + \beta Int_{ijt} + \epsilon_{ijt}$$

Where $Int \in \{\log\left(\frac{K}{N+L}\right), \log\left(\frac{N}{N+L}\right)$ is alternately the capital or nonproduction intensity.

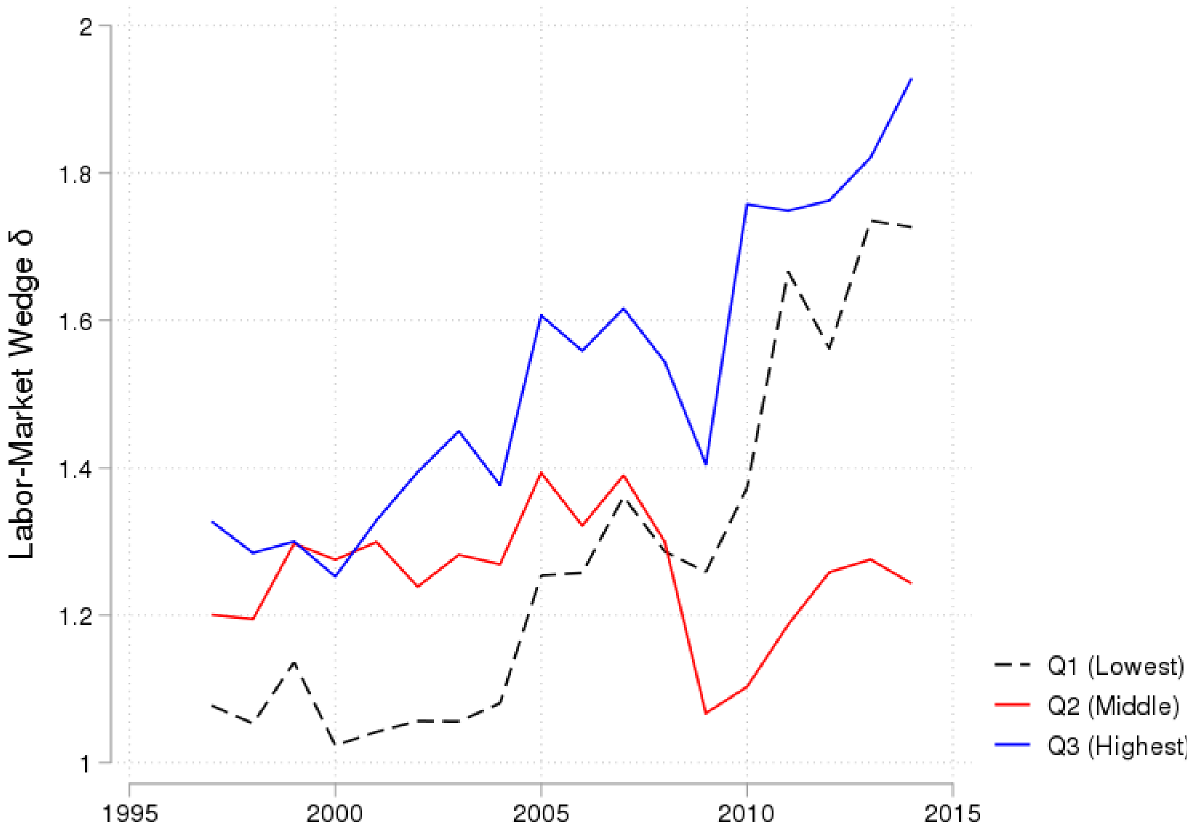
Figure 1.11: Evolution of Labor-Market Wedge by 1997 Tercile of Software Expenditures Per Worker



Notes: 1997-2014 evolution of labor-market wedge δ after 1997 by tercile of (real cost of software) / (worker). Plants are sorted into three terciles by per-worker software expenditures, then subsequent evolution of average δ is plotted within each 1997 tercile. Workers are defined as production workers plus nonproduction workers. Software expenditures are deflated using a NAICS6 NBER-CES shipments price index.Terciles are weighted by ASM sampling weight and labor expenditures. Beginning year is 1997 due to data availability.

K is a plants' capital stock, and N is its number of managers (nonproduction workers). We include year and NAICS6 fixed-effects as in the previous section 1.5.3. Also as in our previous analysis, we normalize capital stocks and nonproduction workers by total plant employment $N + L$.

Figure 1.12: Evolution of Labor-Market Wedge by 1997 Tercile of Communications Expenditures Per Worker



Notes: 1997-2014 evolution of labor-market wedge δ after 1997 by tercile of (real cost of purchased communications) / (worker). Plants are sorted into three terciles by per-worker communications expenditures, then subsequent evolution of average δ is plotted within each 1997 tercile. Workers are defined as production workers plus nonproduction workers. Communications purchases are deflated using a NAICS6 NBER-CES shipments price index. Figure is weighted by ASM sampling weight and labor expenditures. Beginning year is 1997 due to data availability.

Table 1.6 shows that higher nonproduction or capital intensity correlates strongly with a higher labor-market wedge at the plant level. The elasticity of δ with respect to either intensity is approximately 0.2. When considered jointly (column 3 of Table 1.6), the elasticity of δ with respect to either measure is approximately 0.17. These are significant effects both statistically and economically. Figure 1.13 and Figure 1.14 show that these correlations are not driven by outliers, but hold across the range of δ and each intensity. The results in Table 1.6 are robust to using revenues as a measure of firm size, and to exlcluding the $L+N$ normalization. Appendix E shows that they are also robust to including different sets of fixed effects.

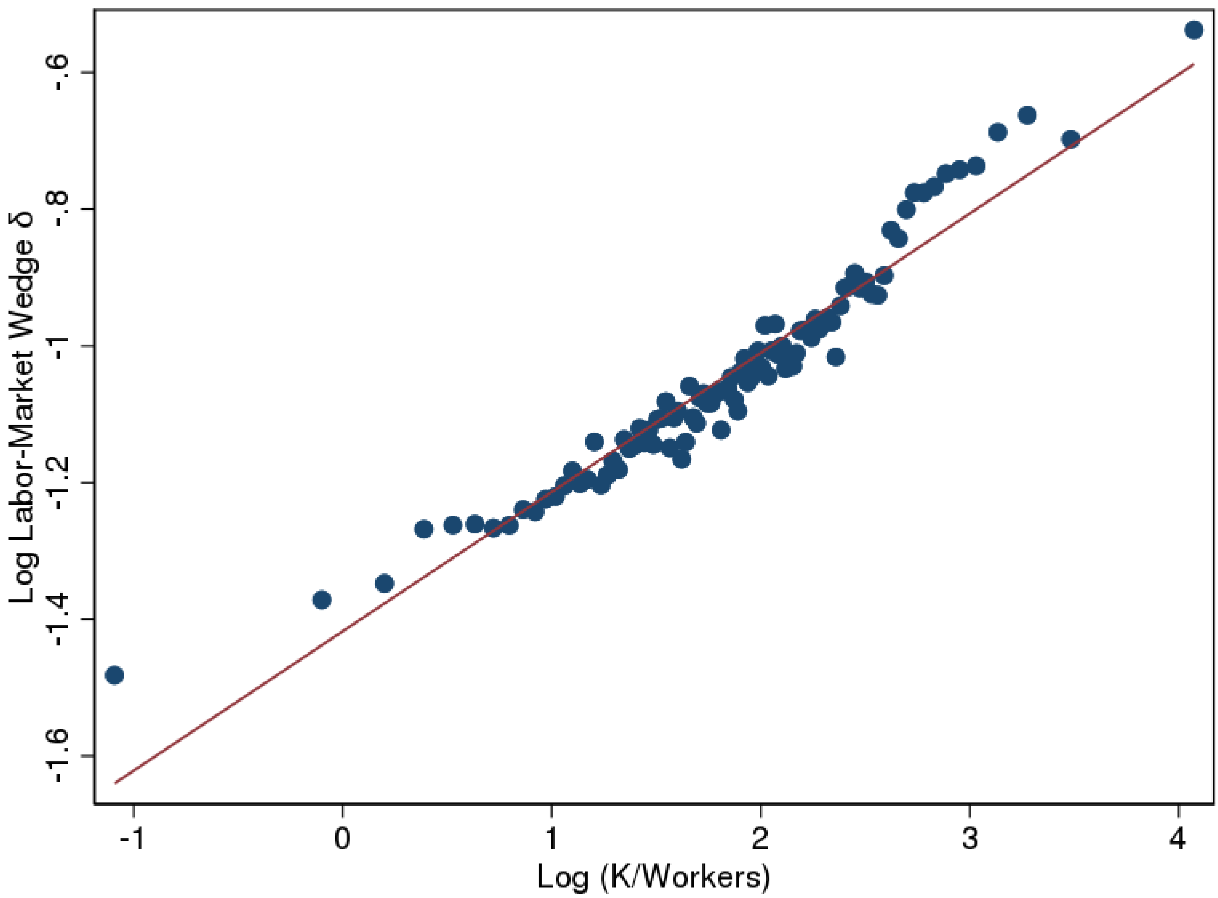
Table 1.6: Regression of Labor-Market Wedge on Input Intensity Measures

	Year, NAICS6 FEs		
	(1)	(2)	(3)
	$\log(\delta)$	$\log(\delta)$	$\log(\delta)$
$\log\left(\frac{K}{N+L}\right)$	0.204***		0.168***
	(0.00112)		(0.00114)
$\log\left(\frac{N}{N+L}\right)$		0.213***	0.168***
		(0.00127)	(0.00129)
N	646,000	646,000	646,000
R^2	0.813	0.812	0.818
AIC	1,360,800	1,365,000	1,344,000

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$
Labor-expenditure- and ASM-sampling-weight- weighted regression of log labor-market wedge δ on log capital intensity (column 1), log nonproduction intensity (column 2), and both intensities jointly (column 3). Input intensity in each case is defined as $\frac{X}{N+L}$, with X the input in question and $N+L$ total plant employment.
Source: ASM/CMF.

Thus plants with greater managerial or capital intensity likely also have higher labor-market wedges. These intensities in turn are outcomes of technological change: plants which invest in new technologies tend to have more capital and more managers. Technological progress is associated with increased intangibles and organizational complexity at the plant level. Therefore we see this section as providing robust supporting evidence to the technology story in the previous section 1.5.3. Higher technological investment leads directly to more spending on computers, software, and communications, which are associated with greater δ. Such investment also leads to capital deepening and greater managerial capital, which are also positively correlated with δ. Technological change, in other words, is positively correlated with plants' labor market wedges both directly and indirectly.

Figure 1.13: Binscatter of Labor-Market Wedge against Capital Intensity

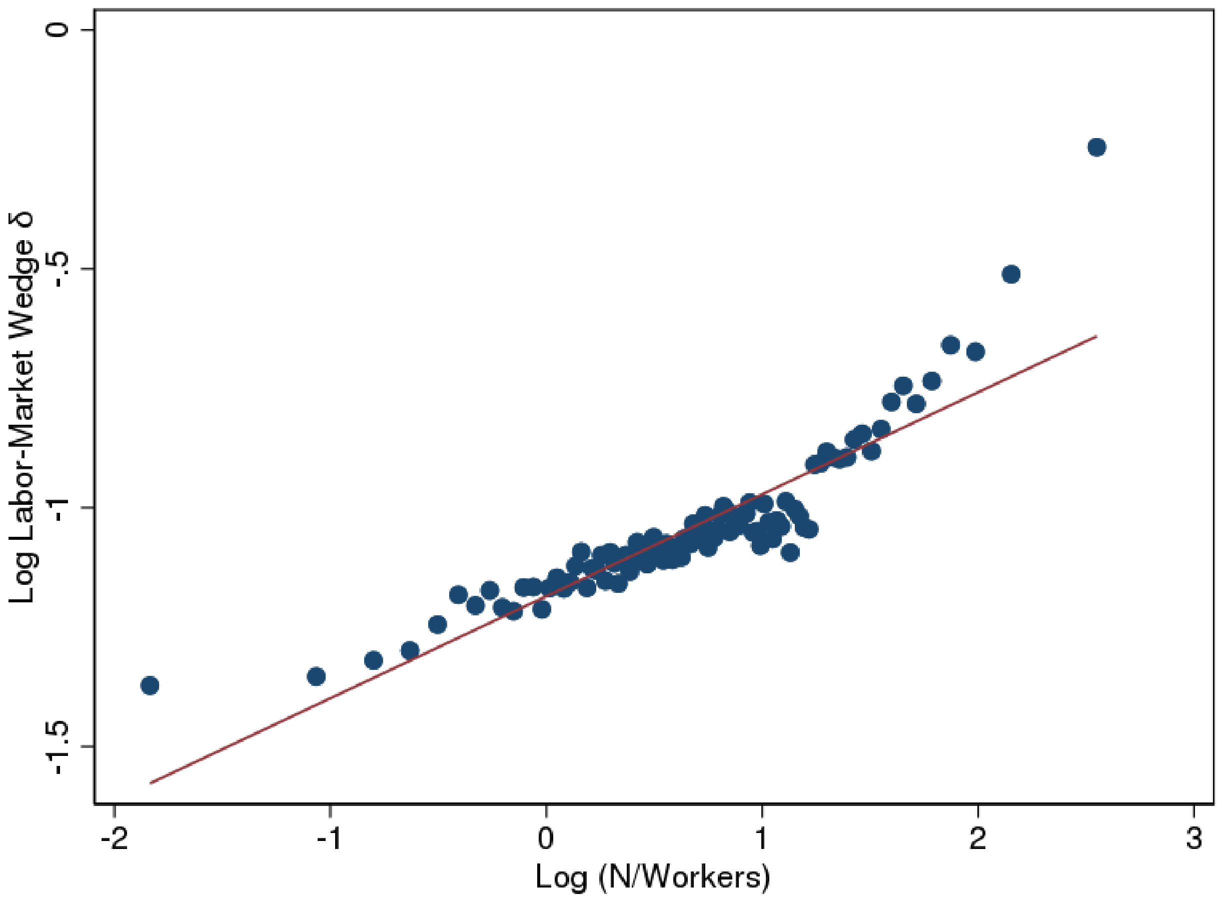


Notes: Binscatter of $\log(\delta)$ against log capital per worker $\log\left(\frac{K}{N+L}\right)$ at the plant level with year & NAICS6 (jt) fixed effects. The figure residualizes $\log(\delta)$ and per-worker capital intensity against year and NAICS6 fixed-effects, then bins each into 100 equally-sized bins and plots the mean of $\log(\delta)$ within each bin along with a (red) line-of-best-fit. Workers are defined as production workers plus nonproduction workers. Figure is weighted by ASM sampling weight and labor expenditures.

Concentration

We now investigate the relationship between labor-market wedges and concentration. Do more concentrated labor markets imply greater δ? A simple nested CES model of labor supply implies that firms which are large in their local labor markets have greater markdowns (see Appendix D). Under static Cournot competition in the labor market (Benmelech et al. (2018), Azar et al. (2020), Berger et al. (2019)), the labor HHI is proportional to the

Figure 1.14: Binscatter of Labor-Market Wedge against Nonproduction Intensity

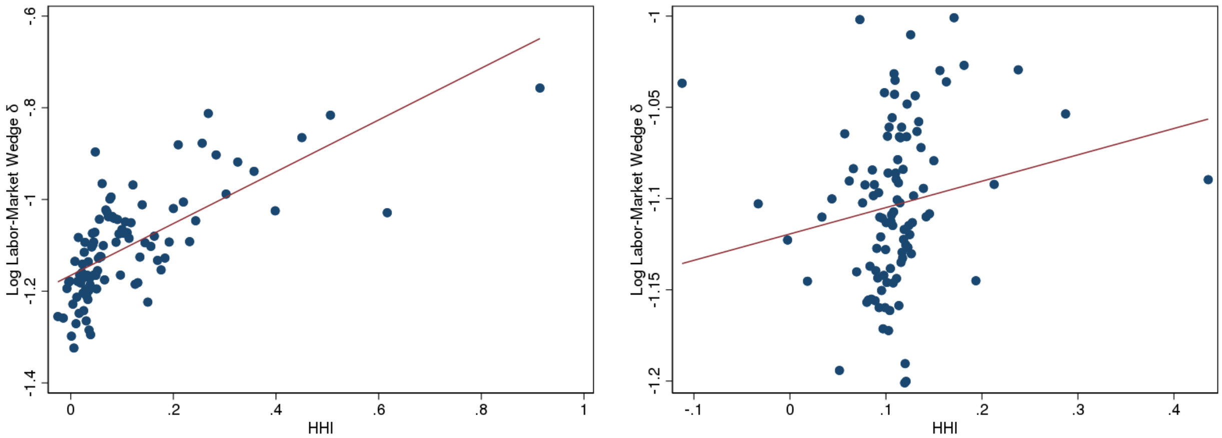


Notes: Binscatter of $\log(\delta)$ against log nonproduction workers per worker $\log\left(\frac{N}{N+L}\right)$ at the plant level with year & NAICS6 (jt) fixed effects. The figure residualizes $\log(\delta)$ and per-worker nonproduction intensity against NAICS6 fixed-effects, then bins each into 100 equally-sized bins and plots the mean of the $\log(\delta)$ within each bin along with a (red) line-of-best-fit. Workers are defined as production workers plus nonproduction workers. Figure is weighted by ASM sampling weight and labor expenditures.

wage-MRPL margin.

In our sample of US manufacturing plants from 1973 to 2014, we find that labor market concentration and wage markdowns are only weakly related. As a measure of concentration, we compute a labor quantity HHI as the sum of the squared share of production workers within

Figure 1.15: Binscatters of Labor-Market Wedge against Labor HHI



Year FEs

Year, CZ FEs

Notes: Binscatters of $\log(\delta)$ against labor quantity HHI at the commuting-zone level with year (t, left) and year, CZ (ct, right) fixed effects. Labor quantity HHI takes the sum of the squared shares of production employees within each commuting zone and year. Each figure then residualizes $\log(\delta)$ and the labor HHI against t or ct fixed-effects, then bins the labor HHI into 100 equally-sized bins and plots the mean of $\log(delta)$ within each bin along with a (red) line-of-best-fit. Figure is weighted by ASM sampling weight and labor quantities.

each commuting zone-year: $\sum_{i \in ct} \left(\frac{\ell_{ict}}{\mathcal{L}_{ct}} \right)^2$[33]. We use a commuting zone as an approximate measure of a local labor market. In Figure 1.15, we plot each CZ-year HHI against its average labor-market wedge[34] with year (left) and year, CZ (right) fixed effects. We find a weak but positive cross-sectional relationship between labor HHIs and δ. Past an HHI of approximately 0.2, there is no meaningful relationship between concentration and the labor-market wege. Within the entire sample, the relationship is weak. Meanwhile, the right panel of Figure 1.15 shows that changes in a commuting zone's HHI are approximately uncorrelated with changes in its average labor-market wedge (including CZ fixed effects implies that the regression is in changes). Thus, labor markets which become more concentrated do *not* increase their labor-market wedges, on average. This suggests that the workhorse assumptions behind

33. Here $\mathcal{L}_{ct}$ is the total production labor within a CZ-year.

34. We use labor *quantity* weights for this average, since the HHI uses quantity shares. The results below are robust to using labor expenditure weights.

some of the labor-market power literature might need reevaluation. Concentration may not be an appropriate measure of labor market power: the relationship between the two is not monotonic and increases in concentration do not correlate with increases in labor-market power.

1.5.4 Macro Sources of Labor-Market Power

We now turn to macroeconomic determinants of labor-market wedges. The rise in the aggregate wedge coincides with a fall in manufacturing unionization, the trade opening to China, and significant changes in manufacturing technology. How important are these factors in driving δ? In this section, we investigate this question using aggregate data from the American Community Survey (ACS) and Current Population Survey (CPS), as well as data on job tasks from Autor and Dorn (2013). We aggregate our data up to the industry level, then merge in our demographic and task data (for details, see Section 1.4). We consider several explanatory variables of interest. First, we consider the fraction of prime-aged workers in the labor force which are unionized as a potential driver of labor-market wedges[35]. Second, the fraction of prime-aged workers in the labor force who work in computer occupations[36] as a proxy for technological change. Third, we consider an indicator for being in the top third of routine-task intensity as defined by Autor and Dorn (2013): this takes the value 1 if an industry is particularly intensive in routine tasks, and 0 if not[37]. Routine tasks are moderately-skilled, repeatable tasks which are easily automated (for instance bank tellers; Autor and Dorn (2013)). Routine-task intensity therefore measures susceptibility to technological change. Finally, we include an indicator variable for top-third of task offshorability. We see these explanatory variables are broadly reflecting institutional

35. As in Section 1.4, we do not limit our analysis to the manufacturing sector only, because workers may substitute between sectors. Our results are robust to limiting our macro variables to the manufacturing sector.

36. Computer occupations are defined as occupation codes 64 and 299. See Appendix A for details.

37. See Appendix A for details about the construction of this variable.

and technological changes in labor-market conditions. Increases in the labor-market wedge coincide temporally with both the early China shock and the ICT/automation revolution. The secular nature of the rise in δ also suggests a role for changing unionization trends. Our macro analysis runs these explanatory factors against each other.

We exploit sub-national variation in labor-market wedges to regress δ on our macro measures at the industry (here defined with ind1990 Census industry codes) level, including year fixed-effects to control for overall changes:

$$\log \delta_{jt} = \alpha + \gamma_t + \beta_1(computer\ occ._{jt}) + \beta_2(union\ member_{jt}) + \beta_3 RTH_{jt} + \beta_4 OSH_{jt} + \epsilon_{jt} \tag{1.13}$$

Where γ_t is a year fixed-effect, RTH is the fraction of highly-routine tasks (top third) within an industry, and OSH is the fraction of highly-offshorable tasks within an industry (top third)[38]. We run (1.13) for years 1980, 1990, 2000, and 2010, pooling the two years before and after each decadal year together. We do this to reduce measurement error and due to data availability with the ACS sample. The regressions are weighted by labor expenditure weight (for the δ aggregation) and the individual observation weight `perwt` in the ACS/CPS (for the macro variable aggregation; see Appendix A for details).

Table 1.7 shows the results. Column 1 includes all explanatory variables, while Columns 2-5 include each variable separately. There are two main points to take from these results. First, industries with greater union membership do not have lower labor-market wedges; indeed, more-unionized industries have *larger* δ. Higher unionization rates may be a response to rising δ, or automation may have proceeded faster in these industries. Therefore deunionization is not likely to be a significant cause of the rise in the aggregate wedge. Second, industries with higher routine-task intensity have lower labor-market wedges. A one percentage-point

38. RTH and OSH are indicators aggregated with observation weights, so represent fractions.

decrease in an industry's routine-intensive occupations leads to a 2.3% fall in its average labor-market wedge. This is consistent with a 'hollowing-out' narrative (Autor and Dorn (2013)) in which automation and ICT reduces demand for routine (intermediate-skilled) jobs while increasing demand for manual (low-skilled) and abstract (high-skilled) jobs. Industries with fewer routine jobs are therefore likely to have adopted more ICT technology and have higher labor-market wedges.

Table 1.7: Regression of Labor-Market Wedge on Industry Aggregates

			Year FEs		
	(1)	(2)	(3)	(4)	(5)
	$\log(\delta)$	$\log(\delta)$	$\log(\delta)$	$\log(\delta)$	$\log(\delta)$
Frac. Computer Occs.	2.719	-2.296			
	(2.668)	2.673)			
Frac. Union Member	0.965*		1.20**		
	(0.437)		(0.405)		
RTH	-2.257***			-2.492***	
	(0.496)			(0.402)	
OSH	0.000375				0.0348
	(0.429)				(0.361)
(Constant)	0.501	-0.159*	-0.445***	0.841***	-0.199*
	(0.293)	(0.066)	(0.118)	(0.174)	(0.0982)
N	200	280	200	280	280
R^2	0.151	0.026	0.06	0.143	0.024
$R^2_{adj.}$	0.125	0.012	0.046	0.131	0.009
AIC	510	746	525	710	747

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$
Industry-level labor-expenditure- and perwt-weighted regression of log labor-market wedge δ on the fraction of workers in computer-based occupations, the fraction of prime-aged workers in the labor force which are unionized, and indicators for high routine-task intensity (RTH) and offshorability (OTH) by task content. RTH and OTH are 1 if in top third of routine-task intensity or offshorability, and 0 otherwise. Computer occupations are defined as `occ1990` occupation codes 64 and 229. Task measures are from Autor and Dorn (2013). Sample is limited to prime-aged (25-64) workers in the manufacturing labor force. Includes years 1980, 1990, 2000, and 2010, pooling each decadal year and the two years before and after together. All variables aggregated to industry (ind1990) and year level.
Union data not available for 1980.
Source: ASM/CMF, IPUMS (ACS/CPS), Autor and Dorn (2013).

Table 1.8 shows the same regression in long-differences; it aggregates the data to the industry and year level, take differences between 1990 and 2010, and run the first-differences regression:

$$\Delta \log \delta_{jt} = \alpha + \beta_1(\Delta computer\ occ._{jt}) + \beta_2(\Delta union\ member_{jt})$$
$$+ \beta_3 \Delta RTH_{jt} + \beta_4 \Delta OSH_{jt} + \epsilon_{jt}$$

Where the difference operator is between 1990 and 2010 (we choose these decadal years for ease of comparison with the cross-sectional analysis above). Industries which were more likely to move out of the top-third of routine-task intensity saw their labor-market wedges decrease (Columns 1 and 4). The semi-elasticity of changes in δ with respect to high-routine-task status is approximately -3.8: a 1 percentage point reduction in routine occupations in an industry decreases its labor-market wedge by 3.8%. In changes, higher union membership is weakly associated with a lower labor-market wedge. These results are robust to alternative weighting[39].

Thus our aggregated analysis provides further suggestive evidence of the technological origins of the rise in the manufacturing labor-market wedge. ICT innovations eliminate manual-intensive jobs while creating routine- and abstract-intensitve jobs, and we find that this pattern is correlated positively with the labor-market wedge. Moreover, unionization rates are not cross-sectionally negatively correlated with δ, which suggests that institutional and legal factors are less important than technological change in creating labor-market wedges.

1.5.5 Discussion & Mechanism

The manufacturing labor-market wedge increased from approximately 1 in 1973 to 2.1 in 2014, with upward inflection points around 1990 and 2002. This wedge reflects marginal

39. In unreported results, we tried weighting δ by number of production workers with similar results.

Table 1.8: Long-Differences Regression of Labor-Market Wedge on Industry Aggregates

	(1)	(2)	(3)	(4)	(5)
	$\Delta \log(\delta)$	$\Delta \log(\delta)$	$\Delta \log(\delta)$	$\Delta \log(\delta)$	$\Delta \log(\delta)$
Δ *Frac. Computer Occs.*	-7.542	-6.697			
	(3.667)	(3.82)			
Δ *Frac. Union Member*	-1.664**		-1.651**		
	(0.535)		(0.56)		
Δ *RTH*	-3.772***			-2.47	
	(1.363)			(1.391)	
Δ *OSH*	-0.756				-1.036
	(0.824)				(0.885)
(Constant)	0.0141	-0.451***	0.0786	0.249***	0.367***
	(0.138)	(0.0818)	(0.115)	(0.0887)	(0.0642)
N	65	70	65	70	70
R^2	0.253	0.043	0.12	0.044	0.02
$R^2_{adj.}$	0.204	0.029	0.106	0.03	0.005
AIC	97	111	102	111	113

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$
1990-2010 long-differenced industry-level labor-expenditure- and perwt-weighted regression of log labor-market wedge δ on the fraction of workers in computer-based occupations, the fraction of prime-aged workers in the labor force which are unionized, and indicators for high routine-task intensity (RTH) and offshorability (OTH) by task content. RTH and OTH are 1 if in top third of routine-task intensity or offshorability, and 0 otherwise. Computer occupations are defined as `occ1990` occupation codes 64 and 229. Task measures are from Autor and Dorn (2013). Sample is limited to prime-aged (25-64) workers in the manufacturing labor force. All variables aggregated to industry (ind1990) and year level for years 1990 and 2010, then differenced.
Union data not available for 1980.
Source: ASM/CMF, IPUMS (ACS/CPS), Autor and Dorn (2013).

revenue products rising faster than wages. It is driven by new plants and, after 2002, by rises in the individual wedges of incumbent plants. And it is largely a within-industry (rather than within-geography) phenomenon. These facts suggest a technological explanation for the rise in the labor-market wedge: the rise in the wedge coincides with the timing of the ICT shock.

Indeed, we find that plants which spent more on technology in the late 1990s had higher levels and greater increases in their labor-market wedges. Plants with more physical and managerial capital relative to their workforce have higher wedges. At the aggregate level, we find that higher wedges are associated with greater intensity of manual and abstract jobs, consistent with the idea that computer-oriented technological changes hollows out routine jobs but boosts the lower and upper ends of the skill distribution. We find no evidence that unionization plays a significant role in rising labor-market wedges, and we find that geography-specific variables do not explain much of the variation in these wedges.

Therefore, we find that changing technology in the manufacturing sector lead to rising labor market power. As firms invest in technologies, they improve labor productivity, which creates a surplus. However, workers' wages are not determined solely by this surplus but are also mediated through local labor markets. Thus, workers do not capture much of the value created from technological change. In a Nash bargaining framework, we can view this as an increase in the employer's outside option or a decrease in worker bargaining power. In either case, technological change creates additional surplus but also creates the conditions for firms to capture most of it.

What is the specific mechanism by which technology enables firms to capture most of the additional surplus? We have seen (Figure 1.15) that changes in concentration alone do not predict changes in the labor-market wedge. But perhaps technological change polarizes local labor markets. This, combined with reductions in worker mobility, reduce workers' outside options or bargaining power and enable firms to capture most of the value created by recent technological progress. That is, perhaps the threat of replacement by another worker (or a robot) forces manufacturing workers to accept lower wages at high-tech manufacturing plants. To link firm outside options explicitly to worker choices requires a model of search, which is outside the scope of this paper. Nonetheless, we find - with both timing, direct, and indirect evidence - that the late-1990s technological revolution greatly increased labor market power in manufacturing.

1.5.6 Comparison to Existing Literature

We now benchmark our results against recent literature. Broadly, we find that our labor-market wedges are similar to those implied in several recent papers despite their not directly analyzing labor market power. One line of literature has focused on the impact of automation on labor markets. Acemoglu and Restrepo (2020) find that commuting zones with greater robot penetration saw lower wages and reduced employment. Automation (as proxied by robot penetration) is a close cousin of the technological change we study. Similarly, Acemoglu and Restrepo (2021) find that automation technologies displace workers with routine skills - and that neither deunionization nor markup power significantly increased worker displacement. This aligns closely with our findings, although Acemoglu and Restrepo (2021) do not directly study markdowns and use a different measure of markups. Similarly, Autor and Dorn (2013) find that automation reduces routine jobs and polarizes local labor markets. We expand this analysis by showing that automation technologies also increase labor-market power: even workers who remained employed after the large technological change captured less of job surplus.

Another line of literature focuses on labor-market rents and rent-sharing as a proxy for imperfect competition. These studies typically fit employment and wage data to models of preferences, worker mobility, and wage determination to quantify the extent to which firms' labor supply curves are imperfectly elastic. Lamadon et al. (2020) use matched worker-firm tax data to show that there is significant imperfect competition in the US labor market. In particular, they find inverse labor-supply elasticities of between 0.17 and 0.23. Our estimate of δ implies an inverse labor supply elasticity $\nu = \delta - 1$, so our estimates are somewhat larger than those in Lamadon et al. (2020). However, they study all sectors; labor-market wedges are plausibly higher in manufacturing[40]. Our estimates of δ are not directly comparable to rent-sharing elasticities, but Card et al. (2018) survey recent research

40. For instance, the labor share of value added is lower in the manufacturing sector.

and find that rent-sharing elasticities on the order of 0.1-0.2. These imply imperfect labor-market competition, though likely slightly lower labor-market wedges than we find. The precise relationship between rent-sharing elasticities and labor-market wedges depends upon the nature of production and the structure of labor markets[41].

A third strand of empirical literature focuses on running regressions of industry level Herfindahl-Hirschman indices on wages or wage growth. HHI most directly measures market power in a symmetric Cournot model. In such a model, it can be shown that:[42]

$$\frac{MRPL - w}{w} = HHI \cdot (\nu) \tag{1.14}$$

Where ν is the inverse of the labor-supply elasticity. Solving for w and taking logs yields:

$$\log w = \log MRPL - \log(1 + HHI \cdot \nu) \tag{1.15}$$

Differentiating this with respect to HHI yields:

$$\frac{\partial \log w}{\partial HHI} = \frac{\nu}{\nu \cdot (HHI) + 1} \approx \nu$$

Benmelech et al. (2018) (BBK) find that $\frac{\partial \log w}{\partial HHI} \approx -0.025$ in the US manufacturing industry from 1977-2009. This implies that $\nu \approx 0.975 \Rightarrow \delta \approx 2$. We estimate a monopsony wedge between 1.1 and 1.6 over this period. This is a slightly smaller wedge to that implied by BBK (though this comparison is imprecise since they do not specify a model; we heuristically back out the markdown implied by their HHI-wage relationship).

BBK also find that the relationship between HHI and wages has become more negative over time. They estimate that this 'concentration effect' grows 1.6-fold from 1977 to 2009 and

41. For example, with a Cobb-Douglas production function, an increase in TFP would be shared between L and K depending on their relative contribution to output even absent any labor-market power.

42. This is the input market analogy to the typical Cournot output market analysis.

twofold from 1991 to 2009 within 3-digit SIC industries in manufacturing. Our measure of δ grows by 59 percent from 1972 to 2011, which is roughly in line with the increase in monopsony power implied by BBK.

Azar et al. (2020) (AMS) find that $\frac{\partial \log w}{\partial \log HHI} \approx -0.141$ in the US services industry from 2010-2013. Differentiating (1.15) with respect to $\log HHI$ yields:

$$\frac{\partial \log w}{\partial \log HHI} = \frac{\nu \cdot HHI}{\nu \cdot HHI + 1}$$

Inserting the AMS estimate yields:

$$\nu \approx \frac{-0.14}{1.14} \cdot \frac{1}{HHI}$$

Inserting their appropriately-weighted mean HHI yields $\nu \approx 0.39$, and therefore $\delta \approx 1.4$, which is close to our estimates. Note that we focus on the manufacturing sector rather than the services sector.

Finally, Berger et al. (2019) show that local labor-market HHIs in US manufacturing have fallen since 1977. They use a nested CES model of labor-market competition to translate falling HHIs into falling labor-market wedges (see Appendix D). Our estimated labor-market wedge δ is not correlated with local HHI in changes (Figure 1.15). We obtain our wedges from a production model, rather than inferring them from concentration measures.

In sum, we find labor input wedges close to those implied by the HHI-focused BBK and AMS despite the fact that neither paper directly analyzes markdowns or labor wedges. We find signfiicant evidence of imperfect competition in labor markets, similar to Lamadon et al. (2020) and the literature surveyed in Card et al. (2018). We find strong evidence that changing technology increases labor-market wedges, in line with Acemoglu and Restrepo (2020) and Acemoglu and Restrepo (2021).

1.6 Conclusion

How important is market power in US manufacturing? We found that labor market power in US manufacturing is large and has risen substantially: on average, manufacturing firms pay production workers around 48% of their marginal revenue products in 2014. The rise in the labor-market wedge was especially sharp in the late 1990s and early 2000s, and came almost exclusively from new manufacturing plants. Oligopsonistic power decreases plant employment and, to a lesser extent, wages. It accounts for as much as two-thirds of the decrease in the manufacturing labor share.

We found that plants' labor-market wedges are strongly correlated with technology-related expenditures on computers, software, and communications in both levels and changes. At the commuting-zone and industry level, higher labor-market power is positively correlated with higher manual and abstract tasks, but not correlated with union membership. We conclude that technological change was a significant contributing factor to the increase in manufacturing labor-market wedges.

Finally, we developed a flexible technique for estimating labor-market wedges. This method, which uses a modified proxy-variable framework, enables researchers to estimate arbitrary numbers of input- and output-market wedges so long as one input is competitive.

Our results suggest two main conclusions. First, labor-market power is a very important (indeed, the dominant) component of an aggregate market-power wedge. Rather than rising markups enabling monopolistic firms to raise prices and underutilize labor inputs, we explored an alternative possibility: that market power is a composite of output and input wedges; of markup and markdown power. Of the two, markdowns are much the more importnat. We found that manufacturing labor-market wedges have doubled since 1973, while markup wedges were approximately flat. Labor-market power is widespread in the manufacturing sector. This explains the sharp fall in the manufacturing labor share and rise in the profit share. Second, labor-market power arises in large part due to the adoption of automation technology. Changing technology enables firms to capture most of the surplus from increased labor productivity and TFP. This suggests that there may be scope for policy action to counteract technology-driven rises in monopsony power.

Many questions remain about the nature of labor-market power. How does it interact with skill-biased technological change? How has it evolved in other sectors? Can its rise be corroborated in other quasi-experimental settings? What are its welfare implications? We leave these questions to future research.

Chapter 2

WHAT DO PRODUCTION-BASED MARKUP ESTIMATORS ACTUALLY MEASURE?

2.1 Introduction

Markets efficiently allocate goods and services only when there is sufficient competition. For researchers and policymakers alike, the nature of frictions to competitive markets matters. If market power is mainly in the output market, then consumers experience welfare losses. On the other hand, if the main source of competitive frictions is in the labor market, then the welfare of workers is reduced. Finally, if adjustment costs are the main source of friction, then the structure of search markets may matter more for welfare. These three cases have distinct implications for welfare, economic modeling, and competitive regulation and policy. In this paper, we argue that researchers and policymakers should broadly reinterpret existing production-based markup estimates as input-market frictions.

While output market power represents a wedge between prices and marginal costs, input market power[1] is a wedge between input costs and marginal input revenues. Current production-based ratio estimators estimate from the ratio of marginal cost to average cost: $\mathcal{M} = \frac{\theta^L}{\frac{WL}{R}}$. Following Bond et al. (2020), we call the empirical implementation of this approach the ratio estimator. Average cost is generally observable, but marginal cost is not. Empirical papers rely on different assumptions and approximations to recover marginal cost, most commonly by estimating a production function. However, empirical application of the ratio estimator is often limited by observable data. Researchers typically do not observe physical output and must rely on revenues as a proxy. In this paper, we show that using revenues to proxy

1. Often called 'monopsony power,' though the term is somewhat of a misnomer as it implies a single buyer. Input market power can exist just as well with few or many buyers, just as output market power can exist with oligopoly or monopolistic competition.

for physical output means that the ratio estimator measures distortions in input markets rather than output markets.

We first show the result in theory by working through a firm's optimization problem. We infer a composite output-market and input-market wedge from the difference between average and marginal cost. With revenue data the output wedge exactly cancels, leaving only the input wedge. In particular, our results suggest that the non-markup portion of the composite wedge is better interpreted as input market power rather than a nuisance parameter implied by misspecification. We then apply our result in practice to show that past research suggests that trade shocks likely induce rises in labor market power or adjustment costs. Our method suggests that several results in the literature can be reinterpreted as implications on input-market wedges.

The difficulty in estimating market power with revenue data has a rich history in the industrial organization and international trade literature. Klette and Griliches (1996) first emphasized the problem: absent data on prices, one cannot disentangle whether an increase in revenue is due to an increase in market power or an increase in physical productivity[2]. In other words, because markups distort both prices and input use, the ratio of prices to inputs provide no information about output market power when only revenue data is available[3]. Recent work has revisited the Klette and Griliches (1996) critique in the context of the ratio estimators and shown that these estimators do not recover markups (Jaumandreu (2018) and Bond et al. (2020)).

We build on this result by showing that ratio estimators applied to revenue data instead recover input market wedges. We show that, since the distortions discussed in Klette and Griliches (1996) apply symmetrically to all flexible inputs, the ratio estimator recovers the extent to which a firm under-utilizes a monopsonistic input relative to a competitively-

2. Competitive settings rule out the first possibility.

3. That is, $R = P(Q)Q$ in imperfectly competitive settings.

supplied one. If this under-utilized input is subject to static monopsonistic distortions in its input market, then the revenue ratio estimator recovers an input wedge (rather than a markup wedge). Alternatively, these input wedges may be interpreted as adjustment costs or factor-augmenting technology depending on the empirical setting. This result has two important implications. First, existing work using ratio estimators to measure markups should be reinterpreted as measuring input market power. For instance, De Loecker et al. (2020) discuss rising markups across the U.S. and global economies. Our results suggest that, to the extent that they use revenue elasticities, their paper can be interpreted as measuring input market power. Raval (2020) compares markups from different inputs to test the consistency of the ratio estimator; but our paper implies that this compares input market distortions. The second implication of our findings is that there is much scope for measuring important market frictions. For example, several recent papers (Dobbelaere and Mairesse (2013), Lamadon et al. (2020), and Card et al. (2018)) have focused on the importance and implications of labor market power. This paper shows the consistency of these estimates even absent data on output prices. It also suggests potentially fruitful avenues for future research on labor-market imperfections.

Much other recent research examines whether rising market power can explain changes in economic dynamism and the fall in the labor share of income (Card et al. (2018), Berger et al. (2019), Dobbelaere and Mairesse (2013), and Morlacco (2020)). For example, Karabarbounis and Neiman (2013) and Elsby et al. (2013) connect rising industry concentration to falling labor shares since the early 2000s. De Loecker et al. (2020) suggest a similar connection. Evaluating these competition hypotheses requires proper measurement of market power. In this paper, we show that researchers and policymakers can reinterpret a common estimator of pricing power as measuring input market power when using revenue data.

2.2 Ratio Estimators with Revenue Data

In this section, we discuss the interpretation of ratio estimators of market power in the context of revenue data. We show that, when firms face imperfect competition in input markets, the ratio estimator is the correct expression for *input* market power. By contrast, as emphasized in Bond et al. (2020), the ratio estimator says nothing about output market power.

2.2.1 Statement of Problem

Firms produce output Q using an intermediate input L[4]. They face imperfectly elastic supply schedules for L, with an input-price elasticity ϵ^L. They pay a single unit price W per unit of L. Thus L is subject to static monopsonistic input distortions: firms optimally choose to underhire L to keep their inframarginal input costs lower. Firms sell output in imperfectly competitive markets. In particular, the elasticity of output price with respect to quantity is ϵ^P[5].

Firms' production function is given by $Q = F(L)$. Firm revenue is given by $R \equiv Q \cdot P(Q)$, where price depends on quantity because of imperfectly elastic demand.

Define the elasticity of output *quantity* with respect to L as $\theta^L \equiv \frac{\partial F(\cdot)}{\partial L}\frac{L}{Q}$. Similarly, the elasticity of output *revenue* with respect to L is $\theta^L_R \equiv \frac{\partial R(\cdot)}{\partial L}\frac{L}{R}$. Since a single parameter determines the relationship between price and quantity demanded, the two elasticities are related by[6] (see Appendix F for details):

$$\theta^L_R = (1 + \epsilon^P)\theta^L \tag{2.1}$$

4. This analysis may be expanded to arbitrarily many intermediate inputs.

5. This is the inverse of the demand elasticity.

6. To obtain this result, totally differentiate the definition of revenue with respect to L.

Revenue elasticities depend on both physical (production) elasticities and the elasticity of demand. Since demand slopes downward, $\epsilon^P < 0$ and revenue elasticities are always lower then revenue elasticities. In other words, more production implies lower prices, which in turn reduces revenues. Equivalently, define the markup $\mathcal{M} \equiv \frac{P}{MC}$. Cost-minimization implies a markup pricing rule $\mathcal{M} = \frac{1}{1+\epsilon^P}$ (see Appendix F for details). Applying this markup pricing rule to (2.1) implies:

$$\theta_R^L = \frac{\theta^L}{\mathcal{M}} \tag{2.2}$$

So that the markup completely determines the relationship between revenue and quantity output elasticities.

2.2.2 Computing and Interpreting the Ratio Estimator

We assume that firms minimize costs. The cost-minimization problem is:

$$\begin{aligned} \min_L C(L) &= W(L)L \\ s.t. \quad F(L) &= Q \end{aligned} \tag{2.3}$$

The input price $W(L)$ depends on L because of imperfectly elastic input supply. The envelope theorem implies that the Lagrange multiplier of (2.3) is $\lambda \equiv C'(Q)$; that is, the firm's marginal cost is the shadow value of relaxing the cost-minimization constraint. Then the markup $\mathcal{M} \equiv \frac{P}{MC} = \frac{P}{\lambda}$. Now consider the first-order condition of (2.3) with respect to

L:

$$
\begin{aligned}
[L] \qquad W + \frac{\partial W}{\partial L} L - \lambda \frac{\partial F(L)}{\partial L} &= 0 \\
w + \frac{\partial W}{\partial L} L - C'(Q) \frac{\partial F(L)}{\partial L} &= 0 \\
1 + \frac{\partial W}{\partial L} \frac{L}{W} &= C'(Q) \frac{\partial F(L)}{\partial L} \frac{1}{W} \\
1 + (\epsilon^L)^{-1} &= P(1 + \epsilon^P) \frac{\partial F(L)}{\partial L} \frac{1}{W} \\
1 + (\epsilon^L)^{-1} &= (1 + \epsilon^P) \frac{\partial F(L)}{\partial L} \frac{L}{Q} \frac{PQ}{WL} \\
1 + (\epsilon^L)^{-1} &= \theta^L \frac{PQ}{WL} (1 + \epsilon^P) \\
1 + (\epsilon^L)^{-1} &= \theta^L_R \left(\frac{WL}{PQ} \right)^{-1} \\
\delta &= \theta^L_R \frac{PQ}{WL} \qquad (2.4) \\
\delta \mathcal{M} &= \theta^L \frac{PQ}{WL} \qquad (2.5)
\end{aligned}
$$

Where the next-to-last line uses identity (2.1) and the last line uses identity (2.2). The RHS of (2.5) is the commonly-used ratio estimator of markups: the product of a flexible input's quantity output elasticity and its inverse cost ratio. The LHS of (2.4) $1 + (\epsilon^L)^{-1}$ is (one plus) the inverse elasticity of input supply. This object is the monopsonistic analogue to the markup rule $(1 + \epsilon^P) \equiv \mathcal{M}$: it is a *monopsonistic pricing rule*. In other words, $1 + (\epsilon^L)^{-1}$ represents an *input-market wedge* δ.

Typically, researchers assume that input markets are competitive and so $\delta \equiv 1$. Then (2.5) recovers markups $\mathcal{M}$, but only if one can estimate the quantity elasticity θ^L[7]. But this typically requires knowledge of output *quantities*, which are unavailable in most production datasets. For example, the US Census of Manufactures only reports establishment sales R (similarly, Compustat, the Chilean ENIA, and the Colombian and Slovenian manufacturing

7. Typically, this is done with either a proxy-style estimator as in Ackerberg et al. (2015) or a dynamic panel estimator as in Blundell and Bond (1998).

surveys only have revenue data). Ignoring this and using the revenue elasticity θ_R^L leads to misspecification: the markup should then be identically unity (Bond et al. (2020)), which defeats the purpose of estimating market power.

If the supply of L is perfectly elastic, so that δ is identically unity, then (2.5) recovers $\mathcal{M}$. If L is not competitively supplied, then $\delta > 1$ and (2.5) implies that the ratio estimator recovers a composite input- and output-market wedge $\delta \cdot \mathcal{M}$. In this case markup wedges (perhaps the result of market structure, barriers to entry, or innovation disparities) are isomorphic to static input-market wedges (perhaps the result of input-market institutions or upstream market structure).

On the other hand, if one can estimate revenue elasticities θ_R^L, then (2.4) implies that the ratio estimator recovers input-market wedges δ rather than markups (the ratio literature assumes instead that this estimator recovers $\mathcal{M}$). With revenue data and an appropriate estimation strategy, a researcher can estimate oligopsonistic power in input markets. Thus using a ratio estimator with a revenue output elasticity gives researchers important information about the extent of monopsonistic power in input markets.

2.2.3 Interpretation of Input Market Power Estimator

Markups cause firms to reduce output and hence expenditures on inputs. They do this because they face downward-sloping demand curves, so that decreases in price must be applied to more-profitable inframarginal sales. Input expenditures are reduced proportionally to input elasticities for each input. Because the input-market wedge in (2.4) is a ratio, this output-limiting effect of markup power exactly cancels. What remains is a wedge from input use alone.

To see this more precisely, suppose a firm produces output using both a monopsonistic input L and a competitively-supplied input M. Since the supply of M is perfectly elastic,

$(\epsilon^M)^{-1} = 0$, which is to say $\delta^M = 1$. Let the price of M be C. Consider the equivalent of (2.4) for the competitive input M, and substitute into it the expression for $(1 + \epsilon^P)$ using the monopsonistic input L to yield:

$$
\begin{aligned}
1 &= \theta^M \left(\frac{CM}{PQ}\right)^{-1} (1 + \epsilon^P) \\
1 &= \theta^M \left(\frac{CM}{PQ}\right)^{-1} (1 + (\epsilon^L)^{-1})(\theta^L)^{-1} \frac{WL}{PQ} \\
1 + (\epsilon^L)^{-1} &= \left(\frac{\theta^L}{WL}\right) \left(\frac{\theta^M}{CM}\right)^{-1} \\
&= \left(\frac{\theta^L_R}{WL}\right) \left(\frac{\theta^M_R}{CM}\right)^{-1} \qquad (2.6)
\end{aligned}
$$

So that the input-market wedge $\delta \equiv 1 + (\epsilon^L)^{-1}$ reflects the extent to which the firm uses less of imperfectly-competitive input L relative to L's output elasticity, as compared competitive to input M relative to input M's output elasticity. This 'double ratio' formula holds for both revenue and quantity elasticities, since the additional elasticity terms from (2.1) cancel.

Put another way, markup wedges and monopsonistic wedges both cause firms to reduce input use. Markup power induces a reduction in input use across all inputs proportional to their output elasticities. Markdown power induces an additional reduction in input use, but only for a monopsonistic input. The ratio estimator 'nets out' the symmetric quantity reduction and allows the researcher to separately identify input-market power as the extent to which a monopsonistic input is used less than a competitive input, relative to their respective values in production. Markup power alone, however, cannot be identified.

The appropriate formula for input market wedges depends upon the output elasticities available. If the researcher can only esimate revenue elasticities, then (2.4) should be used. If quantity elasticities are available, then (2.6) gives input-market wedges and (2.5) applied to competitively-supplied input M with quantity elasticities gives markup wedges (this is simply the ratio estimator for $\mathcal{M}$).

2.2.4 Extending the Framework

Our analysis extends to an arbitrary number of inputs, so researchers can readily add capital or other productive inputs to the production function. In this case, the cost-minimization problem becomes:

$$\min_{\{X_i\}_{i=1}^N} C(\{X_i\}_{i=1}^N) = \sum_{i=1}^N W_i X_i$$
$$s.t. \quad F(\{X_i\}_{i=1}^N) = Q$$

The FOCs and ratio formulas (2.4) are unchanged. To maintain the interpretation of the ratio estimator (2.4), the firm must have at least one monopsonistically supplied input and at least one competitively supplied input. These inputs must also be flexible, in the sense that they may be freely adjusted before the firm engages in production in each period. If the inputs are not flexible, then the firm cannot adjust input use in response to changes in demand or input elastiticies and we cannot infer markup or markdown power from input use.

Input adjustment costs complicate the interpretation of the monopsony parameter. To the extent that these inputs are subject to static and proportional adjustment costs, the ratio estimator reflects both these adjustment costs and monopsonistic input market power[8]. The estimated monopsony parameter is then $\omega\left(1 + (\epsilon^L)^{-1}\right)$, where ω is an adjustment cost parameter proportional to the quantity of input L. More complicated adjustment cost specifications may also be considered, but are outside the scope of this paper[9].

8. In particular, such adjustment costs must be unpredictable to the firm when planning production.

9. For instance, dynamic adjustment costs require the researcher to specify a firm's dynamic problem, since input choice this period depends on the future sequence of anticipated adjustment costs.

2.3 Reinterpreting Results in the Literature

Our reinterpretation of the ratio estimator has implications for several findings in the literature. We highlight two in this section.

2.3.1 De Loecker & Warzynski: Markups and Trade

De Loecker and Warzynski (2012) pioneered the use of the ratio estimator in the markups literature[10]. They used data on Slovenian manufacturing firms to estimate production functions and markups using the ratio estimator. They then regressed markups on export status to investigate whether firms connected to global supply chains had greater pricing power. The idea was that exporting firms learn from exporting, get a productivity advantage over competitors, and incorporate this into their pricing decisions. Figure 2.1 reprints the main table in De Loecker and Warzynski (2012): they find that export status predicts about 12-18% higher markups.

The data in De Loecker and Warzynski (2012) are revenue data. Supposing that their estimated output elasticities are revenue elasticities, we use (2.4) to conclude that the premia in Figure 2.1 represent *input-market wedges*. For instance, in their Cobb-Douglas specification, we infer that the labor-market wedge increased by about 16% for exporting establishments. This interpretation implies that new exporters use the competitive global market to decrease wages (relative to $MRPL$). Firms may use the threat of closure or increased foreign competition to justify lower wages in negotiations with their unions or employees.

Thus, rather than implying that trade increases productivity and markup power, De Loecker and Warzynski (2012) implies that trade increases competitive pressures and therefore markdown power. Markups may still rise, but to evaluate this effect we would need to estimate quantity elasticities. Nonetheless, the broad conclusion of the paper is significantly

10. Adapting Hall (1988).

Table 3—Markups and Export Status I: Cross-Section

Methodology	Export Premium
Hall	0.0155 (0.010)
Klette	0.0500 (0.090)
Specification	
I (Cobb-Douglas)	0.1633 (0.017)
II (I w/ endog. productivity)	0.1608 (0.017)
IV (Translog)	0.1304 (0.014)
V (II w/ export input)	0.1829 (0.017)
VIII (First difference)	0.1263 (0.013)

Notes: Estimates are obtained after running equation (21) where the different specifications refer to the different markup estimates, and we convert the percentage markup difference into levels as discussed above. The standard errors under specifications **I–V** are obtained from a nonlinear combination of the relevant parameter estimates. All regressions include labor, capital, and full year and industry dummies as controls. Standard errors are in parentheses.

Figure 2.1: Reprinted from De Loecker and Warzynski (2012)

altered.

2.3.2 Brandt et al.: Materials Markups and Tariff Reductions

Brandt et al. (2017) investigate the effects of tariff reductions following China's entry into the WTO on the markups, prices, and productivity dynamics of Chinese industrial firms. They compute markups as in De Loecker and Warzynski (2012), then regress markups on prior-year input and output tariffs (as well a set of firm, year, and sector fixed effects). Both input and output tariffs fell after China entered the WTO. Figure 2.2 reproduces the main result on firm-level markups in Brandt et al. (2017). They interpret these findings as suggesting that: (1) lower input tariffs increases markups because marginal costs fall; (2) lower output tariffs reduce markups because the pro-competitive effect of trade opening outweighs markup-increasing efficiency gains (efficiency gains increase markups because they reduce marginal costs).

TABLE 3—EFFECT OF TARIFFS ON FIRM-LEVEL MARKUPS AND PRODUCTIVITY (*logs*)

	No weights		Within-industry output share weights		Output weights	
	(1a)	(2a)	(3a)	(4a)	(5a)	(6a)
Panel A. Markups, 1998–2007						
Output tariff (lagged)	0.045	0.058	0.07	0.095	0.375	0.543
	(0.041)	(0.044)	(0.043)	(0.051)	(0.110)	(0.120)
Input tariff (lagged)	−0.227	−0.969	0.016	−0.733	−0.651	−1.173
	(0.106)	(0.074)	(0.169)	(0.094)	(0.354)	(0.181)
Firm fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Year fixed effects	Yes		Yes		Yes	
Sector-year fixed effects		Yes		Yes		Yes
Observations	1,293,495	1,295,372	1,211,861	1,213,586	1,293,495	1,295,372

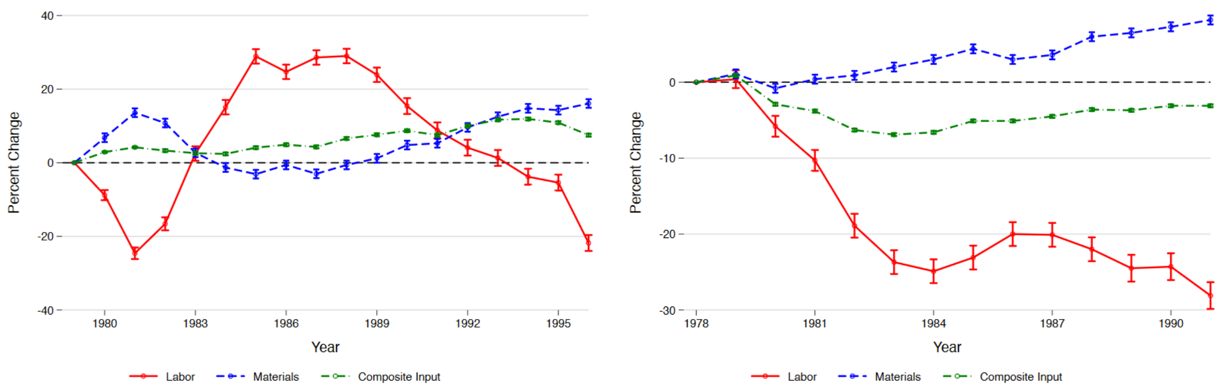
Figure 2.2: Reprinted from Brandt et al. (2017)

Brandt et al. (2017) use revenue data and the ratio estimator to estimate markups. They use materials expenditures and implied materials elasticities in their ratio formula. Our paper (in particular (2.4)) implies that - to the extent that their elasticities are revenue elasticities - these results should be interpreted as measuring materials-market wedges. Thus, lowering input tariffs increases materials wedges, perhaps because state planners pressure Chinese firms to use domestically-sourced inputs to avoid currency outflows. Lowering output tariffs decreases materials wedges, possibly as a result of firms pressuring suppliers once they face tougher foreign competition. Interestingly, this latter result holds for both labor markets (De Loecker and Warzynski (2012)) and materials markets (Brandt et al. (2017)). This suggests scope for future research investigating the impact of WTO accession on Chinese labor markets.

2.3.3 Raval: Comparing Ratio Results

Raval (2020) tests the ratio estimator by comparing implied markups from different inputs. If all inputs are competitive ($\delta_i = 1 \forall i$) and quantity elasticities are available, then (2.5) implies that the markup $\mathcal{M}$ should be identical across all inputs. Therefore comparing estimated $\mathcal{M}_i$ between inputs is a test of the ratio estimator: if $\mathcal{M}_i \neq \mathcal{M}_j$, the ratio estimator may be biased or inconsistent. Raval (2020) applies this test using manufacturing data for Chile and Colombia. We reproduce the time series of implied markups in Figure 2.3. The red lines represent markups computed from labor $\mathcal{M}_L$ and the blue lines markups computed from materials $\mathcal{M}_M$.

Figure 2 Markup Time Trends using Translog Estimates: Chile and Colombia



(a) Chile (b) Colombia

Note: Estimates based on (10), and include 95% Confidence Intervals (vertical bars) based on clustering at the establishment level. All estimates relative to the first year, which is set to zero.

Figure 2.3: Reprinted from Raval (2020)

Raval (2020) concludes that the ratio estimator is inconsistent across inputs and therefore misspecified. However, his data for both Chile and Colombia are also revenue data. If we assume that his estimated elasticities are revenue elasticities, (2.4) implies that his estimates should be reinterpreted as input-market wedges. Figure 2.3 therefore implies that labor-market wedges fell from 1980-1995 in both Chile and Colombia, while materials wedges rose. This suggests that the labor market in both countries became more competitive, but fuel and electricity markets less so. Perhaps institutional changes, upstream concentration, increased capital adjustment costs, or limits to entry in materials markets increased in this period. Rather than implying misspecification, the estimates in Raval (2020) compare input-market wedges across different inputs.

2.4 Conclusion

The ratio estimator recovers input market power rather than output market power when computed using revenue output elasticities. If one input is competitive and quantity elasticities are available, that input may be used to compute markups. If firms use (at least) one competitive input and additional imperfectly-competitive inputs, one can separately identify static input-market frictions and markups. These input-market frictions can be interpreted as static adjustment costs, or incorporated into dynamic models of adjustment costs.

Hence researchers do not need to entirely abandon attempts to estimate market power when only revenue data is available. Rather, the interpretation of the ratio estimator changes. Thus existing literature should be reinterpreted, and researchers have a valuable tool for studying imperfections in labor and other input markets. We believe this tool clarifies the tools and expands the scope of the market-power literature.

Chapter 3

MEASURING MARKUPS WITH REVENUE DATA

3.1 Introduction

Measuring market power is a vital task of economics. Firms that do not face sufficient threats of competition will price above marginal cost, resulting in inefficiencies and welfare losses. Understanding market power is therefore important to evaluate competition, efficiency, welfare, firm dynamics, and antitrust policies. Researchers and policymakers typically use the gap between price and marginal cost - the markup - to measure output market power. A growing body of research applies structural estimators to production data to measure markups [1]. Relative to the traditional demand-based approach in industrial organization, this production-based approach imposes few assumptions on the competitive process, and scales easily to increasingly-available production data that cover large segments of the economy.

However, existing production-based approaches to measuring markups require data on output quantity rather than only output revenue. The most widely-used such approaches, the structural estimators of Ackerberg et al. (2015) and Blundell and Bond (1998), were designed to study firm productivity in competitive settings. In imperfectly competitive settings, they require information on output quantity, but the standard data setting only has revenues (the product of quantity and price)[2] Without additional structure, we cannot disentangle whether firms have higher revenues because they are more productive (and increase quantity produced) or because they have higher market power (and raise prices). Put another way, using revenue data creates an omitted-variable problem since we cannot observe firm prices.

1. See De Loecker and Warzynski (2012), or more recently De Loecker et al. (2020) or Traina (2018)

2. For instance, the US Census of Manufactures, Compustat, and the Colombian, Chilean, and Slovenian manufacturing surveys all only have revenue data.

Because of this problem, applying existing estimators using revenue data produces biased and inconsistent estimates of markups (Klette and Griliches (1996) and Bond et al. (2020))[3]. In this paper, we outline a method to consistently estimate markups and production functions with revenue data.

We solve the omitted price bias from revenue data by pairing an intermediate 'share regression' of input spending on determinants of markups with a timing assumption on the evolution of productivity. The share regression separates the unobserved effect of productivity on output. The second stage then uses informational assumptions to isolate the effect of prices and inputs on output. Combined, our method identifies physical output elasticities that may be used to estimate markups. Our approach applies to a wide range of settings and market conditions. In particular, even if we had information on physical output quantities, we would need to adjust for quality to ensure that outputs are comparable; quality data is very rare in practice. Moreover, many firms produce more than one product, which complicates the interpretation of quantity data[4]. After presenting our model, we give several examples of settings in which it may be used.

The literature on production-function estimation has long noted this omitted-price problem. Klette and Griliches (1996) emphasize that using revenue data produces biased estimates of markups. Bond et al. (2020) apply this insight to contemporary ratio estimators. Our paper generalizes the solution in Klette and Griliches (1996), which relies on monopolistic competition and a parametric demand system. More recently, De Loecker et al. (2016) use observed output price to control for unobserved input price biases, but they operate in a setting in which output price may be computed from available data. Foster et al. (2008) suggest that omitted price bias is so important that it changes the observed correlation between physical productivity and revenue productivity. Meanwhile, Mairesse and Jaumandreu (2005) suggest

3. Dividing by price indices does not solve this problem, since the indices aggregate to a higher level. They cannot give us quantity estimates at the level of estimation.

4. Thus, our method is attractive even in the presence of quantity data.

that the bias may not be large. In sum, the literature has not yet converged on a method for estimating physical production functions from revenue data. We outline a method to consistently do so in this paper.

Our paper also relates to the literature that addresses simultaneity bias: the collinearity that comes from unobserved productivity determining both inputs and outputs (Marschak and Andrews (1944)). The literature uses timing and information assumptions on a firm's production process to use past inputs to control for present productivity (since past inputs were chosen before an innovation to a productivity process, they should be independent of such innovations). Broadly, dynamic panel estimators (Blundell and Bond (1998)) use linearity assumptions on the productivity process for identification, while proxy variable estimators (Olley and Pakes (1996), Levinsohn and Petrin (2003), Ackerberg et al. (2015), and Gandhi et al. (2020)) impose weaker timing assumptions but stronger informational assumptions. Our method is related to proxy variable estimators, and so does not require that the productivity process is linear. But it allows researchers to relax some of the more onerous assumptions of the proxy model. Consequently, our estimator may be applied to study production technology using the strengths of both approaches.

Our estimator builds on this line of work by simultaneously addressing the omitted price bias. It can be thought of as a version of the control-function "proxy" estimator (Olley and Pakes (1996), Levinsohn and Petrin (2003), Ackerberg et al. (2015), and Gandhi et al. (2020)), but with markups themselves as the object to be proxied. It generalizes the use of the share regression in Gandhi et al. (2020). Unlike Gandhi et al. (2020), we do not identify output elasticities directly from the share regression. Rather, we use it to partially identify a combination of technology and conduct by proxying directly for markups. We then use this composite to control for unobserved revenue productivity (profitability) in a second stage. This allows us to identify physical output elasticities even in the presence of revenue data. We view the main contribution of our model as providing estimates of market power in settings with revenue data.

The rest of the paper proceeds as follows. In Section 2, we briefly summarize a standard competitive production model. In Section 3, we describe our modified production model for imperfectly competitive settings, and describe how to use our proposed share regression as a control function to estimate markups. In Section 4, we compare our estimator to existing production-function estimation methods. We provide specific examples of markup controls in Section 5. Section 6 concludes.

3.2 A Standard Production Model

In this section, we outline the typical structural model and data that researchers use to estimate production functions and markups. We suppose throughout that the researcher wants to estimate firm-level markups, and only has data on firm revenues (that is, not quantities). Throughout the rest of the paper, we use uppercase to refer to levels of variables and lowercase to refer to logs of variables. One approach is to ignore the distinction between revenues and quantities, estimate a production function, and apply optimization assumptions to recover markups (for instance, the commonly-used ratio estimator). This empirical strategy confounds demand with productivity (Foster et al. (2008)) and results in inconsistent estimates of markups (Klette and Griliches (1996) and Bond et al. (2020)).

3.2.1 Definitions

We observe data for a panel of firms over periods $t = 1, 2, ..., T$. We omit panel subscripts and assume the data take a short panel form, so that the number of firms grows large for a fixed T. For each firm, we observe output Q, expenditures on a competitively-supplied flexible input X with cost C, and a vector of nonflexible inputs K. X is flexible in the sense that it is variable and static: firms may adjust it in each period after observing the relevant state variables discussed later, and its choice has no dynamic implications. K does not satisfy one of these properties.

Firms have some market power in the output market, so that each firm's residual demand curve is not perfectly elastic. Markups can come from a mix of cost, demand, conduct, or other sources. They may, for instance, be due to inelastic demand for a product due to advertising, a technological advantage over competitors, or a concentrated market. We are agnostic about the source of markups, and we do not require markups to be orthogonal to productivity. Cost-minimizing firms internalize inframarginal price reductions and choose a price and output $\{P, Q\}$ along their residual demand curves. We define firm revenue as $R \equiv Q \cdot P(Q)$ and the elasticity of R with respect to input X as $f_x^R \equiv (1 + \varepsilon^D) f_x$[5]. Firms optimally choose to price above marginal cost and we define the markup $\mathcal{M} \equiv \frac{P}{MC}$.

3.2.2 Information and Timing

The timing of the problem is as follows: first, a firm uses its expectation about a shock to its productivity, along with latent information on its residual demand curve, to choose an optimal markup. Second, the firm chooses inputs X, K to realize its optimally-chosen markup given its residual demand curve and its expectation about current productivity. Third, a productivity shock ε occurs. Productivity is realized, production occurs, and markups are realized.

Markups are not orthogonal to productivity, since the firm uses its expectation about productivity to select an optimal markup. We are otherwise agnostic about the source of output market power.

3.2.3 Technology & Productivity

Each firm's production function is given by $Q = F(X, K)$. Define the elasticity of $F(\cdot)$ with respect to input X as $f_x \equiv \frac{\partial F(\cdot)}{\partial X} \frac{X}{Q}$. Inputs generate output according to a constant returns

5. To obtain this result, totally differentiate the definition of revenues and apply Hicks' formula.

to scale production function with Hicks-neutral productivity: $Q = AF(X,K)$[6]. A represents physical productivity (TFPQ in the parlance of Foster et al. (2008)). However, firms make decisions on the basis of revenues, and so must also consider the evolution of prices. We discuss this in a following subsection.

We assume that the productivity term A is additively separable in logs into a part known to the firm when making input decisions ω, and an i.i.d. error term ε[7]. The physical productivity process then evolves as:

$$a = \omega + \varepsilon$$

Firms' technology in logs is (again denoting logs in lowercase letters):

$$q = f(x,k) + \omega + \varepsilon \tag{3.1}$$

3.2.4 Optimization

We assume that firms minimize costs. Each firm uses expected output in its minimization problem because it knows that it must account for an as-yet-unknown portion of productivity ε. The firm's cost-minimization problem for input X is[8]:

$$\min_X CX$$
$$s.t. \quad Q = E[AF(X,K)]$$

6. Any returns-to-scale parameter may be substituted for CRS, but some assumption is needed because markups are not separately identified from the scale parameter (Basu (2019)). Hicks-neutrality of productivity ensures that the production function is additive in logs.

7. More specifically, we may assume that $\varepsilon \sim \mathbb{N}(0, \sigma_\varepsilon^2)$, though this is not necessary for what follows.

8. We omit K from the optimization problem since K is quasi-fixed.

Denoting optimal output by Q^*, the Lagrangian is:

$$CX + \lambda(Q^* - E[AF(X,K)]$$

With first-order condition for X:

$$[X] \quad C = \lambda E[A]\frac{\partial F(X,K)}{\partial X}$$

Denote P as the firm's (endogenous) output price and note that the markup $\mathcal{M} \equiv \frac{P}{mc} \equiv \frac{P}{\lambda}$[9]. Our definition of the production function implies that $\frac{A}{Q} = F(X,K)$. Manipulating the FOC using these definitions yields:

$$\begin{aligned}
C &= \lambda E[A]\frac{\partial F(X,K)}{\partial X} \\
\frac{P}{\lambda} &= E[A]\frac{\partial F(X,K)}{\partial X} \cdot \frac{P}{C} \\
\mathcal{M} &= E[A]\frac{X}{Q} \cdot \frac{\partial F(X,K)}{\partial X} \cdot \frac{P}{C} \cdot \frac{Q}{X} \\
\mathcal{M} &= \frac{E[A]}{A} \cdot AXQ \cdot \frac{\partial F(X,K)}{\partial X} \cdot \frac{PQ}{CX} \\
\mathcal{M} &= \frac{E[A]}{A} \cdot XF(X,K) \cdot \frac{\partial F(X,K)}{\partial X} \cdot \frac{PQ}{CX} \\
\mathcal{M} &= \frac{E[\exp(\varepsilon)]}{\exp(\varepsilon)} f_x \frac{R}{CX}
\end{aligned}$$

Where the last line comes from the definition $R \equiv PQ$ and from the fact that $\frac{E[A]}{A} = \frac{\exp(\omega)E[\exp\varepsilon]}{\exp\omega\exp\varepsilon} = \frac{E[\exp(\varepsilon)]}{\exp(\varepsilon}$. Writing this result in logs, we obtain:

$$\mu = \log(f_x) + r - cx + b - \varepsilon \tag{3.2}$$

9. The Envelope Theorem implies that the Lagrange multiplier λ is the marginal cost of production.

Where $b \equiv \log(E[\exp(\varepsilon)])$. Recovering markups from (3.2) requires an estimate of f_x. We cannot simply regress R on X to obtain this estimate, for two main reasons. First, there is simultaneity bias through productivity: higher A induces the firm to increase X, which increases Q and thus R. This 'transmission bias' was first noted by Marschak and Andrews (1944). The second reason OLS of revenue on inputs generates biased estimates of f_x is that determinants of price are omitted from the right-hand-side regressors (emphasized by Klette and Griliches (1996)). We discuss our solutions to these issues in the next section.

3.3 The Share Regression with Imperfect Competition

In this section, we discuss the assumptions needed to estimate markups and production function parameters with revenue data. We first interpret the first-order condition as a markup function, and use it to identify revenue elasticities. We then combine it with timing and informational assumptions on productivity to identify physical output elasticities. Finally, we discuss examples of markup identification.

3.3.1 Assumption 1: The Markup Control Function

Return to the first-order condition (3.2) and rewrite it as:

$$cx - r = \log(f_x) - \mu + b - \varepsilon \tag{3.3}$$

The left-hand side is the log intermediates share of revenues. The term $(\log(f_x) - \mu)$ on the right-hand side is the elasticity of revenues with respect to input X, or f_x^R. The elasticity term f_x in (3.3) is a function only of inputs[10]: $f_x = f_x(x, k)$. Let markups be a function of inputs, firm i and time t fixed effects, and a vector of other observables $\mathbf{S}$: $\mathcal{M} = \mathcal{M}(x, k, i, t, \mathbf{S})$.

10. f_x is not a function of TFP A by virtue of (assumed) Hicks-neutrality. Since increases in A increase output and input expenditures symmetrically, they cancel out of the elasticity formula.

Then, making the timing explicit, (3.3) can be written as:

$$cx_t - r_t = s(x_t, k_t, i, t, \mathbf{S}_t) + b - \varepsilon_t \tag{3.4}$$

So that a researcher may nonparametrically regress the intermediates share on inputs, fixed effects, and a vector of markup determinants $\mathbf{S}$ to obtain an estimate of the revenue elasticity[11]. This share regression (3.4) can be seen as a *markup function*: it describes the determinants of output wedges[12]. Importantly, it also recovers an estimate of the productivity error $\hat{\varepsilon}_t$. Estimating $\hat{\varepsilon}$ is the main function of the first stage of proxy estimators (Olley and Pakes (1996), Levinsohn and Petrin (2003), and Ackerberg et al. (2015)). Estimating it here allows us to drop some of the assumptions of these models (see below for more details). However, the share regression cannot separate the impact of markups from output elasticities, since it still contains the unknown f_x. We now turn to identifying the physical elasticity f_x by adding timing and informational assumptions on the productivity process.

3.3.2 Assumption 2: Markovian TFPR

To separately identify markups and physical elasticities, we combine the share regression with a timing assumption on profitability. With this assumption, the model can separate (physical) productivity from input use.

Assumption: Prices p and productivity innovations ν jointly follow a Markov process with mean-zero shocks η: defining $p + \omega \equiv \nu$, this implies that $\nu_t = g(\nu_{t-1}) + \eta$.

11. The constant represents a Jensen nuisance term in planned output.

12. Our share regression is similar to that in Gandhi et al. (2020), but adapted for use in the revenue context.

Now write firm revenues as:

$$
\begin{aligned}
r_t &= p_t + q_t \\
&= f_t + p_t + \omega_t + \varepsilon_t \\
&= f_t + \nu_t + \varepsilon_t \\
&= f_t + g(\nu_{t-1}) + \eta_t + \varepsilon_t
\end{aligned} \tag{3.5}
$$

Where the second line comes from the definition of $f(\cdot)$, the third line uses the definition of ν_t, and the fourth line comes from Assumption 1. Now rewrite the markup function (3.3) as:

$$
\begin{aligned}
\log(\mathcal{M}_t) &= \log(f_x) + r_t - cx_t + b - \varepsilon_t \\
&= \log(f_x) + (f_t + \nu_t + \varepsilon_t) - cx_t + b - \varepsilon_t \\
&= \log(f_x) + (f_t + \nu_t) - cx_t + b \\
\nu_t &= cx_t - f_t - \log(f_x) + \log(\mathcal{M}_t) - b
\end{aligned}
$$

Insert this definition into the solution for revenues (3.5) to yield:

$$
r_t = f_t + g(cx_{t-1} - f_{t-1} - \log(f_x) + \log(\mathcal{M}_{t-1}) - b) + \eta_t + \varepsilon_t \tag{3.6}
$$

Which implies that the model identifies the elasticity f_x so long as markups can be controlled for. This is because r_t and cx_{t-1} are observed in revenue production data, $f_t(\cdot)$ is a parametric function of observable inputs, $g(\cdot)$ is a nonparametric function which is independent of current inputs, and f_x is a function of observed inputs. We control for $\mathcal{M}_{t-1}$ in (3.6), there must be some observed independent variation in $\mathcal{M}(\cdot)$ which does not enter $f(\cdot)$. This is the $\mathbf{S}$ vector in (3.4), which affects only $\mathcal{M}(\cdot)$. $\mathbf{S}$ could include firm or time fixed effects, or any other observable (we discuss specific examples below). With this variation, the model identifies quantity elasticities and markups. The markup function solves the omitted-variables problem, while the TFPR timing assumption solves the transmission bias. Since returns-to-scale are isomorphic to markups (Klette and Griliches (1996) and Flynn et al. (2019)), we require a returns-to-scale assumption.

In sum, our method uses the first-order condition

3.3.3 Estimation

To implement our estimation technique:

1. Regress the log intermediates share of revenues $(cx - r)$ on log inputs and fixed effects, along with markup determinants $\mathbf{S}$. Obtain the predicted share of expenditures $\hat{s} = \log(f_x) - \mu + b$ and the predicted residual $\hat{\varepsilon}$

2. Regress revenues on inputs, a nonparametric function of lagged $\hat{s}$, and $\hat{\varepsilon}$; that is, $r = f(X_t, K_t; \boldsymbol{\theta}_\mathbf{f}) + g(cx_{t-1} - \hat{s}_t - f(X_{t-1}, K_{t-1}; \boldsymbol{\theta}_\mathbf{f}); \boldsymbol{\theta}_\mathbf{g}) + \hat{\varepsilon}_t + \eta_t$, where $g(\cdot)$ is a polynomial of arbitrary degree and $f(\cdot)$ is a prespecified production function (e.g. translog)

3. Use the moment condition

$$\hat{\eta}_t \cdot \begin{bmatrix} 1 \\ X_{t-1} \\ K_t \\ \mathbf{S}_{t-1} \end{bmatrix} = 0$$

   to recover η as a function of the vector of elasticities $\boldsymbol{\theta}_\mathbf{f}$ and $g(\cdot)$ parameters $\boldsymbol{\theta}_\mathbf{g}$

4. Minimize the distance $\hat{\eta}(\boldsymbol{\theta}_\mathbf{f}, \boldsymbol{\theta}_\mathbf{g})' \times W \times \hat{\eta}(\boldsymbol{\theta}_\mathbf{f}, \boldsymbol{\theta}_\mathbf{g})$ by GMM to recover the quantity elasticities $\boldsymbol{\theta}_\mathbf{f}$ and production process parameters $\boldsymbol{\theta}_\mathbf{g}$ using any weight matrix W

The specific moments in Step 3 depend upon the models one chooses for $f(\cdot)$ and $g(\cdot)$. For instance, additional moments for a translog production function $f(\cdot)$ can be added by including squares and interactions of the inputs X and K. Additional lags and interactions of the markup determinant vector $\mathbb{S}$ may be added to control for more complicated specifications of the productivity process $g(\cdot)$. For instance, one can add $\mathbb{S}^2$ to control for a second-order $g(\cdot)$ process. By construction, these additional moments are orthogonal to productivity innovations. In economic terms, costs, demand, and conduct are determinants of markups, which co-evolve with the TFPR process. Innovations to this process are therefore orthogonal to any functions of the costs, demand, or conduct embedded in the $\mathbf{S}$ vector.

This two-step approach can also be implemented as a single GMM problem by minimizing the residuals in the share regression together with the moment conditions from the second step (Wooldridge (2009)). This adds one moment condition ($E[\hat{\varepsilon}|s] = 0$) to the GMM problem in Step 3.

3.4 Comparison to Existing Methods

In this section, we compare our estimator to two commonly-used methods to estimate production functions. We first discuss the proxy class of model, then the dynamic panel class.

3.4.1 Proxy Variable Estimators

The proxy model of production (Olley and Pakes (1996), Levinsohn and Petrin (2003), Ackerberg et al. (2015), Gandhi et al. (2020)) uses timing assumptions on productivity to instrument for current productivity with past input use. Proxy models assume that productivity is Markovian and that intermediates use is strictly monotonic with respect to productivity. These two assumptions allow one to use lagged inputs to proxy for productivity, solving the transmission bias. More formally, the assumptions of the proxy model are:

Proxy Assumption 1: Productivity innovations follow a Markov process with mean-zero shocks: $\omega_t = g(\omega_{t-1}) + \xi$.

Proxy Assumption 2: Intermediates X are strictly monotonic in productivity A conditional on other inputs K.

Proxy Assumption 2 ensures that the mapping from productivity ω to inputs X is invertible: $\omega_t = \mathbb{X}(X_t, K_t)$. Then past inputs can proxy for current productivity through the productivity process and the fact that the firm cannot perfectly forsee ω innovations. Then estimation can proceed from $y_t = f(X_t, K_t; \boldsymbol{\theta}_\mathbf{X}) + \mathbb{X}(X_t, K_t) + \varepsilon_t$.

The proxy approach requires that there be no other unobservable affecting intermediates demand: intermediates must be deterministic conditional on other inputs and productivity. Our approach relaxes this *scalar unobservable* assumption: we replace it with the assumption that no other inputs enter the markup determinant vector $\mathbf{S}$. In other words, we impose scalar unobservability on $\mathcal{M}$. Markups and productivity are both fundamentally unobservable, but researchers typically impose some markup-setting process or rule in modeling (e.g., monopolistic competition with a given elasticity of substitution). Therefore, a scalar unobservability assumption on $\mathcal{M}$ is possibly less limiting. We do not require the monotonicity assumption (Proxy Assumption 1) at all.

More generally, our model can be thought of as a version of the proxy estimator, in which we proxy for markups directly instead of proxying for productivity. It is built to estimate markups, and also allows us to relax some of the proxy assumptions.

3.4.2 Dynamic Panel Estimators

The dynamic panel approach (Blundell and Bond (1998) and Arellano and Bond (1991)) is a commonly-used alternative to proxy approaches. Dynamic panel models maintain the same basic structure of production. They assume that productivity innovations ω_t follow an AR-1 process with innovations ξ_t: $\omega_t = \rho\omega_{t-1} + \xi_t$. The unobserved term ξ_t is uncorrelated with all past and future input choices. Then ρ-differencing the production function yields:

$$r_t - \rho r_{t-1} = f_t - \rho f_{t-1} + p_t - \rho p_{t-1} + \omega_t - \rho\omega_{t-1} + \varepsilon_t - \rho\varepsilon_{t-1} \tag{3.7}$$

$$r_t = \rho r_{t-1} + (f_t - \rho f_{t-1}) + \nu_t - \rho\nu_{t-1} + (\varepsilon_t - \rho\varepsilon_{t-1})$$

We cannot proceed further without price data. However, note that (3.3) gives us an expression for ν_t. Then (3.8) becomes:

$$r_t = \rho r_{t-1} + (f_t - \rho f_{t-1}) + (cx_t - f_t - \log(f_x)\log(\mathcal{M}_t) - b) - \rho(cx_{t-1} - f_{t-1} - \log(f_x) + \\ + \log(\mathcal{M}_{t-1}) - b) + (\varepsilon_t - \rho\varepsilon_{t-1})$$

$$r_t = \rho r_{t-1} + (cx_t - \rho cx_{t-1}) + \log(f_x)(\rho - 1) + b(\rho - 1) + (\log(\mathcal{M}_t) - \rho\log(\mathcal{M}_{t-1})) \\ + (\varepsilon_t - \rho\varepsilon_{t-1})$$

$$(\varepsilon_t - \rho\varepsilon_{t-1}) = (r_t - \rho r_{t-1} + \log(f_x)(\rho - 1) + b(\rho - 1) + (\log(\mathcal{M}_t) - \rho\log(\mathcal{M}_{t-1})) \tag{3.8}$$

Estimation can proceed by using twice-lagged (or more) values of inputs as instruments for $(\varepsilon_t - \rho\varepsilon_{t-1})$. On the other hand, if TFPR is AR-1 with $\nu_t = \rho\nu_{t-1} + \xi_t$, then (3.8) becomes:

$$r_t = \rho r_{t-1} + (f_t - \rho f_{t-1}) + \xi_t + (\varepsilon_t - \rho\varepsilon_{t-1})$$

And again estimation can proceed using twice- or more lagged inputs. In other words, our share regression allows researchers to use dynamic panel models to identify quantity elasticities with revenue data even assuming that only TFPQ (and not TFPR) is AR-1. If a researcher is comfortable assuming that TFPR is AR-1, then dynamic panel models may be used directly.

3.5 Examples

In Section 3.3.1, we showed that our markup function identifies market power $\mathcal{M}$ so long as $\mathcal{M}$ is determined (partially) independently of inputs. In this section, we offer several commonly-used parametric examples of markup functions.

3.5.1 Monopolistic Competition with CES Demand

Suppose that firms are monopolistic competitors facing a constant price elasticity of demand σ. Suppose further that these firms compete in a number of industries j. In this Dixit-Stiglitz environment, a firm in industry j faces an implicit demand curve given by $p_{ij} = P_j \left(\frac{q_{ij}}{Q_j}\right)^{-\frac{1}{\sigma_j}}$. Firm optimization implies that markups are constant within industries and given by $\mathcal{M}_j = \frac{\sigma_j}{\sigma_j - 1}$.

In the context of our markup function, assuming CES demand and monopolistic competition implies that $\mathbf{S} = c_j$: markups are determined by a constant within industry. One may recover markups and elasticities by adding a constant in the share regression (3.4) and the productivity control function in (3.6).

This is the setting in Klette and Griliches (1996). They show that in this case, one can estimate (industry-level) production functions by including controls for the industry indices P_j and Q_j. Klette and Griliches (1996) use the estimating equation

$$r_{it} = \beta_0 + \frac{\eta + 1}{\eta}(\beta_1 x_{it} + \beta_2 k_{it}) - \frac{1}{\eta} q_{It} + \nu_{it} \tag{3.9}$$

Where q_{It} is an *industry-level* price index which comes from the Dixit-Stiglitz monopolistically-competitive environment. Estimation can then proceed using (observed) industry-level output. In this model, $\mathcal{M} = \frac{\eta}{\eta - 1}$, so equation (3.9) is similar to our (3.6) but without the time-series structure of productivity. The level of markups can be recovered from the coefficient on industry output $q_I It$.

3.5.2 Market Characteristics

If the researcher has data on market characteristics which determine $\mathcal{M}$, then our model identifies market power and production elasticities by putting these into $\mathbf{S}$. For instance, advertising expenditure or export status have been used in the literature (Hall (2014) and De Loecker and Warzynski (2012)) as controls for markups[13]. In our context, we would then have $\mathbf{S} = \mathbf{S}(ads, export)$. Any other observable market characteristics - geography, demographics, market structure, product type, number of competitors - may be added into $\mathbf{S}$, depending on the researcher's model. Then (3.4) and (3.6) identify the output elasticities and markups.

3.5.3 Market Shares

If markups are determined by market shares, as for example in a Cournot model, then in the context of our model we have $\mathbf{S} = q_i/Q_j$, where Q_j is industry output[14]. Then (3.4) and (3.6) may again be used to identify markups and output elasticities by adding market shares to each. In contrast to models such as nested CES, this approach does not impose a parametric relationship between market shares and markups. Rather, the data determine the relationship. Of course, we still make assumptions about the productivity process and the timing of firm information (but not about the parametric form of the production function).

3.5.4 Conduct Instruments

One readily constructed choice of $\mathcal{M}$ instrument is *competitors' input choices* within industries j. Markups depend upon industry conduct, and therefore upon competitors' choices.

13. Broadly speaking, the idea is that advertising can reduce the price elasticity of demand, while exporting forces firms to adopt better production techniques, which then can drive markups in a less-competitive domestic market.

14. Researchers may define industries j in whatever way appropriate. They may be classified industries, or broadly- or narrowly-defined product markets.

We sketch here a straightforward application of this approach. Suppose that that each firm operates under Markovian "market conditions" d_{ijt}:

$$d_{ijt} = \mathbb{D}(d_{ij,t-1}) + u_{ijt}$$

Which evolve over time with error terms u_{ijt} orthogonal to the productivity innovation η_{ijt}. Firms attempt to forecast the evolution of market conditions and use other firms' input choices as proxies. Then we have $E[\eta_{ijt} \times d_{ijt-1}] = 0$. Firms use competitor choices to form predictions about market conditions, but these choices do not directly affect productivity.

These conduct instruments are valid in many models of imperfect competition. For instance, in a Cournot setting each firm is small relative to its industry and takes into account other firms' pricing decisions when setting its output. More generally, any model in which demand and supply conditions are persistent is consistent with using lagged d as instruments. For instance, a model of habit formation is consistent with our timing assumptions, as is a model with serially correlated input and output shocks. These instruments do not rule out perfect competition in intermediates markets or strategic interactions between firms (indeed, they rely on such interactions). They do, however, rule out dynamic strategic interactions.

These conduct instruments are closely related to Bartik instruments, and are readily constructed in production datasets: all that is required is input choices by industry.

3.6 Conclusion

This paper shows how to estimate markups with revenue data. We combine firm cost-minimization and an assumptions on the timing of firm decision-making to recover quantity elasticities and estimates of market power. We show that, so long as a researcher has some variable which affects markups independently of inputs, they can identify both physical production elasticities and markups. This is less restrictive than it seems, for this independent variation could come from an industry-level constant (as in the CES-monopolistic competition case), fixed effects, market shares, or other inputs recoverable from production data. What is required is for the researcher to spell out the determinants of markups in their model and relate them to observables in the data.

Our method is a variant of proxy-style production-function estimators, but proxies explicitly for markups rather than for productivity. We believe it is a tool which can be applied to a wide variety of settings. We hope that it can be used to clarify and expand the scope of the competition literature.

Bibliography

Acemoglu, Daron, and Pascual Restrepo. 2020. "Robots and Jobs: Evidence from US Labor Markets." *Journal of Political Economy* 128 (6): 2188–2244.

———. 2021. *Tasks, Automation, and the Rise in US Wage Inequality.* Working Paper, Working Paper Series 28920. National Bureau of Economic Research.

Ackerberg, Daniel A, Kevin Caves, and Garth Frazer. 2015. "Identification properties of recent production function estimators." *Econometrica* 83 (6): 2411–2451.

Aghion, P., A Bergeaud, T Boppart, P Klenow, and H. Li. 2020. *A Theory of Falling Growth and Rising Rents.*

Arellano, Manuel, and Stephen Bond. 1991. "Some Tests of Specification for Panel Data: Monte Carlo Evidence and an Application to Employment Equations." *Review of Economic Studies* 58 (2): 277–297.

Atalay, Enghin, Ali Hortaçsu, and Chad Syverson. 2014. "Vertical Integration and Input Flows." *American Economic Review* 104 (4): 1120–48.

Autor, David H., and David Dorn. 2013. "The Growth of Low-Skill Service Jobs and the Polarization of the US Labor Market." *American Economic Review* 103 (5): 1553–97.

Autor, David, David Dorn, Lawrence F Katz, Christina Patterson, and John Van Reenen. 2020. "The Fall of the Labor Share and the Rise of Superstar Firms." *The Quarterly Journal of Economics* 135, no. 2 (February): 645–709.

Azar, José, Ioana Marinescu, and Marshall I Steinbaum. 2020. "Labor Market Concentration." *The Journal of Human Resources.*

Baqaee, David Rezza, and Emmanuel Farhi. 2019. "Productivity and Misallocation in General Equilibrium." *The Quarterly Journal of Economics* 135, no. 1 (September): 105–163.

Barkai, Simcha. 2020. "Declining Labor and Capital Shares." *The Journal of Finance* 75 (5): 2421–2463.

Basu, Susanto. 2019. "Are Price-Cost Markups Rising in the United States? A Discussion of the Evidence." *Journal of Economic Perspectives* 33 (3): 3–22.

Benmelech, Efraim, Nittai Bergman, and Hyunseob Kim. 2018. *Strong employers and weak employees: How does employer concentration affect wages?* Technical report. National Bureau of Economic Research.

Berger, David, Kyle Herkenhoff, and Simon Mongey. 2019. *Labor Market Power.* Technical report.

Blundell, Richard, and Stephen Bond. 1998. "Initial Conditions and Moment Restrictions in Dynamic Panel Data Models." *Journal of Econometrics* 87 (1): 115–143.

Bond, Steve, Arshia Hashemi, Greg Kaplan, and Piotr Zoch. 2020. "Some Unpleasant Markup Arithmetic: Production Function Elasticities and their Estimation from Production Data."

Brandt, Loren, Johannes Van Biesebroeck, Luhang Wang, and Yifan Zhang. 2017. "WTO Accession and Performance of Chinese Manufacturing Firms." *American Economic Review* 107 (9).

Card, David, Ana Rute Cardoso, Joerg Heining, and Patrick Kline. 2018. "Firms and Labor Market Inequality: Evidence and Some Theory." *Journal of Labor Economics* 36 (S1): S13–S70.

De Loecker, Jan, Jan Eeckhout, and Gabriel Unger. 2020. “The Rise of Market Power and the Macroeconomic Implications.” *The Quarterly Journal of Economics* 135, no. 2 (January): 561–644.

De Loecker, Jan, Pinelopi K. Goldberg, Amit K. Khandelwal, and Nina Pavcnik. 2016. “Prices, Markups, and Trade Reform.” *Econometrica* 84, no. 2 (March): 445–510.

De Loecker, Jan, and Frederic Warzynski. 2012. “Markups and firm-level export status.” *American Economic Review* 102 (6): 2437–71.

Decker, Ryan A., John Haltiwanger, Ron S. Jarmin, and Javier Miranda. 2016. “Declining Business Dynamism: What We Know and the Way Forward.” *American Economic Review* 106 (5): 203–07.

Dobbelaere, Sabien, and Jacques Mairesse. 2013. “Panel data estimates of the production function and product and labor market imperfections.” *Journal of Applied Econometrics* 28 (1): 1–46.

Doraszelski, Ulrich, and Jordi Jaumandreu. 2013. “R&D and Productivity: Estimating Endogenous Productivity.” *The Review of Economic Studies* 80, no. 4 (March): 1338–1383.

Edmond, Chris, Virgiliu Midrigan, and Daniel Yi Xu. 2021. *How Costly Are Markups?* Technical report. Journal of Political Economy (Revise & Resubmit).

Eggertsson, Gauti B, Jacob A Robbins, and Ella Getz Wold. 2018. *Kaldor and Piketty's facts: the rise of monopoly power in the United States.* Technical report. National Bureau of Economic Research.

Elsby, Michael WL, Bart Hobijn, and Ayşegül Şahin. 2013. “The Decline of the US Labor Share.” *Brookings Papers on Economic Activity* 2013 (2): 1–63.

Flynn, Zach, Amit Gandhi, and James Traina. 2019. “Measuring Markups with Production Data.” Working Paper.

Fort, Teresa C, and Shawn D Klimek. 2016. “The effects of industry classification changes on US employment composition.” Working Paper.

Foster, Lucia, John Haltiwanger, and Chad Syverson. 2008. “Reallocation, Firm Turnover, and Efficiency: Selection on Productivity or Profitability?” *American Economic Review* 98 (1): 394–425.

Gandhi, Amit, Salvador Navarro, and David Rivers. 2020. “On the Identification of Gross Output Production Functions.” *Journal of Political Economy* 128 (8): 2973–3016.

Hall, Robert E. 2018. *New evidence on the markup of prices over marginal costs and the role of mega-firms in the US economy.* Technical report. National Bureau of Economic Research.

———. 1988. “The relation between price and marginal cost in US industry.” *Journal of Political Economy* 96 (5): 921–947.

———. 2014. “What the Cyclical Response of Advertising Reveals about Markupsand other Macroeconomic Wedges.” *NBER Working Paper.*

Hershbein, Brad, Claudia Macaluso, and Chen Yeh. n.d. “Monopsony in the U.S. Labor Market.” *American Economic Review (Revise & Resubmit).*

Jaumandreu, Jordi. 2018. *Dangerous Shortcuts: Ignoring Marginal Cost Determinants in Markup Estimation.* Technical report.

Karabarbounis, Loukas, and Brent Neiman. 2019. “Accounting for factorless income.” *NBER Macroeconomics Annual* 33 (1): 167–228.

Karabarbounis, Loukas, and Brent Neiman. 2013. "The global decline of the labor share." *Quarterly Journal of Economics* 129 (1): 61–103.

Kehrig, Matthias. 2015. "The Cyclical Nature of the Productivity Distribution." *The Quarterly Journal of Economics (Revise & Resubmit).*

Kehrig, Matthias, and Nicolas Vincent. 2021. "The Micro-Level Anatomy of the Labor Share Decline." *The Quarterly Journal of Economics* 136, no. 2 (March): 1031–1087.

Kirov, Ivan, and James Traina. 2021a. "Measuring Markups with Revenue Data." Working Paper.

———. 2021b. "What Do Production-Based Markup Estimators Actually Measure?" Working Paper.

Klette, Tor Jakob, and Zvi Griliches. 1996. "The Inconsistency of Common Scale Estimators when Output Prices are Unobserved and Endogenous." *Journal of Applied Econometrics* 11 (4): 343–361.

Lamadon, Thibaut, Magne Mogstad, and Bradley Setzler. 2020. "Imperfect Competition, Compensating Differentials and Rent Sharing in the U.S. Labor Market." *American Economic Review (Revise & Resubmit).*

Levinsohn, James, and Amil Petrin. 2003. "Estimating production functions using inputs to control for unobservables." *review of economic studies* 70 (2): 317–341.

Mairesse, Jacques, and Jordi Jaumandreu. 2005. "Panel-data Estimates of the Production Function and the Revenue Function: What Difference Does It Make?*." *The Scandinavian Journal of Economics* 107 (4): 651–672.

Manning, Alan. 2021. "Monopsony in Labor Markets: A Review." *ILR Review* 74 (1): 3–26.

Marschak, Jacob, and William H Andrews. 1944. “Random simultaneous equations and the theory of production.” *Econometrica,* 143–205.

Morlacco, Monica. 2020. “Market Power in Input Markets: Theory and Evidence from French Manufacturing.”

Neil, Bailey Martin, Charles Hulten, David Campbell, Timothy Bresnahan, and Richard E. Caves. 1992. “Productivity Dynamics in Manufacturing Plants.” *Brookings Papers on Economic Activity. Microeconomics* 1992:187–267.

Olley, G Steven, and Ariel Pakes. 1996. “The Dynamics of Productivity in the Telecommunications Equipment Industry.” *Econometrica,* 1263–1297.

Raval, Devesh. 2020. “Testing the Production Approach to Markup Estimation.”

Rinz, Kevin. 2018. *Labor Market Concentration, Earnings Inequality, and Earnings Mobility.* Technical report. Center for Administrative Records Research and Applications Working Papers.

Robinson, Joan. 1933. “The Economics of Imperfect Competition.” *The Economic Journal* 43, no. 172 (December): 657–661.

Rossi-Hansberg, Esteban, Pierre-Daniel Sarte, and Nicholas Trachter. 2021. “Diverging Trends in National and Local Concentration.” *NBER Macroeconomics Annual* 35:115–150.

Syverson, Chad. 2019. “Macroeconomics and Market Power: Facts, Potential Explanations and Open Questions.” *Brookings Economic Studies Report.*

Traina, James. 2018. *Is Aggregate Market Power Increasing? Production Trends Using Financial Statements.* Technical report. Stigler Center for the Study of the Economy and the State New Working Paper Series.

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

Appendix A

DATA APPENDIX

A.1 ASM/CM data

Our data consist of inputs, outputs, and plant characteristics from the Annual Survey of Manufactures and Census of Manufactures. The CM is a required survey sent to all manufacturing plants conducted in years ending in -2 and -7. The ASM is sent to a subset of establishments, with larger establishments having a higher probability of selection into the ASM (the precise cutoff for inclusion with certainty has changed over time). ASM observations include a sampling weight which is the inverse of an establishment's probability of selection into the ASM sample. All results in this paper include ASM sampling weights. We subset our data to establishments which are observed in the ASM. This ensures that we have panel data throughout (indeed, we would have to drop establishments which are not observed in the ASM in estimation, since we have no lagged inputs for them). We use the Total Factor Productivity Beta version 2 database of the ASM within the Census database. Our main geographical identifier for plants is the commuting zone (CZ).

For price indices and depreciation rates, we merge in the NBER-CES manufacturing database by NAICS6 and year. For capacity utilization, we merge in the Fed's Capacity Utilization Survey by NAICS3 and year. For scaling, we merge in the KLEMS database by DJA industry and year. To do the latter merge, we first construct a NAICS6-DJA bridge using the appropriate weights from KLEMS.

We measure output as revenues from produced goods (e.g., sales less resales and goods kept in inventory: `Q = (TVS - CR) + (FIE - FIB) + (WIE - WIB)`, with each component deflated by an appropriate price deflator: `PISHIP` from NBER-CES for the first, and intermediate price indices for the second two).

We measure intermediates as materials plus energy plus fuels. The idea is to capture a broad range of intermediate inputs, so that the bundle is approximately competitive even if each component of it is not. For quantity measures, we deflate each of materials, energy, and fuels with an appropriate deflator from the NBER-CES Manufacturing Database (materials and fuels with $PIMAT$, energy with $PIEN$).

We measure production labor as total production-worker hours at a plant, and nonproduction labor as the total number of nonproduction workers (we have no data on nonproduction hours).

We construct capital using the perpetual inventory method (PIM) backwards and forwards from an establishment's first CM observation (the idea here is that capital may be measured more accurately in CM years). Following Kehrig (2015), we multiply ASM/CM capital stocks by the ratio of current to historical cost of capital from the BEA capital tables (these costs are reported at the 3-digit industry level). Using investment data in the ASM/CM and industry-specific depreciation rates from the NBER-CES manufacturing database, we iterate backwards and fowards using

$$K_{t+1}^i = (1 - \delta_i) K_t^i + I_t^i$$

for $i \in \{E, S\} = \{\text{Equiment, Structures}\}$. Capital and investment data is available separately for structures and equipment in the ASM/CM. We run PIM for each type of capital separately, then add them to get an establishment's total capital stock. Finally, we adjust capital stocks by 3-digit industry capital utilization rates from the Fed's Plant Capacity Utilization survey and scale capital at the DJA industry level using the KLEMS database. Our final measure of capital is then effective (utilized) capital scaled to an industry aggregate.

A.1.1 Industry Codes

The definition of industry codes changes every 5 years in our sample. There are particularly large changes in 1987 (when SIC codes change from the 1972 to 1987 vintage) and in 1997 (when SIC codes change to NAICS codes for all establishments). There are then significant changes in manufacturing NAICS codes in 2012. To get a consistent set of industry codes, we crosswalk establishment-year observations to their corresponding Fort-Klimek 2007 NAICS code (Fort and Klimek (2016)). This crosswalk is available within the Census. The Fort-Klimek crosswalk assigns a consistent 2007 NAICS code to each establishment.

A.1.2 Scaling to KLEMS

To account for mismeasured inputs, we scale each of M, L, N, K to their industry-level equivalents in KLEMS. The idea is that the CM/ASM may not measure total input expenditure because some expenditure is done at the firm level (rather than the establishment level). This may be a particular concern for management inputs N or capital K. KLEMS hosts aggregate data on productive inputs and outputs for a number of countries.

To implement our scaling, we first generate KLEMS aggregates. We aggregate output, intermediates, labor, and capital indices in both gross and quantity terms (using a base year of 1997). Then in the ASM/CM data we generate industry aggregates $ASM_j = \sum_{kt} J_{kt}$ for $J \in \{M, L, N, K\}$ and KLEMS industries k (KLEMS industries are named DJA codes, and are approximately equivalent to the NAICS-3 level of aggregation). We account for sampling weights when doing this ASM_j aggregation. We then compute $scale_j = \frac{KLEMS_j}{ASM_j}$ and multiply inputs by $scale_j$. We scale ASM revenues to KLEMS output values, ASM intermediates expenditures to KLEMS intermediates values, labor and nonproduction labor inputs to KLEMS labor quantities, and all other inputs to KLEMS quantities.

To scale labor inputs, we first generate total hours using the Foster et al. (2008) adjustment as:

$$th = L \cdot \frac{WL + SN}{WL}$$

We then scale th to labor quantities in KLEMS by year, and total labor expenditure ($WL + SN$) to KLEMS labor expenditure by year. We use the same scaling factor for both L and N, as well as for both WL and SN. This assumes that aggregate trends in mismeasured production and nonproduction inputs track each other. We do this in practice because we have no data on nonproduction workers N in KLEMS.

To generate implied wages, we first scale labor hours (L) and labor expenditure (WL) to KLEMS aggregates. We then take the ratio of scaled labor expenditures to scaled labor quantities as our approximate wage.

A.1.3 Trims

Before running our share regression, we trim establishment-year observations with zero, missing, or negative revenue, production worker hours, capital, intermediates, nonproduction workers, and total production workers. We also trim all observations with inverse materials ratio $\frac{PQ}{CM}$ greater than the 99th percentile or lower than the 1st percentile of this ratio among all plants within each year. We also trim the 1st and 99th percentiles of $\frac{CM}{WL}$ by year. We do this to eliminate outliers in our estimation of μ and δ. These ratios are the data portion of the relevant ratio formulas for our market-power wedges. These $(p1, p99)$ ex ante trims remove establishments which have extremely high or low automation (e.g., high or low materials to labor expenditures) as well as those who report extremely high or low output relative to inputs (e.g. high or low revenue to materials expenditure). Such firms are unlikely to have sufficiently similar technologies, so should be excluded from our estimation groups.

A.2 CPS, ACS and Task Data

We import demographic data from the CPS and ACS into our ASM/CM sample. The CPS and ACS data were obtained from IPUMS. We pull the 1% metropolitan ACS sample in 1970, 1980 and 1990, as well as the 1% sample in 2000 and 2005-2016. We also pull the 5% state ACS sample in 1980, 1990, and 2000. From the ACS, we obtain occupation, education, and labor-force participation status. We limit our sample to individuals not in group quarters, not self-employed, prime-aged (18-64), and in the 50 continental states.

From the CPS, we pull CPS ASEC and monthly surveys for every year from 1988-2012. We again limit the CPS to individuals not in group quarters, not self-employed, prime-aged (18-64), and in the 50 continental states. We obtain union membership and coverage data from the CPS. Since the CPS sample is small relative to the ACS sample, we pool years into decadal years to correct for sampling or mismeasurement. We pool five years around each decadal year into the decadal year. Thus 1988-1992 are pooled into 1990, 1998-2002 into 2000, and 2008-2012 into 2010.

We merge the CPS union data with the ACS demographic data by state, occupation code (using the `occ1990` codes in the IPUMS data), year, industry (using the `ind1990` codes in the IPUMS data), and education category (segmenting education into no college, some college, and college graduate).

To get commuting-zone identifiers for our ACS and CPS data, we merge David Dorn's Public Use Microdata Area (PUMA) to commuting zone crosswalks for 1990, 2000, and 2010. For ACS/CPS observations with county codes, we use a county-to-commuting-zone crosswalk from the Census to identify CZs. For observations with no county code, we use county group (for 1980) or PUMA (1990, 2000, and 2010) crosswalks from David Dorn. We then reweight ACS person weights (perwt) using the adjustment factors provided in these crosswalks.

We merge data on task intensity from Autor and Dorn (2013) using occ1990dd occupational codes. This requires first merging in an occ1990 - occ1990dd crosswalk. Both the crosswalk and the task data are available on David Dorn's website. This task data contains manual, abstract, and routine task intensities by occ1990dd occupation. It also contains data on task offshorability.

We construct our measure of high routine-task intensity RTH as follows: first, we aggregate up manual, routine, and abstract task indicators to the industry (commuting-zone) level. Then, we winsorize abstract and manual tasks below the 66th percentile of each: we replace task_abstract by 0.4638442 if task_abstract ¡ 0.4638442 and we replace task_manual by 0.0340417 if task_manual ¡ 0.0340417. Next we construct the measure of routine-task intensity RTI as in Autor and Dorn (2013):

$$RTI = \log(\text{task\_routine}) - \log(\text{task\_manual}) - \log(\text{task\_abstract})$$

Finally, we generate RTH as an indicator for $(RTI > 1.826)$ (the 66th percentile of RTI).

Similarly, we construct OSH as an indicator for $task\_offshorability > 0.63$, the 66th percentile of the Autor and Dorn (2013) offshorability metric.

Our final ACS/CPS data contains demographic information, labor-force participation rates, union membership, and task/offshorability intensity data by commuting zone and ind1990 industry.

When merging our ACS/CPS data into our ASM/CM data, we aggregate the ASM/CM data at the geographical (CZ) or industry (ind1990) level first. To do the latter, we first create a NAICS6-ind1990 crosswalk using the files provided by the Census. With these two datasets, we can investigate the causes and consequences of labor-market power dynamics on the labor-supply side (geography) or on the labor-demand side (industry).

Appendix B

EMPIRICAL ESTIMATION APPENDIX

After constructing our data, we log and demean output, inputs, revenue, the share of revenue accounted for by intermediates expenditures, intermediates expenditures, and wages $\{Q, M, L, K, N, R, S, CM, W\}$. We demean each of these variables within industries to eliminate the need to include a constant in our estimation (after deameaning, each of these objects is now mean-zero within industry). Within each 6-digit Fort-Klimek NAICS industry j, we then run the regression:

$$s_j = \alpha_j + \beta_{m,j} m_j + \beta_{l,j} l_j + \beta_{n,j} n_j + beta_{k,j} k_j + \xi_I + \xi_T + \epsilon_j \tag{B.1}$$

Thus our estimation groups are the 473 NAICS-6 in the ASM/CM data: each NAICS-6 industry has its own (time-invariant) output elasticities. We experimented with alternative estimation groups with similar results. ξ_I and ξ_T are plant-year fixed effects. We use these constants to control (proxy) for markups, as explained in the main text.

We run our GMM procedure in Python, using the `numba` package to precompile all our functional definitions. We implement our CRS assumption by estimating output elasticities for $\{L, K, N\}$ and defining $\theta^M = 1 - \theta^L - \theta^K - \theta^N$. Any other returns-to-scale assumption may be used[1], but one is required to eliminate a moment from the identification problem (Klette and Griliches (1996) and Flynn et al. (2019)). In practice, we implement our RTS assumption by defining a modified log production function and functional form for the evolution of

1. We use CRS because it is a good approximation to RTS in the US manufacturing industry.

TFPR:

$$f = m_t + \mathbf{J}_t \cdot \theta^{\mathbf{j}}$$

$$g = \Big(m_{t-1} - \hat{s}_{t-1} - m_{t-1} - \mathbf{J}_{t-1} \cdot \theta^{\mathbf{j}} + \mathbf{S_g}\Big)^T \cdot \phi^{\mathbf{g}}$$

Where $\theta^{\mathbf{j}}$ is the vector of output elastictiies and $\phi^{\mathbf{g}}$ is the vector of coefficients in $g(\cdot)$. $\mathbf{S_g}$ are the state variables which determine the evolution of productivity through $g(\cdot)$. Recall that $g(\cdot)$ can take an arbitrary functional form, so the number of parameters in $\mathbf{S_g}$ is undetermined. Anything that researchers believe affects the evolution of the productivity process should in principle be included in the $g(\cdot)$ function. We experimented with including a constant, plant age, the number of plants in a firm, and a measure of the aggregate price index in $\mathbf{S_g}$. Our results are robust to these, but for simplicity the results in the main text only include a constant in the $\mathbf{S_g}$ vector.

We construct shocks η_t as from the residual:

$$\eta_t = r - f(\theta^{\mathbf{j}} \cdot \mathbf{J_t}) - g(\hat{s}_{t-1} - f(\theta^{\mathbf{j}} \cdot \mathbf{J_{t-1}}) - \hat{\epsilon}_t \qquad \text{(B.2)}$$

Where $\hat{\epsilon}_t$ comes from the share regression (B.1). This structural error should have $E[\eta_t = 0]$

under our model, so we can form the moment conditions in the main text:

$$\hat{\eta}_t \cdot \begin{bmatrix} 1 \\ w_{t-1} \\ n_{t-1} \\ k_{t-1} \\ w_{t-1}^2 \\ n_{t-1}^2 \\ k_{t-1}^2 \\ w_{t-1} \cdot n_{t-1} \\ w_{t-1} \cdot k_{t-1} \\ n_{t-1} \cdot k_{t-1} \\ (cm - \hat{s})_{t-1} \end{bmatrix} = 0$$

to recover the η as a function of the vector of elasticities $\theta^{\mathbf{j}}$. We then minimize $\hat{\eta}(\theta^{\mathbf{j}})' \cdot W^{-1} \cdot \hat{\eta}(\theta^{\mathbf{j}})$ by GMM. We use the identity matrix $W \equiv \mathbb{I}$ in our GMM estimation.

As in the main text, the form of these moment conditions comes from our translog production function specification. The quasi-fixed inputs n and k instrument for themselves since the model implies that they must be chosen before the start of each period. Wages instrument for labor.

We initialize the parameters at: $\{f^M, f^L, f^N, f^K, \rho\} = \{0.5, 0.1, 0.1, 0.3, 0.9\}$[2]. ρ is the AR-1 parameter in the evolution of productivity $g(\cdot)$, which we set at 0.9 after testing different starting values. We set each of the output elasticities at the aggregate cross-sectional average cost shares in the NBER-CES manufacturing database (that is, aggregated across all industries and years). We set the initial values of all cross- and squared-term coefficients in our translog specification to 0. Our results are robust to different starting values of these

2. f^j is the coefficient on j in the translog specification, while θ^j is j's output elasticity.

parameters.

We also constrain each output elasticity θ^j to be greater than the first percentile of input j's aggregate cross-sectional cost share in the NBER-CES manufacturing database (again aggregated across all years and industries). In particular, we set $\{\theta^M, \theta^L, \theta^N, \theta^K\} \geq \{0.25, 0.02, 0.01, 0.05\}$. This monotonicity restriction reduces outliers in our estimation (that is, extremely low estimated elasticities) and is quite conservative[3].

In addition, we impose a concavity restriction on the production function in the GMM problem by restricting the eigenvalues of the Hessian of our production function to be negative. The Hessian matrix of our problem is:

$$\mathbb{H} = \begin{pmatrix} f_{ml} + f_{mn} + f_{mk} & -f_{ml} & -f_{mn} & -f_{mk} \\ -f_{ml} & f_{ml} + f_{ln} + f_{lk} & -f_{ln} & -f_{lk} \\ -f_{mn} & -f_{ln} & f_{mn} + f_{ln} + f_{nk} & -f_{nk} \\ -f_{mk} & -f_{lk} & -f_{nk} & f_{mk} + f_{lk} + f_{nk} \end{pmatrix}$$

We restrict the eigenvalues of this matrix to be negative. To improve numerical stability, we add -0.00005 to all diagonal terms of $\mathbb{H}$ (not doing so can produce errors if the eigenvalues are sufficiently close to 0).

To implement our GMM minimization problem, we use the `scipy.optimize` package. This package estimates numerical derivatives using the quasi-Newton method of Broyden, Fletcher, Goldfarb, and Shanno (BFGS). We include both the monotonicity and concavity restrictions in the minimization command.

With `numba` pre-compilation, our estimation routine takes approximately 5-7 minutes.

3. In our gross-output production function θ^M is typically quite large, so 0.25 is not a binding restriction.

Appendix C

AGGREGATION APPENDIX

Our estimates of labor-market wedges and markups are at the firm level, but we would also like estimates of the aggregate labor-market wedge and markup. We define these aggregate objects in the same way as their firm-level counterparts, but computed with an aggregate production function. That is, $\mathcal{M} \equiv \Theta_M \frac{PQ}{cM}$ and $\mathcal{D} \equiv \frac{\Theta_L}{\Theta_M} \frac{cM}{wL}$, where the output elasticities are defined as $\Theta_L = \frac{\partial Y}{\partial L} \frac{L}{Y}$ and $\theta_\ell = \frac{\partial y}{\partial \ell} \frac{\ell}{y}$ and $\mathcal{D}$ represents the aggregate labor-market wedge. If the aggregate production function has the same elasticities as each individual production function, then we have:

$$
\begin{aligned}
\mu_i =& \theta_m^i \frac{p_i q_i}{c_i m_i} \\
\mu_i c_i m_i =& \theta_m^i p_i q_i \\
\int_i \mu_i c_i m_i di =& \int_i \theta_m^i p_i q_i di \\
\int_i \mu_i c_i m_i di =& \Theta_M \int_i p_i q_i di \\
\int_i \mu_i c_i m_i di =& \Theta_M PQ \\
\int_i \mu_i \frac{c_i m_i}{CM} di =& \Theta_M \frac{PQ}{CM} \\
\mathcal{M} =& \int_i \mu_i \frac{c_i m_i}{CM} di
\end{aligned}
$$

So that aggregate markups are a cost-weighted average of firm-level markups. Similarly, for labor-market wedges:

$$
\begin{aligned}
\delta_i =& \frac{\theta_\ell^i}{\theta_m^i} \frac{c_i m_i}{w_i \ell_i} \\
\delta_i \theta_m^i w_i \ell_i =& \theta_\ell^i c_i m_i \\
\int_i \delta_i \theta_m^i w_i \ell_i di =& \int_i \theta_\ell^i c_i m_i di \\
\Theta_M \int_i \delta_i w_i \ell_i di =& \Theta_L \int_i c_i m_i di \\
\int_i \delta_i w_i \ell_i di =& \frac{\Theta_L}{\Theta_M} CM \\
\int_i \delta_i \frac{w_i \ell_i}{WL} di =& \frac{\Theta_L}{\Theta_M} \frac{CM}{WL} \\
\mathcal{D} =& \int_i \delta_i \frac{w_i \ell_i}{WL} di
\end{aligned}
$$

In other words, aggregate labor-market wedges are a labor-cost-weighted average of firm-level labor-market wedges. If firms use heterogeneous technologies, then the aggregate markup becomes:

$$
\begin{aligned}
\frac{p_i q_i}{PQ} =& \frac{c_i m_i \mu_i}{\theta_m^i} \cdot \frac{\Theta_M}{CM\mathcal{M}} \\
1 =& \int_i \left(\frac{m_i}{M} \right) \left(\frac{\theta_m^i}{\Theta_M} \right)^{-1} \frac{\mu_i}{\mathcal{M}} di \\
\mathcal{M} =& \int_i \left(\frac{m_i}{M} \right) \left(\frac{\theta_m^i}{\Theta_M} \right)^{-1} \mu_i di
\end{aligned}
$$

Where the second line uses the fact that $\int_i p_i q_i di \equiv PQ$. The expression for the aggregate labor-market wedge is similar. Thus firm-level market-power wedges are weighted by both cost and the inverse of their relative contributions to the aggregate output elasticity. Firms with larger output elasticities get lower weights, because they increase inputs more when their output rises.

More generally, to aggregate either markups or labor-market wedges, we need a measure of aggregate output elasticities. This is a difficult problem. We have:

$$\begin{aligned}
\theta_i =& \frac{\partial y_i}{\partial \ell_i}\frac{\ell_i}{y_i} \\
\theta_i \partial \ell_i y_i =& \partial y_i \ell_i \\
\sum \theta_i \partial \ell_i y_i =& \sum_i \partial y_i \ell_i = \partial Y L + \sum_i cov(\partial y_i, \ell_i) \\
\frac{\partial Y}{\partial L}\frac{L}{Y} =& \sum_i \left[\theta_i \frac{\partial \ell_i}{\partial L}\frac{y_i}{Y} - \frac{cov(\partial y_i, \ell_i)}{\partial L \cdot Y} \right]
\end{aligned}$$

In other words, aggregate elasticities depend not only on individual elasticities, but also on size, marginal labor intensities, and covariances between labor intensities and marginal revenues. Without data on both the input-output matrix and on the degree to which increases in output are accomodated by each input[1], we cannot compute this object in general. Instead, we aggregate materials and labor output elasticities using input shares:

$$\Theta_M = \int_i \theta_m^i \frac{m_i}{M} di$$
$$\Theta_L = \int_i \theta_\ell^i \frac{\ell_i}{L} di$$

The thought experiment behind this is as follows: consider adding a small increment of intermediates (labor) into the economy at random. How does output change? In other words, we fix the distribution of labor and intermediates usage and consider the inputs needed to produce additional unit of input. This may be biased upward or downward depending on microeconomic (establishment-level) elasticities of substitution between m and ℓ and on the macro covariances between input usage and output expansion. We then define aggregate

1. These counterfactual elasticities are not even knowable, let alone obtainable from data.

labor-market wedges as above (and markups similarly):

$$\begin{aligned}\mathcal{D} =& \frac{\Theta_M}{\Theta_L}\frac{CM}{WL} \\ =& \frac{\int_i \theta_m^i \frac{m_i}{M} di}{\int_i \theta_\ell^i \frac{\ell_i}{L} di}\left(\frac{CM}{WL}\right)\end{aligned}$$

$$\mathcal{M} = \int_i \theta_m^i \frac{m_i}{M} di \left(\frac{PQ}{CM}\right)$$

Each time we compute labor-market wedges or markups at some aggregated level, we aggregate up labor and intermediates elasticities and sum expenditures and revenues at that level. These weights ignore the input-output structure and covariances between output growth and input use emphasized in Baqaee and Farhi (2019).

Appendix D

A NESTED CES APPROACH TO LABOR-MARKET POWER: APPENDIX

Here we lay out a workhorse general-equilibrium model which generates heterogeneous markups and markdowns at the firm level. The model has a nested constant elasticity of substitution structure in both output and labor markets.

D.1 Environment

The economy consists of a continuum of (representative) households and a continuum of firms. Households work, save, and consume, while firms produce output. Households get utility from consumption and leisure. Consumption is aggregated across s sectors, each of which has M_s subsectors. The household consists of many individual household members (it is a 'large family'). Each member supplies labor to firms.

There are a finite number of firms in each subsector[1]. Firms hire labor to work in labor markets in which the firm operates. Firms hire workers from households and sell output to households.

D.2 Households

Households like to consume and dislike work. Different types of households differ in their labor supply preferences. Denote each type by i. Preferences are given by:

$$U(C, L) = E_0 \sum_{t=0}^{\infty} \beta^t u(C_t) + v(L_t)$$

1. Which implies a continuum of firms, since there are continuum of subsectors.

Household consumption C_t is an aggregate bundle of consumption from each one of s sectors. There is a continuum of such sectors and households have a common elasticity of substitution θ between sectors, so that total consumption C_t is:

$$C_t = \left(\int_0^1 C_{st}^{\frac{\theta-1}{\theta}} ds \right)^{\frac{\theta}{\theta-1}}$$

Within each sector s, the consumption bundle C_s is another CES aggregate. Consumers have a (common) elasticity of substitution γ between goods within each sector. Each sector has a mass M_s of firms operating, and each firm produces one good within the sector. Sectoral consumption is then given by:

$$C_{st} = \left(\sum_{n=1}^{M_s} c_{st}(w_{ns})^{\frac{\gamma-1}{\gamma}} \right)^{\frac{\gamma}{\gamma-1}}$$

Where w_{ns} is an index of firm n in sector s (which will be the firm's wage rate). Households' aggregate disutility of labor supply is given by (again dropping the i subscripts for convenience):

$$L_t = \left(\int_0^1 L_{jt}^{\frac{\phi+1}{\phi}} \right)^{\frac{\phi}{\phi+1}}$$

Households supply labor in different markets j, with elasticity of labor supply ϕ between markets. Each firm employs workers in all of the labor markets M_{jt} in which it operates at time t. Household preferences over work in these labor markets can be decomposed as:

$$L_{jt} = \left(\sum_{i=1}^{M_j} \ell_{jt}^{\frac{\eta+1}{\eta}} \right)^{\frac{\eta}{\eta+1}}$$

So that households have elasticity of labor supply η within labor markets. As with product markets, we assume that $\phi < \eta$, so that households are more willing substitute within a labor market than across labor markets. This nested CES labor supply structure mirrors that in Berger et al. (2019). The across-market elasticity of substitution can be thought of as standing in for mobility costs, and the within-market elasticity stands in for job search costs within a market. This structure can also be micro-founded as workers in a large household acting on different idiosyncratic amenity values. These amenity values are distributed as nested Type-1 Extreme Value variables, giving the problem of labor market selection a nested logit form.

D.3 Elasticities and Market Power

The nested CES structure implies that firms can exercise some market power in both the output and labor markets. The inverse demand function for a firm j which chooses price p_j and output q_j in period t is:

$$\frac{p_{jt}}{P_t} = \left(\frac{q_{jt}}{Q_{st}}\right)^{-1/\gamma} \left(\frac{Q_{st}}{Q_t}\right)^{-1/\theta} \tag{D.1}$$

Where P_t is the aggregate price index, Q_{st} is the quantity produced in sector s in period t, Q_t is aggregate output, and q_{jt} is firm level output. Since the demand curve slopes downward, firms can exercise some market power. We assume that products within a sector are more substitutable than products across sectors, so that $\theta < \gamma$. This means that firms which are large in their sector (but small relative to the overall economy) face a smaller demand elasticity. These sector-dominating firms can exercise more market power. Put another way, the source of market power in the output market is the nature of preferences (e.g. differentiated goods). Markups only differ across firms due to their size.

Similarly, the labor supply curve can be written as:

$$\ell_{ijt} \propto \left(\frac{w_{ijt}}{W_{jt}}\right)^{\eta} \left(\frac{W_{jt}}{W_t}\right)^{\theta}$$

Where ijt indexes firm i in labor market j at time t. The source of labor market power is again preferences: workers do not view labor markets (or jobs within a market) as perfect substitutes, and firms can exploit this fact to pay workers less than their marginal products. Firms which are larger in local labor markets have higher monopsony power.

The nested CES product and labor markets imply elasticities of demand and inverse elasticities of labor supply[2] which are linear in market shares:

$$\epsilon^D_{ist} = \left[\theta s_{ist} + \gamma(1 - s_{ist})\right]$$
$$\epsilon^L_{ijt} = \left[\frac{1}{\phi} s_{ijt} + \frac{1}{\eta}(1 - s_{ijt})\right]^{-1}$$

Where s_{ist} is the share of firm i in product sector s, and s_{ijt} is the labor share of firm i in labor market j. Firms which are larger relative to their peers face the smaller across-sector (-market) elasticity, while firms which are smaller relative to their peers face the larger within-sector (-market) elasticity. Thus larger firms exert more market power: larger firms in product sectors s have higher markups, and larger firms in labor markets j have higher markdowns. To be precise, markups and markdowns are given by:

$$\mu_{ist} = \frac{\epsilon^D_{ist}}{\epsilon^D_{ist} - 1}$$
$$\delta_{ijt} = \frac{\epsilon^L_{ijt} + 1}{\epsilon^L_{ijt}}$$

D.4 Firms

Firms produce output by combining intermediates, labor and capital in a constant-returns-to-scale production function:

$$q_i = \omega_i k_i^{1-\beta_\ell-\beta_m} \ell_i^{\beta_\ell} m_i^{\beta_m}$$

2. The inverse elasticity of labor supply is $\frac{\partial w}{\partial \ell}\frac{\ell}{w}$ and is a more computationally straightforward object to work with than the elasticity of labor with respect to the wage.

Intermediates and capital are inelastically supplied. Firm i in labor market j faces an upward-sloping inverse labor supply curve:

$$w_{ijt} = \ell_{ijt}^{\epsilon_{ijt}^{L}} \tag{D.2}$$

Where ϵ_{ijt}^{L} is the inverse elasticity of labor supply as above.

D.5 Equilibrium

In equilibrium, each firm maximizes profits:

$$\pi(w, \epsilon^{L}) = \max_{k,\ell,m} pq - w\ell(w) - cm - rk$$

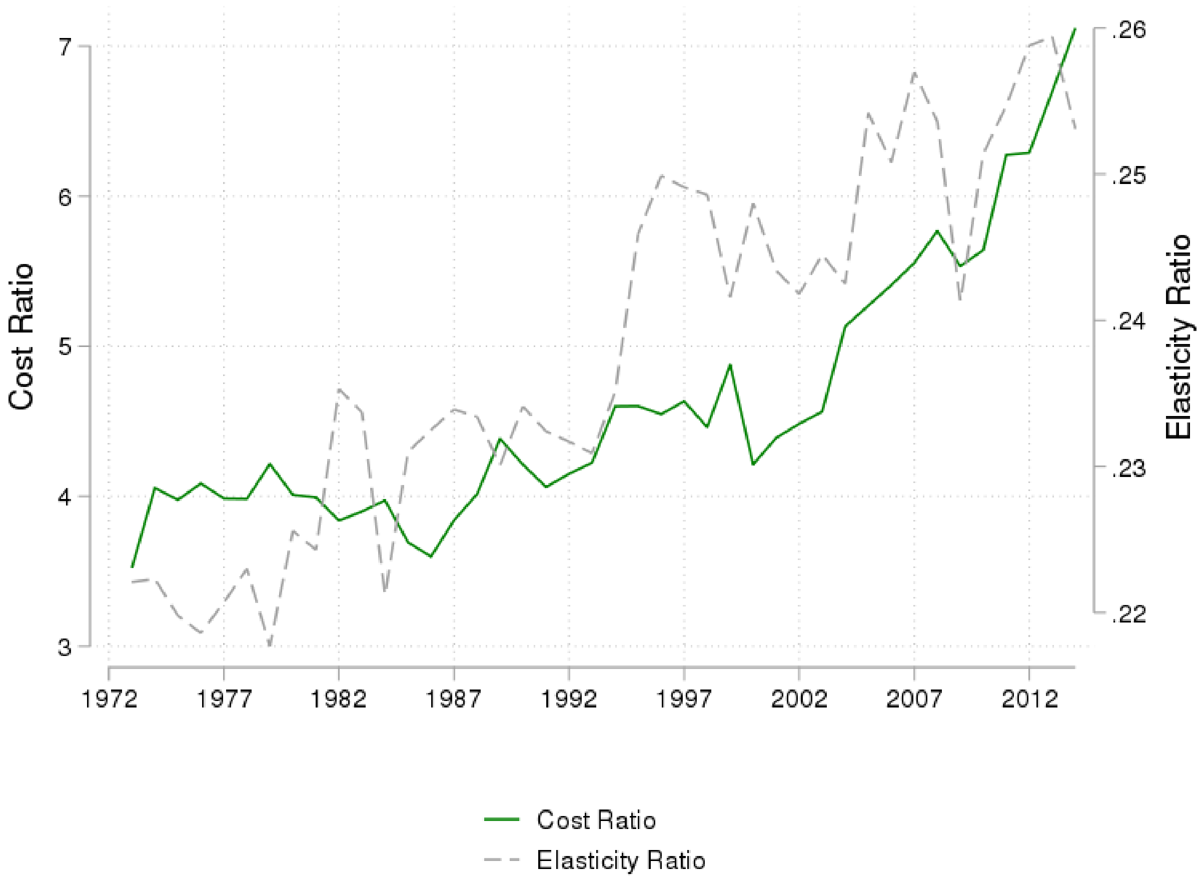
Subject to its demand curve (D.1) and labor supply curve (D.2) (We have suppressed the i, s, and j subscripts for readability). This is precisely identical to the cost minimization problem in Section 1.2 and yields identical first-order conditions.

We can then solve for firm level k, ℓ, m choices and aggregate them using market-clearing conditions. Aggregate variables will depend upon the demand elasticities θ, γ and the supply elasticity ϵ^{L}, as well as the distribution of ϵ_{ijt}^{L} and $cov(\delta_{ijt}, mu_{ist})$

Appendix E

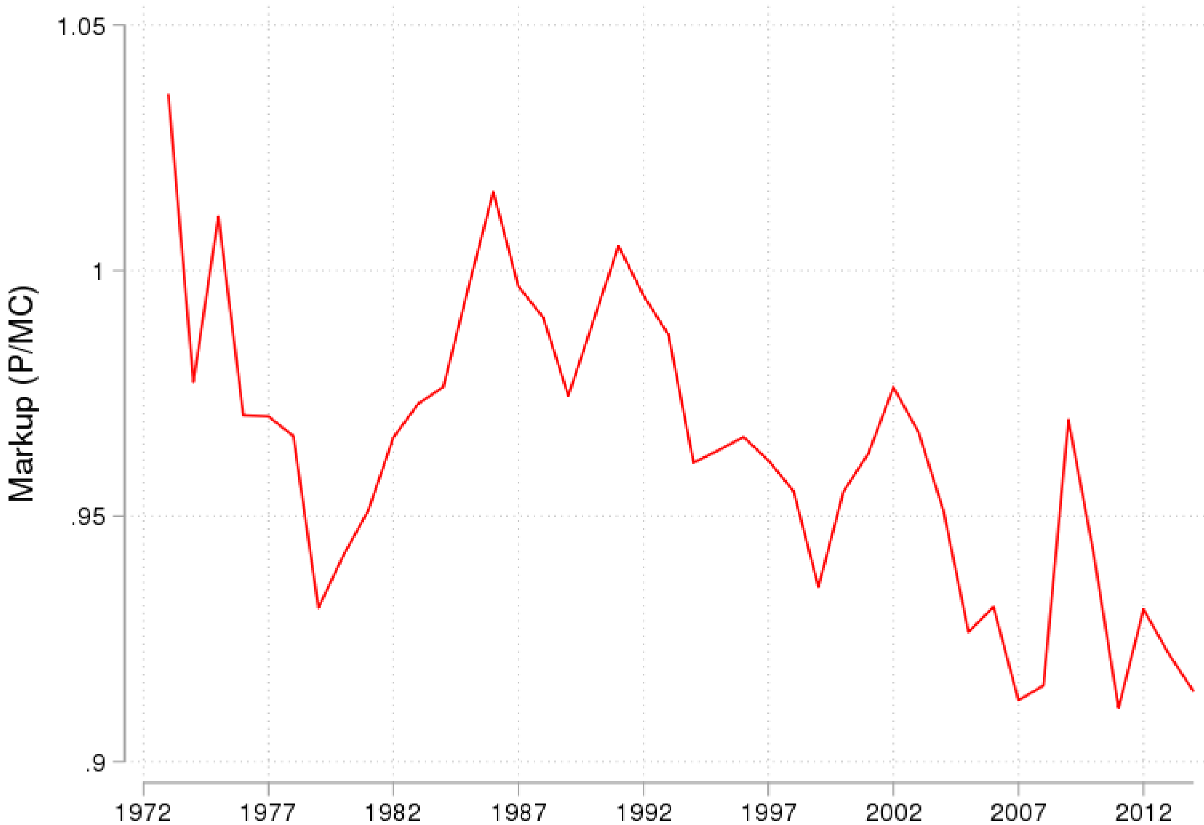
ADDITIONAL EMPIRICAL FIGURES APPENDIX

Figure E.1: Decomposition of Aggregate Labor-Market Wedge into Costs and Elasticities



Notes: Decomposition of labor-market wedge into cost ratio $\frac{C_t M_t}{W_t L_t}$ and elasticity ratio $\frac{\theta_L}{\theta_M}$ for each year. The cost and elasticity ratios are individually aggregated using labor expenditure weights and ASM sampling weights. Because of Jensen's inequality and the fact that there is some covariance between these objects, the product of the cost and elasticity ratios differ from the aggregate labor-market wedge $\mathcal{D}$.

Figure E.2: Aggregate Markup Wedge



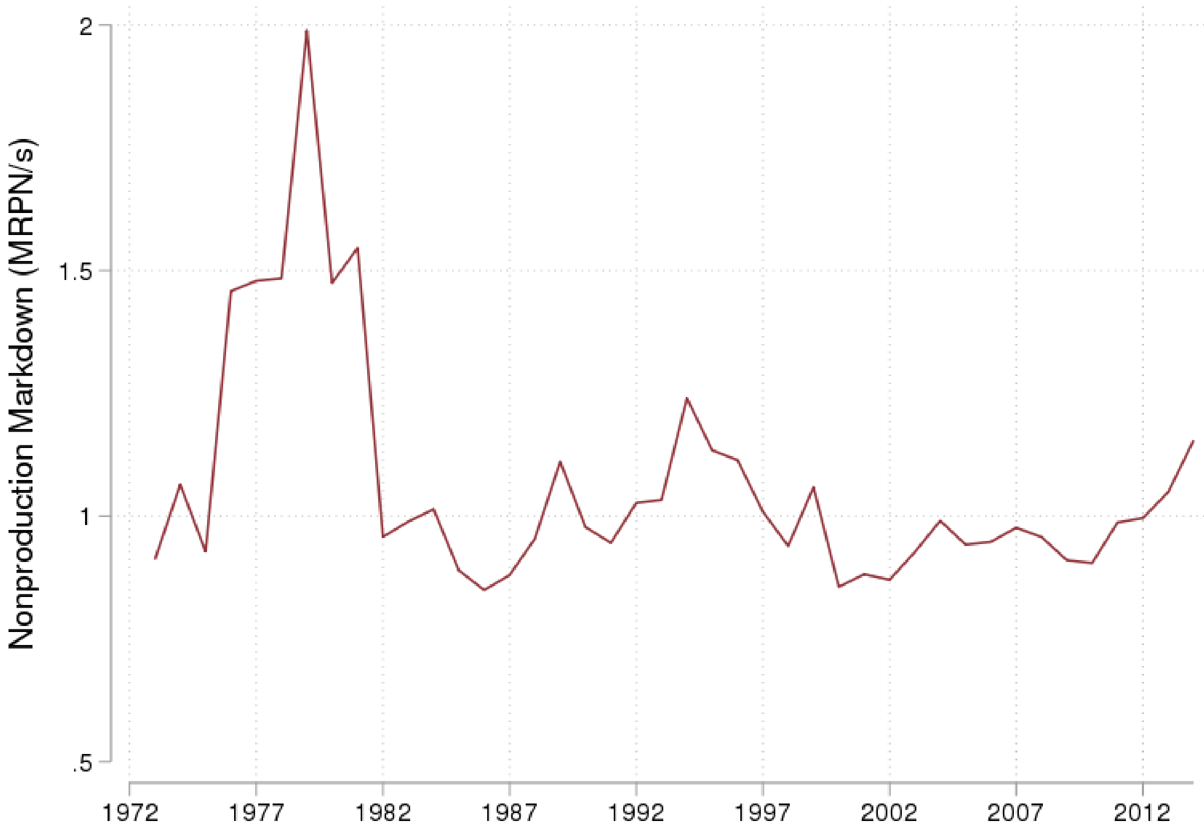
Notes: Markup power over time: Figure shows static aggregate markuip wedge $\mu \equiv \theta_M \cdot \left(\frac{P_t Q_t}{C_t M_t}\right)$ from 1973-2014 for US manufacturing. Aggregation is done using intermediate cost weights and ASM sample weight for each year (see Appendix C).

Table E.1: Additional Fixed-Effect Regression Results

	Year, NAICS6, Firm FEs	Year, NAICS6, Plant FEs	Year, CZ, NAICS6, Firm FEs	Year, CZ, NAICS6, Plant FEs
	(1)	(2)	(3)	(4)
	$\log(\delta)$	$\log(\delta)$	$\log(\delta)$	$\log(\delta)$
R^2	0.894	0.948	0.900	0.948
$R^2_{adj.}$	0.882	0.939	0.888	0.939
AIC	954,200	499,400	919,000	498,000

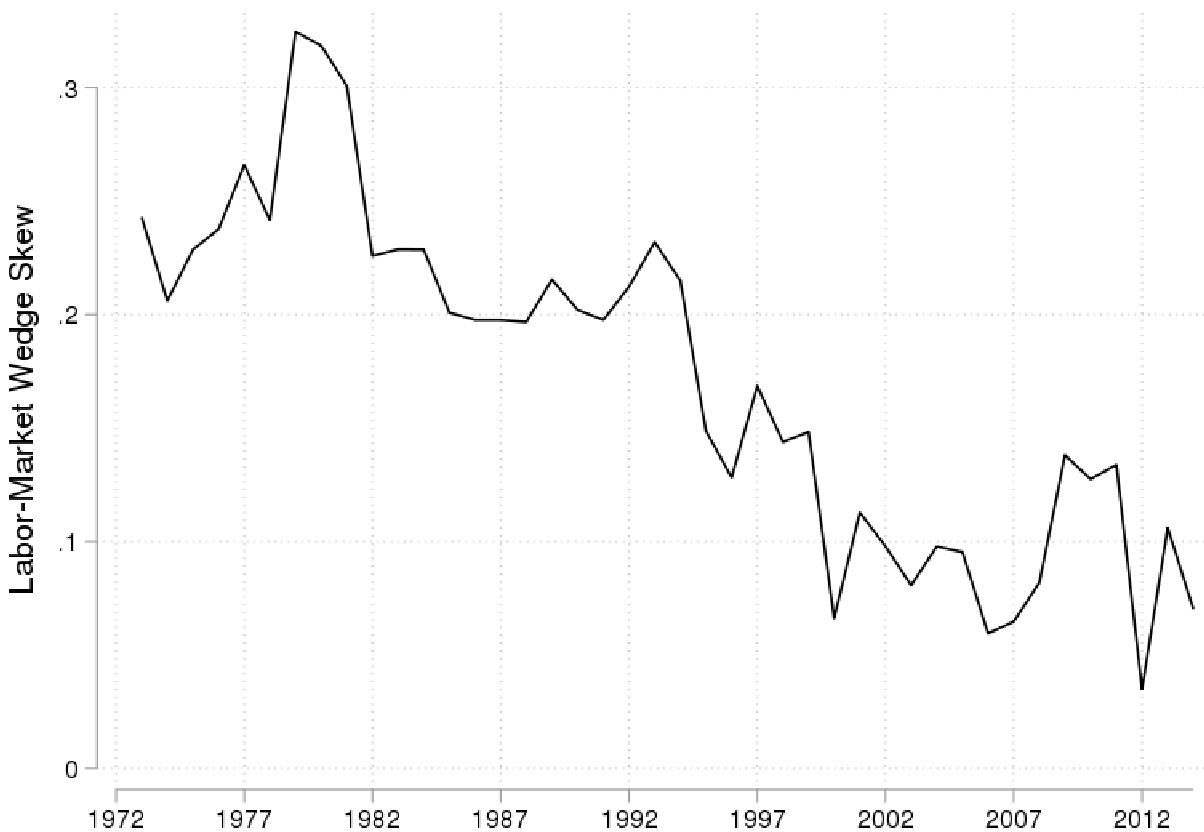
Labor-expenditure- and ASM-sampling-weight- weighted regression of $\log(\delta)$ on different fixed effects. Including Firm f or plant i fixed-effects implies that regressions are in changes.
Source: ASM/CMF.

Figure E.3: Aggregate Nonproduction Wedge



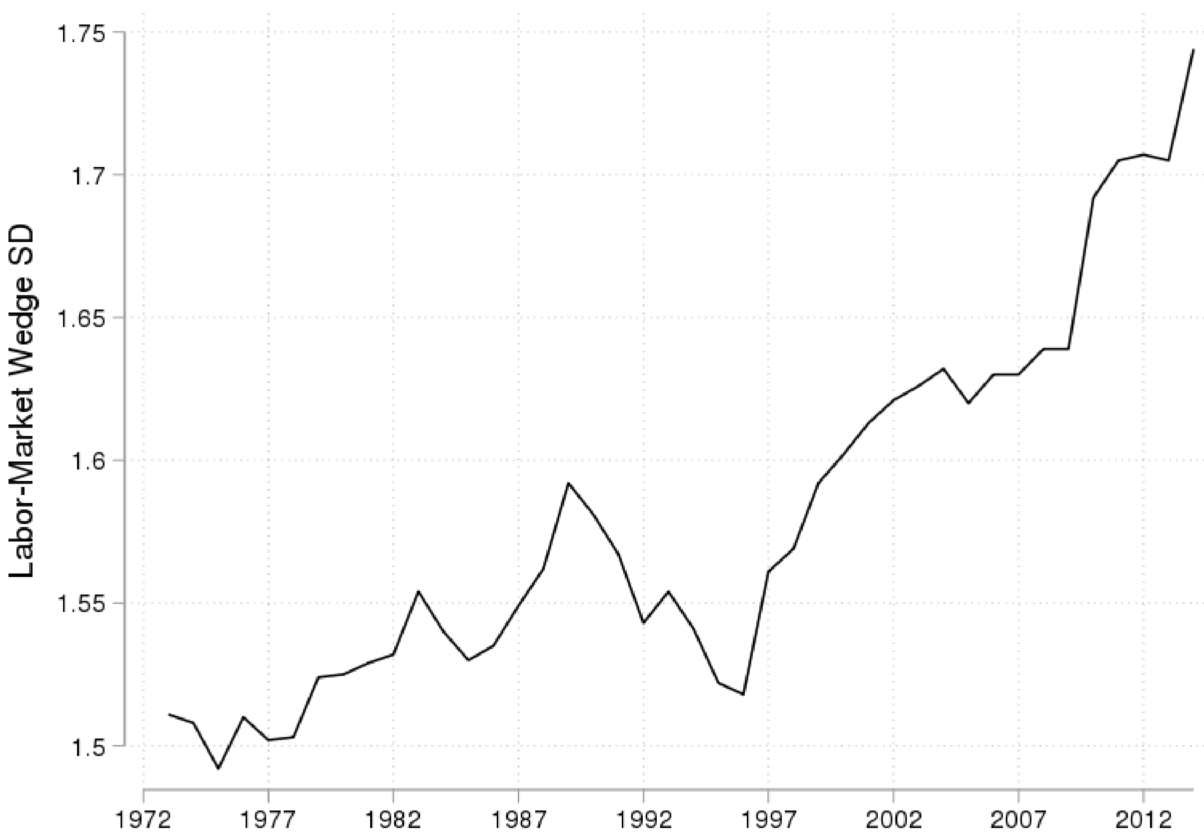
Notes: Nonproduction market power over time: Figure shows static aggregate nonproduction input wedge $\delta_N \equiv \frac{\theta_N}{\theta_M} \cdot \left(\frac{C_t M_t}{S_t N_t}\right)$ over time for US manufacturing. Aggregation is done using nonproduction cost weights and ASM sample weight for each year for each year (see Appendix C).

Figure E.4: Skewness of Aggregate Labor-Market Wedge (Time Series)



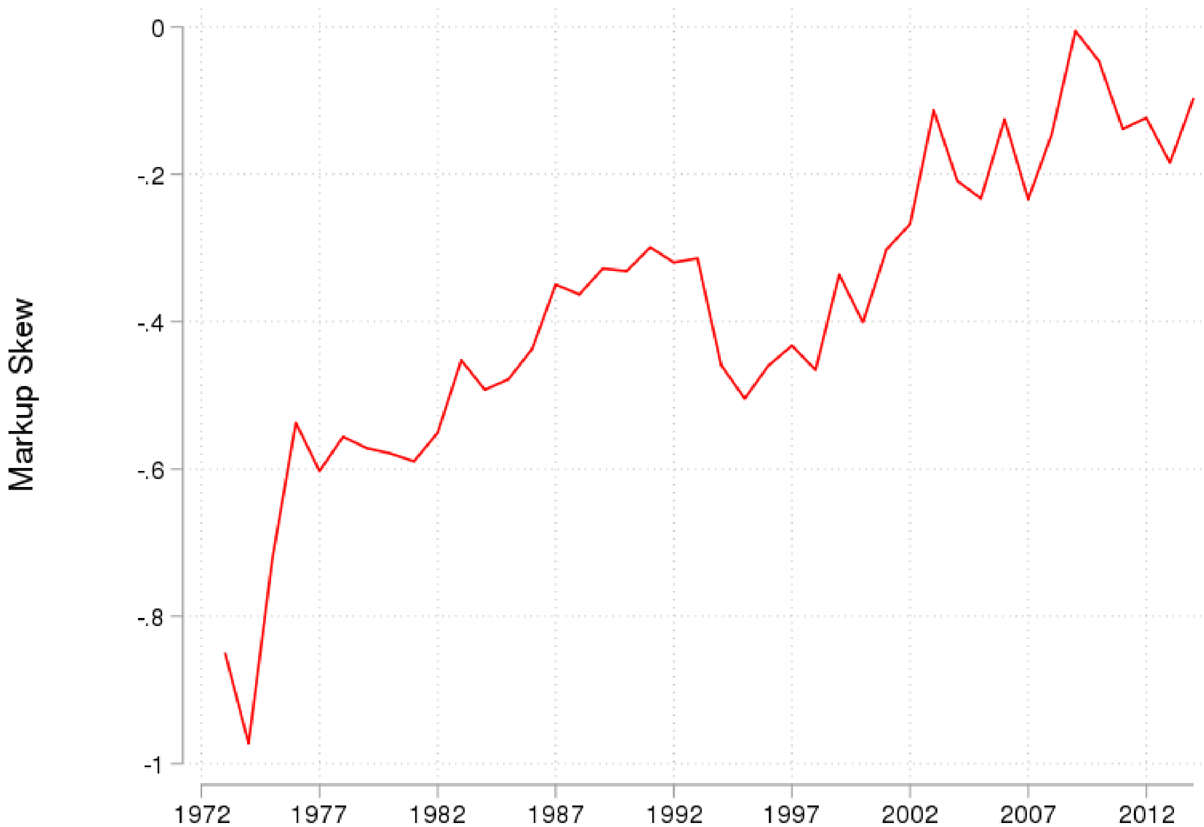
Notes: Time series of the skewness of labor-market wedges over time (1973-2014). Skewness is computed in each year across the distribution of labor-market wedges δ_t using labor expenditure weights and ASM sampling weights.

Figure E.5: Standard Deviation of Aggregate Labor-Market Wedge (Time Series)



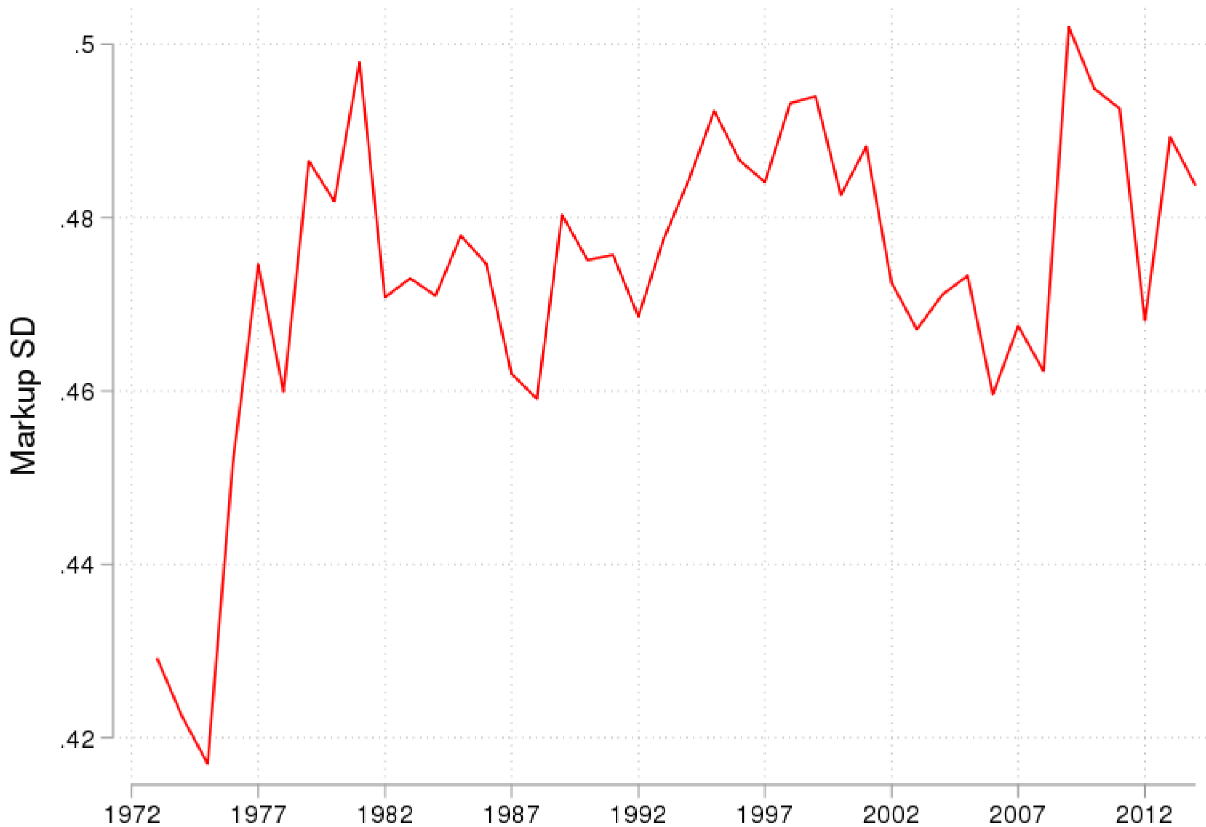
Notes: Time series of the standard deviation of labor-market wedges over time (1973-2014). Standard deviation is computed in each year across the distribution of labor-market wedges δ_t using labor expenditure weights and ASM sampling weights.

Figure E.6: **Skewness of Aggregate Markup Wedge (Time Series)**



Notes: Time series of the skewness of markups over time (1973-2014). Skewness is computed in each year across the distribution of markups μ_t using intermediates expenditure weights and ASM sampling weights.

Figure E.7: Standard Deviation of Aggregate Markup Wedge (Time Series)



Notes: Time series of the standard deviation of markups over time (1973-2014). Standard deviation is computed in each year across the distribution of markups μ_t using intermediates expenditure weights and ASM sampling weights.

Table E.2: Additional Fixed-Effect Regressions of Wages, Labor Inputs, and Labor Expenditures on Labor-Market Wedge

	Year, NAICS6, Firm FEs			Year, NAICS6, Plant FEs		
	(1)	(2)	(3)	(4)	(5)	(6)
	w	ℓ	wl	w	ℓ	wl
$\log(\delta)$	-0.0640***	-0.313***	-0.372***	-0.107***	-0.247***	-0.346***
	(0.000673)	(0.000175)	(0.00181)	(0.000814)	(0.00133)	(0.00128)
Constant	0.160***	-0.019***	7.74***	0.128***	0.0483***	7.783***
	(0.00065)	(0.00168)	(0.00174)	(0.00071)	(0.00116)	(0.00111)
N	621000	621000	621000	621000	621000	621000
R^2	0.644	0.655	0.776	0.783	0.916	0.953
AIC	184,100	1,369,000	1,411,000	-124,600	479,900	431,100

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

(1)-(3) Year, NAICS6, firm fixed effects; (4)-(6) Year, NAICS6, plant fixed effects.

Labor-expenditure- and ASM-sampling-weight- weighted regression of log wages, labor inputs, and labor expenditures on log labor-market wedge δ. Regressions including firm (plant) FEs are implicitly in changes.

Source: ASM/CMF.

Table E.3: Additional Fixed Effects for Regression of Labor-Market Wedge on Direct Measures of Technology

	Year, NAICS6, Firm FEs		NAICS6, Firm FEs
	(1)	(2)	(3)
	$\log(\delta)$	$\log(\delta)$	$\log(\delta)$
New Computer Expenditures/Worker	0.0188***		
(00-01)	(0.00438)		
Purchased Communications/Worker		0.0782***	
(97, 02, 06-14)		(0.00162)	
Software Expenditures/Worker (97)			0.0366***
			(0.00971)
N	9,600	147,000	2,700
R^2	0.950	0.912	0.935
AIC	7,700	210,800	2,831

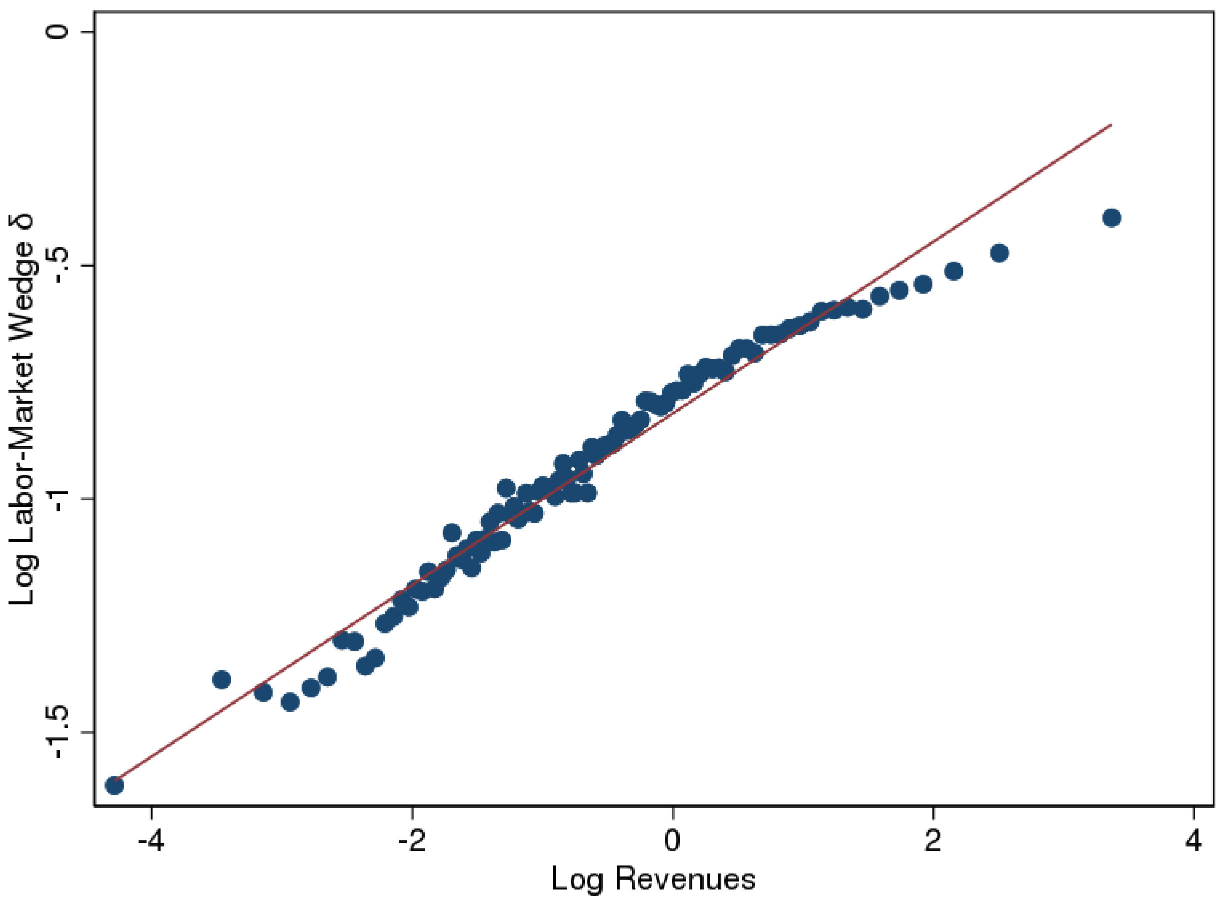
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Labor-expenditure- and ASM-sampling-weight- weighted regression of log labor-market wedge δ on log new computer expenditures per worker (row 1), log cost of purchased communications per worker (row 2), and log software expenditures per worker (row 3). New computer expenditures are available 2000-2001, software expenditures 1997, cost of purchased communications 1997, 2002, 2006-2014.

Technological variables are deflated and normalized by (workers) $\equiv$ (production workers + other employees).

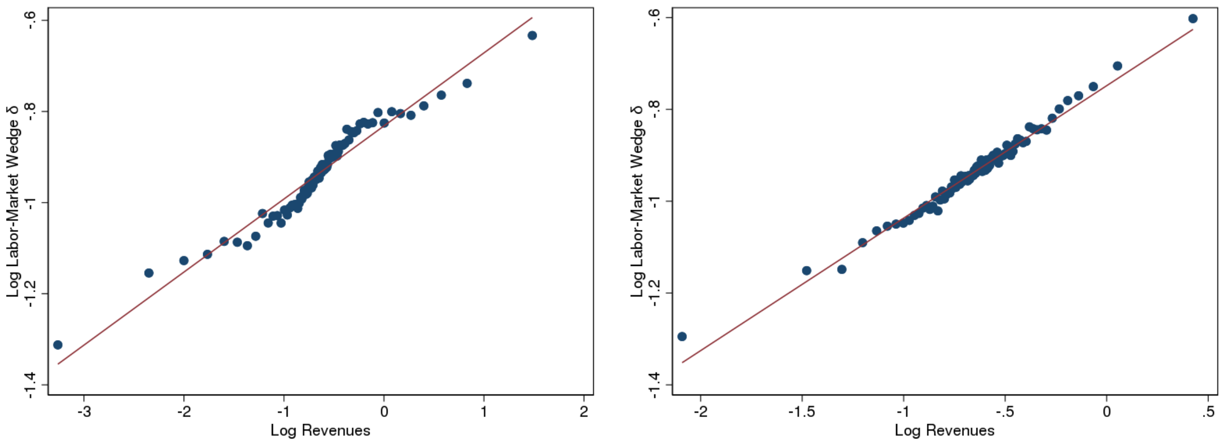
Source: ASM/CMF.

Figure E.8: Binscatter of Labor-Market Wedge against Plant Revenues



Notes: Binscatter of log(δ) against log revenues at the plant level with year, NAICS6 (ft, right) fixed effects. Revenues are scaled to KLEMS aggregates by year and dja industry. Figure residualizes log(δ) and revenues against jt fixed-effects, then bins log($delta$) into 100 equally-sized bins and plots the mean of revenues within each bin along with a (red) line-of-best-fit. Figure is weighted by ASM sampling weight and labor expenditures.

Figure E.9: Binscatter of Labor-Market Wedge against Plant Revenues (Additional Fixed Effects)



(a) Year, NAICS6, Firm FEs

(b) Year, NAICS6, Plant FEs

Notes: Binscatters of $\log(\delta)$ against log revenues at the plant level with year, NAICS6, firm (fjt, left) and year, NAICS6, plant (ijt, right) fixed effects. Revenues are scaled to KLEMS aggregates by year and dja industry. Each figure residualizes $\log(\delta)$ and revenues against t or jt fixed-effects, then bins $\log(delta)$ into 100 equally-sized bins and plots the mean of revenues within each bin along with a (red) line-of-best-fit. Figure is weighted by ASM sampling weight and labor expenditures. Including firm (plant) FEs makes this implicitly a regression in changes.

Table E.4: More Additional Fixed Effects for Regression of Labor-Market Wedge on Direct Measures of Technology

	Year, NAICS6, Plant FEs	
	(1)	(2)
	$\log(\delta)$	$\log(\delta)$
New Computer Expenditures/Worker (00-01)	0.0114***	
	(0.00352)	
Purchased Communications/Worker (97, 02, 06-14)		0.0191***
		(0.00138)
N	8,200	144,000
R^2	0.989	0.96
AIC	5,515	90,320

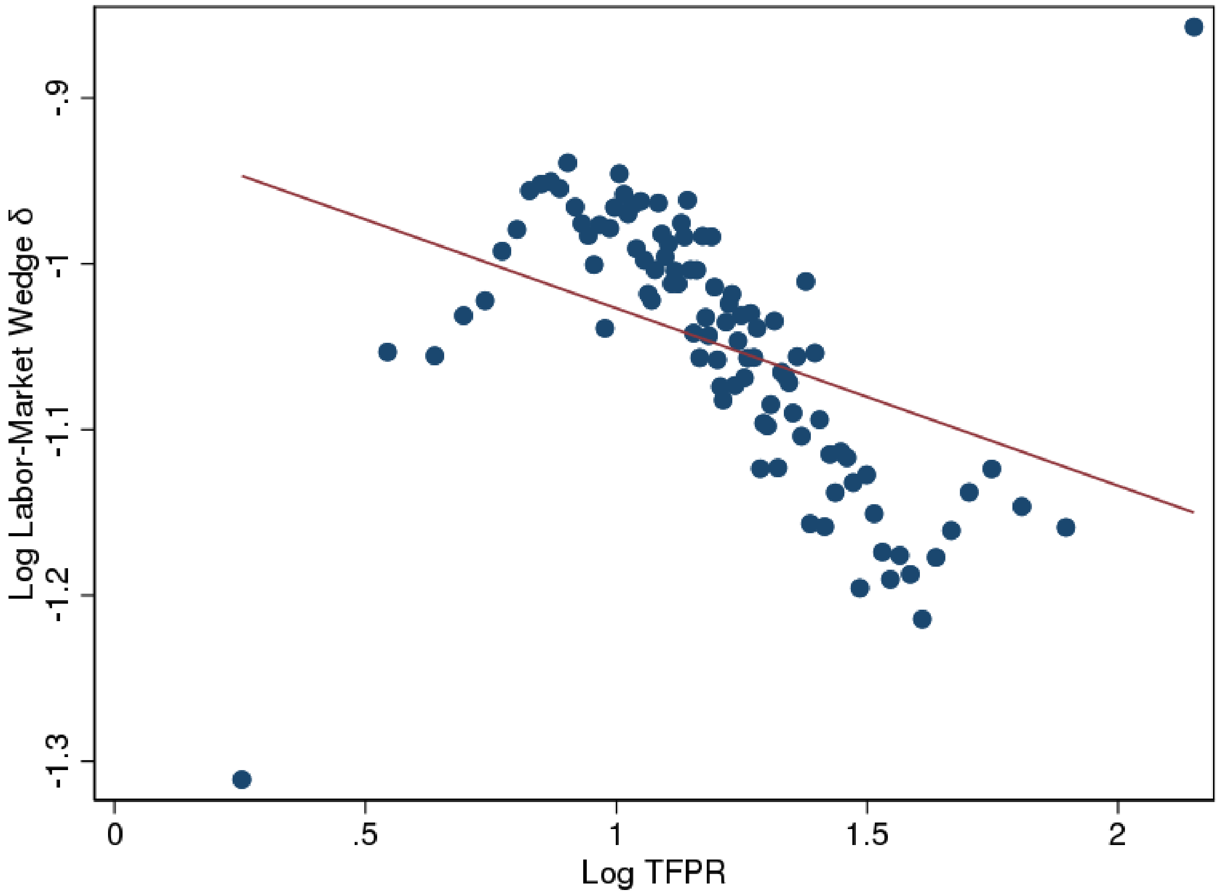
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Labor-expenditure- and ASM-sampling-weight- weighted regression of log labor-market wedge δ on log new computer expenditures per worker (row 1) and log cost of purchased communications per worker (row 2). New computer expenditures are available 2000-2001, cost of purchased communications 1997, 2002, 2006-2014.

Technological variables are deflated and normalized by (workers) $\equiv$ (production workers + other employees).

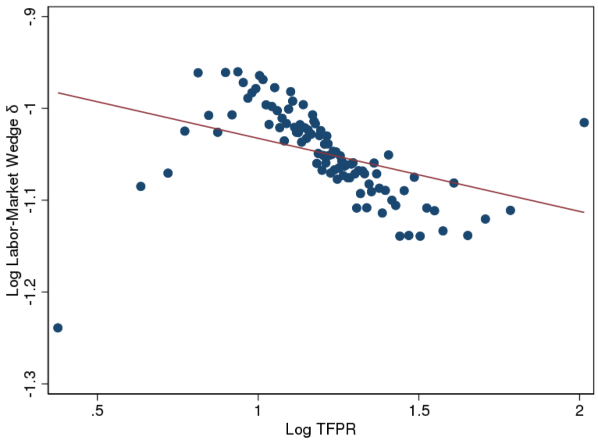
Source: ASM/CMF.

Figure E.10: Binscatter of Labor-Market Wedge against Plant TFPR

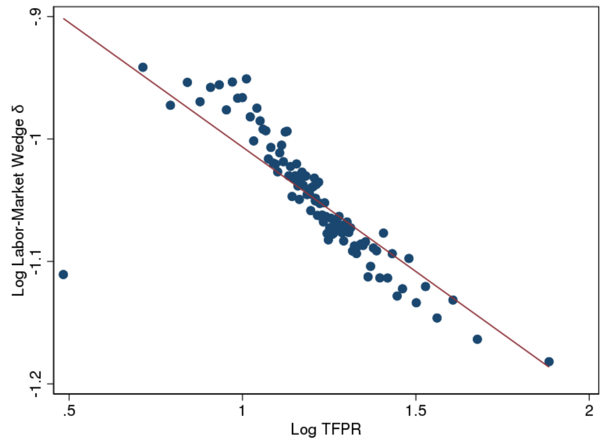


Notes: Binscatter of log(δ) against log revenue total factor productivity (TFPR) at the plant level with year, NAICS6 (ft, right) fixed effects. TFPR is an output of our estimation routine. Figure residualizes log(δ) and TFPR against or jt fixed-effects, then bins log($delta$) into 100 equally-sized bins and plots the mean of TFPR within each bin along with a (red) line-of-best-fit. Figure is weighted by ASM sampling weight and labor expenditures.

Figure E.11: Binscatter of Labor-Market Wedge against Plant TFPR (Additional Fixed Effects)



(a) Year, NAICS6, Firm FEs



(b) Year, NAICS6, Plant FEs

Notes: Binscatters of $\log(\delta)$ against log revenue total factor productivity (TFPR) at the plant level with year, NAICS6, firm (fjt, left) and year, NAICS6, plant (ijt, right) fixed effects. TFPR is an output of our estimation routine. Each figure residualizes $\log(\delta)$ and TFPR against t or jt fixed-effects, then bins $\log(delta)$ into 100 equally-sized bins and plots the mean of TFPR within each bin along with a (red) line-of-best-fit. Figure is weighted by ASM sampling weight and labor expenditures. Including firm (plant) FEs makes this implicitly a regression in changes.

Table E.5: Additional Fixed Effects for Regression of Labor-Market Wedge on Input Intensity Measures

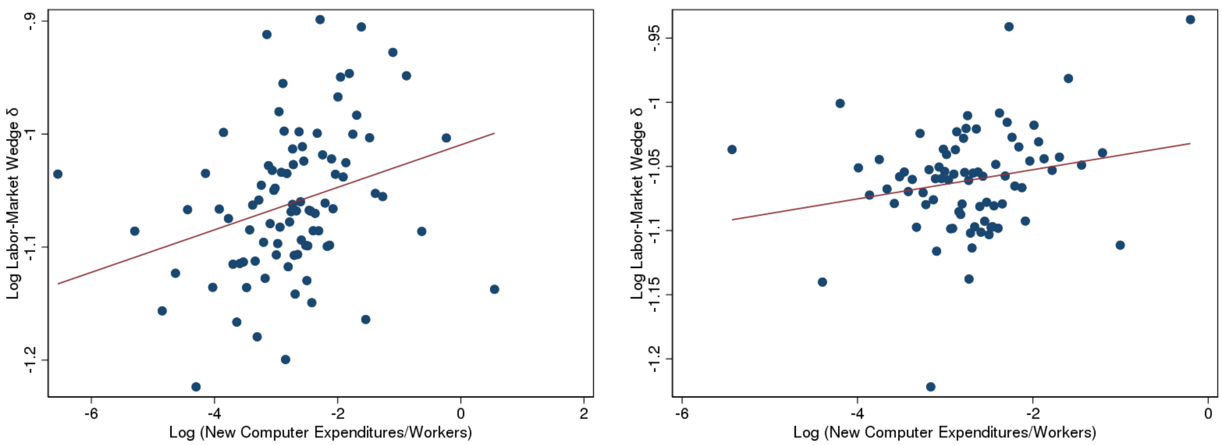
	Year, NAICS6, Firm FEs		
	(1)	(2)	(3)
	$\log(\delta)$	$\log(\delta)$	$\log(\delta)$
$\log\left(\frac{K}{N+L}\right)$	0.169***		0.122***
	(0.00135)		(0.00138)
$\log\left(\frac{N}{N+L}\right)$		0.211***	0.178***
		(0.00136)	(0.00140)
N	621,000	621,000	621,000
R^2	0.897	0.899	0.900
AIC	936,900	927,900	919,200

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Labor-expenditure- and ASM-sampling-weight- weighted regression of log labor-market wedge δ on log capital intensity (column 1), log nonproduction intensity (column 2), and both intensities jointly (column 3). Input intensity in each case is defined as $\frac{X}{N+L}$, with X the input in question and $N + L$ total plant employment.

Source: ASM/CMF.

Figure E.12: Binscatter of Labor-Market Wedge against New Computer Expenditures Per Worker (Additional Fixed Effects)

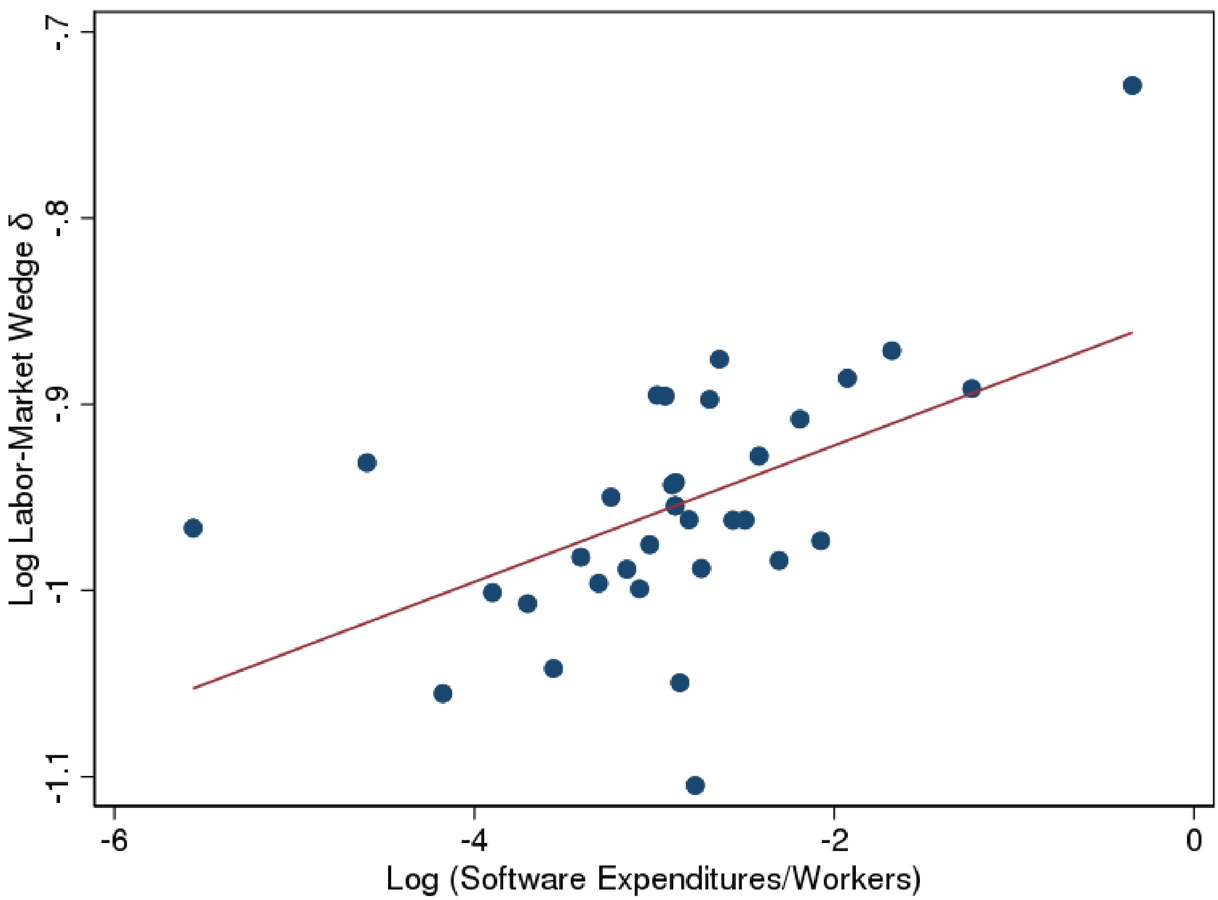


(a) Year, NAICS6, Firm FEs

(b) Year, NAICS6, Plant FEs

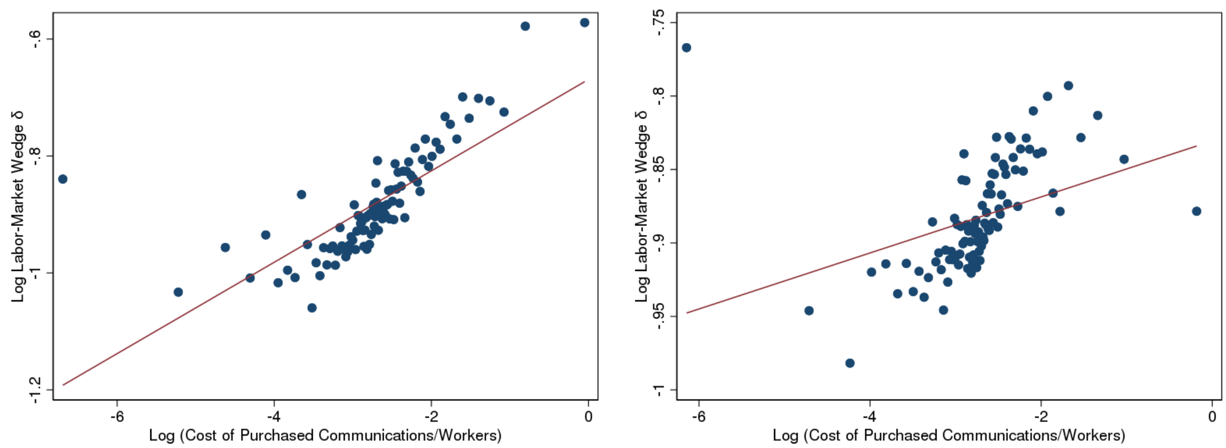
Notes: Binscatter of $\log(\delta)$ against log new computer expenditures per worker $\log\left(\frac{nmc}{workers}\right)$ at the plant level with year, NAICS6, & firm (fjt) fixed effects (left) and year, NAICS6, & plant fixed effects (right). The figure residualizes $\log(\delta)$ and per-worker computer expenditures against each set of fixed effects, then bins each into 100 equally-sized bins and plots the mean of $\log(\delta)$ within each bin along with a (red) line-of-best-fit. Workers are defined as production workers plus nonproduction workers. New computer expenditures are deflated using a NAICS6 NBER-CES shipments price index. Figure is weighted by ASM sampling weight and labor expenditures. For years 2000-2001 due to data availability.

Figure E.13: Binscatter of Labor-Market Wedge against Software Expenditures Per Worker (Additional Fixed Effects)



Notes: Binscatter of $\log(\delta)$ against log software expenditures per worker $\log\log\left(\frac{cs}{workers}\right)$ at the plant level with NAICS6 (fj) & firm fixed effects. The figure residualizes $\log(\delta)$ and per-worker software expenditures against NAICS6 and firm fixed-effects, then bins each into 100 equally-sized bins and plots the mean of $\log(\delta)$ within each bin along with a (red) line-of-best-fit. Workers are defined as production workers plus nonproduction workers. Software expenditures are deflated using a NAICS6 NBER-CES shipments price index. Figure is weighted by ASM sampling weight and labor expenditures. For year 1997 due to data availability.

Figure E.14: Binscatter of Labor-Market Wedge against Communications Expenditures Per Worker (Additional Fixed Effects)



(a) Year, NAICS6, Firm FEs

(b) Year, NAICS6, Plant FEs

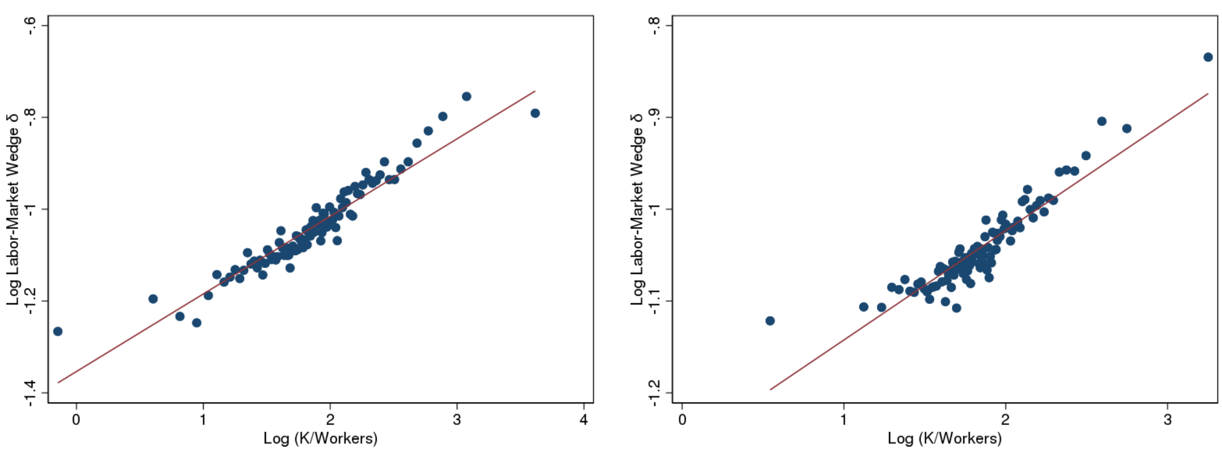
Notes: Binscatter of $\log(\delta)$ against log cost of purchased communications per worker $\log\left(\frac{cpc}{workers}\right)$ at the plant level with year, NAICS6, & firm (fjt) fixed effects (left) and year, NAICS6, & plant fixed effects (right). The figure residualizes $\log(\delta)$ and per-worker communications expenditures against each set of fixed effects, then bins each into 100 equally-sized bins and plots the mean of $\log(\delta)$ within each bin along with a (red) line-of-best-fit. Workers are defined as production workers plus nonproduction workers. Communications purchases are deflated using a NAICS6 NBER-CES shipments price index. Figure is weighted by ASM sampling weight and labor expenditures. For years 1997, 2002, and 2006-2014 due to data availability.

Table E.6: More Additional Fixed Effects for Regression of Labor-Market Wedge on Input Intensity Measures

	Year, NAICS6, Plant FEs		
	(1)	(2)	(3)
	$\log(\delta)$	$\log(\delta)$	$\log(\delta)$
$\log\left(\frac{K}{N+L}\right)$	0.119***		0.0772***
	(0.00141)		(0.00145)
$\log\left(\frac{N}{N+L}\right)$		0.169***	0.150***
		(0.00130)	(0.00135)
N	614,000	614,000	614,000
R^2	0.949	0.950	0.950
AIC	495,300	484,100	480,800

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$
Labor-expenditure- and ASM-sampling-weight- weighted regression of log labor-market wedge δ on log capital intensity (column 1), log nonproduction intensity (column 2), and both intensities jointly (column 3). Input intensity in each case is defined as $\frac{X}{N+L}$, with X the input in question and $N + L$ total plant employment.
Source: ASM/CMF.

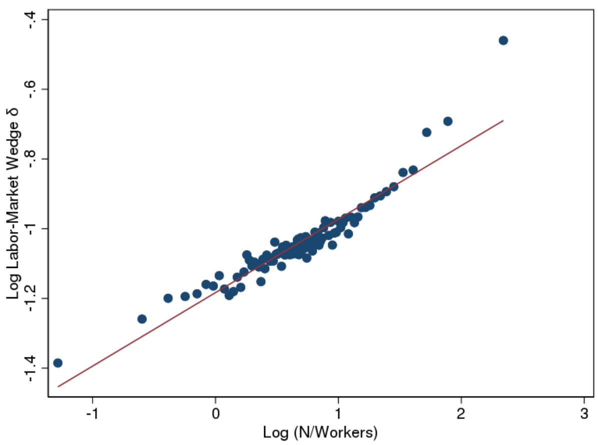
Figure E.15: Binscatter of Labor-Market Wedge against Plant Capital Intensity (Additional Fixed Effects)



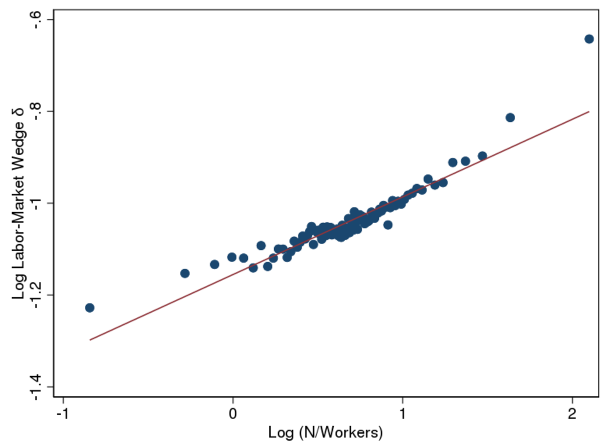
(a) Year, NAICS6, Firm FEs **(b)** Year, NAICS6, Plant FEs

Notes: Binscatter of $\log(\delta)$ against log capital per worker $\log\left(\frac{K}{N+L}\right)$ at the plant level with year, NAICS6, & firm (fjt) fixed effects (left) and year, NAICS6, & plant fixed effects (right). The figure residualizes $\log(\delta)$ and per-worker capital intensity against each set of fixed effects, then bins each into 100 equally-sized bins and plots the mean of $\log(\delta)$ within each bin along with a (red) line-of-best-fit. Workers are defined as production workers plus nonproduction workers. Figure is weighted by ASM sampling weight and labor expenditures.

Figure E.16: Binscatter of Labor-Market Wedge against Plant Nonproduction Intensity (Additional Fixed Effects)



(a) Year, NAICS6, Firm FEs



(b) Year, NAICS6, Plant FEs

Notes: Binscatter of $\log(\delta)$ against log nonproduciton workers per worker $\log\left(\frac{N}{N+L}\right)$ at the plant level with year, NAICS6, & firm (fjt) fixed effects (left) and year, NAICS6, & plant fixed effects (right). The figure residualizes $\log(\delta)$ and per-worker nonproduction intensity against each set of fixed effects, then bins each into 100 equally-sized bins and plots the mean of $\log(\delta)$ within each bin along with a (red) line-of-best-fit. Workers are defined as production workers plus nonproduction workers. Figure is weighted by ASM sampling weight and labor expenditures.

Appendix F

ELASTICITY ALGEBRA APPENDIX

F.1 Output and Revenue Elasticities

This appendix provides the algebra for the relationships between elasticities, and the relationship between elasticities and markups. First, we derive the relationship between revenue and quantity elasticities in (2.1). Consider an input to production L. Define the elasticity of quantity produced respect to L as $f_l \equiv \frac{\partial F(\cdot)}{\partial L}\frac{L}{Q}$ and the elasticity of revenue as $r_l \equiv \frac{\partial R(\cdot)}{\partial L}\frac{L}{R}$. Finally, define the elasticity of output price with respect to quantity $\epsilon^P \equiv \frac{dP}{dQ}\frac{Q}{P}$By definition, revenue $R = P \cdot Q$. Totally differentiating this expression with respect to L yields:

$$
\begin{aligned}
R &= P \cdot Q \\
\frac{dR}{dL} &= Q\frac{dP}{dL} + P\frac{dQ}{dL} \\
\frac{dR}{dL}\frac{L}{R} &= Q\frac{dP}{dL}\frac{L}{R} + P\frac{dQ}{dL}\frac{L}{R} \\
\frac{dR}{dL}\frac{L}{R} &= \frac{dP}{dL}\frac{LQ}{PQ} + \frac{dQ}{dL}\frac{LP}{QP} \\
\frac{dR}{dL}\frac{L}{R} &= \frac{dP}{dL}\frac{L}{P} + \frac{dQ}{dL}\frac{L}{Q} \\
\frac{dR}{dL}\frac{L}{R} &= \frac{Q}{P}\frac{P}{Q}\frac{dQ}{dL}\frac{dL}{dQ}\frac{dP}{dL}\frac{L}{P} + \frac{dQ}{dL}\frac{L}{Q} \\
\frac{dR}{dL}\frac{L}{R} &= \frac{dP}{dQ}\frac{Q}{P}\frac{dQ}{dL}\frac{L}{Q} + \frac{dQ}{dL}\frac{L}{Q} \\
\frac{dR}{dL}\frac{L}{R} &= \frac{Q}{P}\frac{dQ}{dL}\frac{dP}{dQ}\frac{L}{Q} + \frac{dQ}{dL}\frac{L}{Q} \\
r_l &= (1+\epsilon^P)f_l
\end{aligned}
$$

Which is (2.1) in the main text.

F.2 Price Elasticity and the Markup

Next, we show the relationship between the inverse elasticity of demand ϵ^P and the markup $\mathcal{M}$. Assume firms maximize profits and minimize costs. The profit-maximization problem is:

$$\max_Q P(Q) \cdot Q - C(Q)$$

Where $P = P(Q)$ because of imperfectly elastic demand, and $C(Q)$ is the firm's cost function. The FOC of the profit-maximization problem is:

$$\frac{dP}{dQ}Q + P = C'(Q)$$
$$\frac{dP}{dQ}\frac{Q}{P} + 1 = \frac{C'(Q)}{P}$$
$$1 + \epsilon^P = \frac{1}{\mathcal{M}}$$

Where the last line uses the definition $\mathcal{M} \equiv \frac{MC}{P} = \frac{C'(Q)}{P}$. Rearranging this result yields $\mathcal{M} = (1 + \epsilon^P)^{-1}$, which is the result in the main text.