Increasing Returns to Scale and Markups *

Olga Shanks
ostaradu@gmu.edu
George Mason University

December 22, 2023

Abstract

I estimate aggregate and industry-specific elasticities of scale and markups for the U.S. economy over the period from 1980 to 2019 using data on publicly traded companies. I apply Olley-Pakes and Ackerberg-Caves-Frazer estimation methods and find that the aggregate elasticity of scale for the U.S. economy is 1.1 and has been rising. The elasticity of scale in turn serves as an input for calculating industry markups. Increasing returns to scale help explain observed increases in markups over the last decades for broad sectors of the economy. My estimate of 1.2 for the aggregate markup is significantly lower than the estimate of 1.6 found in recent literature. The large disparity in markup estimates stems from differences in the treatment of fixed and variable costs and the methodological approach to the calculation of markups.

Keywords: market concentration, markup, elasticity of scale, increasing returns to scale, production function estimation
JEL Codes: D22, D24, L11, L16

1 Introduction

Theorizing about increasing returns to scale has a rich history among economists, from Smith (1776) who celebrated the division of labor, to Young (1928) and Romer (1986), who brought to light the scalability of knowledge, to Verdoorn (1949) and Kaldor (1978, 1981), who formulated the Kaldor-Verdoorn law, and to Krugman (1979, 1991), who applied the theory of increasing returns to international trade and

*I thank Garett Jones, Thomas Stratmann, Nathan Shanks, and the anonymous reviewers for their helpful comments and suggestions.

geography. Yet, when it comes to explaining widely acknowledged phenomena of the last decades: rising industry concentration (Autor et al., 2020; Grullon et al., 2019), decreasing share of labor in total output (Gutiérrez and Piton, 2020; Karabarbounis and Neiman, 2014), and declining business dynamism (Decker et al., 2014; Akcigit and Ates, 2021), economists revert back to the assumptions of constant or diminishing returns prevalent in the neoclassical economic theory.

In a seminal paper, Autor et al. (2020) provide a theoretical explanation for rising industry concentration while assuming production technology with constant returns to scale. Autor et al. (2020) show that under constant returns, an increase in consumer price sensitivity is necessary to generate a tendency for big firms to keep growing. The heterogeneity in total factor productivity between firms ensures that more productive firms are able to charge lower prices and capture a bigger market share. Increasing returns to scale, on the other hand, would generate similar outcomes of industry concentration and the rise of large firms without requiring a change in consumer preferences.

Increasing returns to scale can also generate the reduction of labor share in total output, if increasing returns are driven by capital due to advancements in automation and computer technologies, for example. Karabarbounis and Neiman (2014) explain the reduction in the share of labor in total output by an exogenous decline in the relative price of investment goods, which compels firms to substitute away from labor and toward capital. This theory requires the elasticity of substitution between capital and labor to be greater than one, which does not find much support in the literature (Lawrence, 2015; Oberfield and Raval, 2021).

Economists have also been concerned with rising industry markups. De Loecker et al. (2020) argue that markups are the cause of the rise in the number of large firms and the drop in labor share. This argument leaves open the question of what creates markups in the first place. The alternative explanation I advance in the present research reverses the markups-to-concentration causality. I argue that the technology of recent decades, such as automation and information technology, allows firms to experience increasing returns to scale, which in turn gives rise to large firms and industry concentration. Large firms incur lower marginal costs causing increased markups.

Despite arguments that increasing returns to scale may bring about industry concentration, market power, a decline in business dynamism and the labor share of output, the frequent reason increasing returns to scale are dismissed as an explanation for these phenomena is because empirical estimation attempts have found the

elasticity of scale to be close to one. Ackerberg et al. (2015), Antweiler and Trefler (2002), Pavcnik (2002) find it to be slightly above one, Fernandes (2007) and Demirer (2020) find it to be either slightly below or slightly above one depending on the industry or model specification, and few studies find it below one (De Roux et al., 2021). The absence of consensus among these estimates and the fact that the estimates are frequently based on select industries in a small group of countries give economists little reason to deviate from the assumption of constant returns to scale. But where should we draw the line between close enough to one and far enough from one to distinguish between constant returns and increasing returns to scale? Basu and Fernald (1997) discuss the plausibility of even a small deviation from one having a large impact on industry structure.

The main purpose of this paper is to contribute a new attempt at estimating the elasticity of scale. I estimate industry-level and aggregate elasticities of scale in the U.S. over the period from 1980 to 2019. I apply the estimation methods of Olley and Pakes (1996) and Ackerberg et al. (2015) to publicly traded companies and find that the aggregate elasticity of scale for the U.S. economy is above one and has been growing over the last four decades, from 1.02 to 1.10. My work differs from many previous attempts at estimating scale elasticities in that it purports to estimate them for the U.S. economy in the aggregate as opposed to a limited number of industries. De Loecker et al. (2020) and Traina (2018) are two recent papers that use Compustat data to estimate industry-level production functions for the U.S. economy, and my choice of data and approach come closest to theirs.

My contribution to the literature on the estimation of aggregate scale elasticities is twofold. First, I make my estimation approach explicit. Although De Loecker et al. (2020) use input-output elasticity estimates, explaining how the estimates are arrived at is not the objective pursued by De Loecker et al. (2020). Second, I estimate output elasticities for all 2-digit NAICS industries excluding only the finance sector. My econometric approach allows me to estimate the changes in the elasticity of scale over time even for industries with a relatively low number of firms. The data-intensive 2-stage estimation approaches of Olley and Pakes (1996) (henceforth "OP") and Ackerberg et al. (2015) (henceforth "ACF") render industry-year-specific estimation infeasible for smaller industries. To solve this problem, I aggregate the data into five-year rolling periods and generate industry-period-specific estimates.

The second objective of this paper is to derive markups using the estimated scale elasticities. De Loecker et al. (2020) argue that the aggregate markups of U.S. firms rose from 1.2 to 1.6 from 1980 to 2016. When the working version of this paper

appeared in print, it sparked responses by Syverson (2019), Basu (2019) and Berry et al. (2019) and separate research projects by Hall (2018) and Traina (2018) among others. Syverson (2019) and Basu (2019) see an inconsistency in De Loecker et al. (2020) calculations. Although markups and elasticity of scale are not directly observed, profit share of revenues is observable. With a simple algebraic relationship connecting these measures, high markup increases would have to be accompanied by extremely high profit margins, which are not supported by empirical evidence. Traina (2018) addresses this "paradox" by challenging the use of Cost of Goods Sold (COGS) as a measure of a firm's variable costs. Instead of using just COGS, Traina also includes Selling, General and Administrative expenses (SG&A) and arrives at markups in the range of 1.1 and 1.2. Hall (2018) uses time-series analysis with instrumental variables to estimate markups based on the Solow's 1957 method for finding the growth rate of technology. Hall's estimation shows growing markups but also markedly lower than in De Loecker et al. (2020).

Despite strong arguments disputing De Loecker et al.'s calculations, many articles (Autor et al., 2020; Blanchard, 2019; Gutiérrez and Philippon, 2017) continue to cite the authors' results as accepted baseline for further analysis and policy recommendations. This paper contributes to the literature disputing claims of high markups. I take Traina's adjustment one step further and include capital into the markup derivation. This approach is justified by two arguments. The first argument is the disappearing distinction between capital and labor when it comes to their fixed vs. variable status as well as the size of their adjustment costs. The second is the longer-term view I take by using the five-year horizon in the data that allows me to use the markup formula that includes total revenue and total cost, which Basu (2019) and Syverson (2019) use to question the validity of De Loecker et al.'s results. When I use total costs including COGS, SG&A and capital, I arrive at lower markup estimates.

The paper proceeds as follows. Section 2 talks about the data. In section 3, I explain the econometric procedures used for estimating the elasticities of scale. Section 4 shows the estimation results. Section 5 covers the theoretical model for markup estimation. In section 6, I show the markup derivation results. Finally, section 7 briefly discusses the macroeconomic implications of the findings and concludes.

2 Data

I use Compustat Fundamentals Annual database as my primary data source. The data set includes financial information on publicly traded companies in the U.S. from 1950 to 2020 (the data was downloaded on 9/24/2021). I exclude firms that do not report an industry code. I also exclude records where a firm's ratio of operating income before depreciation to sales is in the bottom or top 1% of all the firms in that year. I delineate industries using 2-digit NAICS industry codes. I exclude the year 2020 because of economic distortions generated by COVID-19. Although Compustat goes back to 1950, sufficiently detailed industry-level information appears to be available since around the 1980s. Up until the 1980s, certain industries are in nascent stages and are represented by only a few firms, which makes them unusable for 2-stage regressions with multiple controls required by the OP and ACF estimation methods, as will be discussed in detail in section 3. Figure 1 shows how many firms are in Compustat for each industry by year from 1960 to 2020. Some industries, such as arts, entertainment, and recreation (NAICS code 71) or agriculture, forestry, fishing and hunting (NAICS code 11) have a small number of firms in the earlier decades. For example, before 1970, there are fewer than 5 firms in the educational services industry (NAICS code 61) and fewer than 10 firms in warehousing (NAICS code 49). Estimating industry- and year-specific production functions is not feasible for such small sample sizes.

The accounting variables of interest are firm-level sales, COGS, SG&A, and property plant and equipment (PPE). It bears noting that unlike SG&A in firms' financial statements, SG&A in Compustat excludes depreciation. In 31% of observations, SG&A is missing. However, another variable available in Compustat, operating income before depreciation, can be used to impute SG&A by subtracting COGS and operating income before depreciation from sales. This calculation performed on the observations with all the data points present reveals that imputed SG&A and Compustat-provided SG&A are within 0.01% of each other in 98.7% of the cases. This test provides sufficient endorsement to the imputation of missing SG&A, and the imputation recovers half of the missing values.

To calculate user cost of capital, I use $r_t = i_t - \pi_t + \delta_t$, where i_t is the nominal interest rate, π_t is the inflation rate and δ_t is depreciation plus risk premium. I use the Federal Funds rate for i_t and the FRED reported inflation rate for π_t, and I follow De Loecker et al. (2020) in using constant 12% for δ_t.

The data analysis revealed that firms with 2-digit NAICS code 52 belonging to the finance and insurance industry differ significantly from the rest of the firms in

the data set. Specifically, their PPE is frequently not recorded in Compustat. This occurs even if the annual statements filed with the SEC contain information on PPE. PPE is missing in 21% of all observations, and 70% of those come from the finance and insurance industry. Within the finance and insurance industry, 74% of the observations are missing PPE. This absence of PPE in Compustat could lead to bias in the results because the observations with missing data simply get skipped by most statistical software, which would likely violate the assumption of random sampling. In addition, the financial industry revenue figures in Compustat include both interest and non-interest income, while their COGS include only interest expense.

Another factor that makes the finance industry unique is that the financial firms' operating margins appear to be much higher than in other industries as they record many of their expenses as extraordinary items. Since the industry represents over 15% of total revenues in Compustat, this premium in margins can significantly skew the aggregate results for the economy at large. Finally, the finance industry is different from the other industries in the type of exogenous factors that profoundly affect it. Besides the common-to-all factors of customer demand and the overall health of the economy, the finance industry responds uniquely to the Fed's monetary policy, which has been expansionary since the financial crisis of 2008 and through 2019. For all these reasons, I have excluded 2-digit NAICS code 52 from my analysis.

The summary statistics of the core variables are displayed in Table 1.

3 Elasticity of Scale

To estimate the elasticity of scale, I assume a Cobb-Douglas production function with two inputs, variable input V and capital input K: $Y = AK^{\beta_k}V^{\beta_v}$, where A is a Hicks-neutral technology, β_k is the output elasticity of the capital input and β_v is the output elasticity of the variable input. The elasticity of scale is then measured by the sum of the output elasticities $\beta_k + \beta_v$.

The Compustat data set contains data on revenues and costs, which are different from the outputs and inputs required to estimate production functions. Klette and Griliches (1996) found that if output prices are correlated with input choices, then using revenue and cost figures instead of quantities creates an omitted price bias, which biases the estimates for output elasticities downward. To partially address this bias, I follow De Loecker and Warzynski (2012) and Bartelsman et al. (2013) and deflate revenues and costs using industry-level deflators. I use chain-type price indexes for gross output by industry (2- or 3-digit level) published by the Bureau of Economic

Analysis. Bond et al. (2021, Section 3) argue that using industry-level deflators is not an adequate solution to the omitted price bias in the case of market power. However, a market equilibrium with a single market price and heterogeneous market power can be arrived at via a Cournot or a Dixit-Stiglitz models of competition that incorporate asymmetric cost structures (Frank Jr, 1965; Montagna, 1995; Okuguchi, 1973; Uchiyama, 2018). I therefore assume that the firms in an industry charge the same price, in which case deflating revenues and costs by industry-level deflators allows me to arrive at inputs and outputs necessary for estimating the production functions. Even if the assumption of a single market price does not hold perfectly, Klette and Griliches (1996) expect the omitted price bias to lead to the underestimation of the value of output elasticities, so the evidence of increasing returns shown in section 4 would be stronger if data on prices were available.

I use the sum of deflated COGS and SG&A as a measure of variable inputs and deflated PPE multiplied by the estimated user cost of capital as capital inputs. SG&A contain such expenses as marketing, office administration, human resources, and utilities, and there is debate about how to treat this input. One option is to treat SG&A as overhead independent of the production output (De Loecker et al., 2020). A second option is to treat it as a separate input fundamentally different from the variable and capital inputs. I choose a third option, which is to treat SG&A as a variable input as the expenses in SG&A do fluctuate with the output a firm produces. A firm needs more human resources if it has more employees, it will spend more on marketing if it expands output to enter new markets, and it will have higher utility bills to produce more output and have more employees.

There is another compelling reason to include SG&A in the variable input. It has been recorded by multiple authors (De Loecker et al., 2020; Traina, 2018) that the proportion of cost of goods sold in total costs of the firm has been falling and the proportion of selling, general and administrative costs has been rising. One explanation for this trend is the growing importance of marketing and management expenses, which are included in SG&A (Karabarbounis and Neiman, 2019). Traina (2018), conversely, hypothesizes that costs have been shifting from COGS into SG&A over time. To expand on Traina (2018)'s logic, I suggest considering the incentives for firms to re-code their existing costs under a different category. A perfunctory search of SEC's correspondence with firms filing their quarterly and annual reports reveals that firms sometimes justify a shift of costs from COGS into SG&A by making themselves comparable to competitors. Examples are Rite Aid correspondence with the SEC (2006) and McCormick's correspondence with the SEC (2007). Such

justification has little to do with the economic distinction between variable and fixed costs. Additionally, since gross margin is a metric of interest to many investors, firms may find it in their interest to make their gross margin look as high as possible by moving costs formerly categorized as COGS into SG&A (Fan and Liu, 2017). Thus, not including SG&A in variable costs will generate a bias and an inconsistency between time periods, since the classification of costs between the COGS and SG&A can change over time due to purely accounting reasons and not due to changes in production processes.

To run the OP and ACF processes on Compustat data I must solve the challenge of the low number of observations in some years for smaller industries. Compustat has data on individual firms by year, which enables accounting for both the heterogeneity between industries and the changes that production processes undergo over time. However, with some industries having a small number of publicly traded firms (e.g., 17 firms in the 2-digit NAICS industry 11: agriculture, forestry, fishing, and hunting in 2018), running empirical models by industry by year does not always produce meaningful results. I address this problem by looking instead at five-year rolling periods. Thus, I estimate elasticity of scale in period 1980 to 1984, then in period 1981 to 1985, and so on. This way, the number of observations I have for each statistical test is not the number of firms in any given year but roughly five times that number. Besides supplying more observations, this approach has good economic intuition as it allows me to examine a longer-term horizon in which a firm is expected to recoup much of its initial capital investment and reinvest in new capital purchases or upgrades. Using rolling periods has the added benefit of smoothing out the trend line making it easier to interpret.

Data within five-year rolling periods are panel data. To avoid biased results, I need to control for fixed effects. It is easy to control for year fixed effects. It is harder to do the same for firms, because including firm fixed effects would negate the intended data advantage of five-year rolling periods. I employ an imperfect solution to this problem by controlling for sub-industry, which I designate as a 3-digit NAICS industry. There are five sub-industries per industry on average, with the number of sub-industries ranging from one in utilities to ten in information services.

3.1 Ordinary Least Squares

Estimating the production function using the naive OLS suffers from simultaneity and selection biases (Marschak and Andrews, 1944). However, it may be instructive to apply OLS to the existing data to see how much and in what direction im-

provements to the estimation method can change the results. Using a simple Cobb-Douglas production function, I evaluate a separate regression for each five-year rolling period (36 periods) and for each industry (21 industries). The regressions are of the form

$$y_{it} = \beta_0 + \beta_k k_{it} + \beta_v v_{it} + year_t + sub_j + u_{it}, \tag{1}$$

where lower-case letters denote logs of Y_{it} for output, K_{it} for capital, and V_{it} for variable inputs, β_0 is a constant term, $year_t$ is year fixed effects, sub_j is sub-industry fixed effects, and u_{it} is the error term. The coefficients on variable and capital inputs, β_v and β_k, are specific to the industry and the five-year rolling period. I sum them to arrive at the elasticity of scale per period per industry.

3.2 Syverson's Method

In his 2004 paper, Syverson uses a modified version of the OLS regression. He assumes a Cobb-Douglas production function of the form

$$Y = A(K^{\beta_k} V^{1-\beta_k})^{\gamma}, \tag{2}$$

where γ is the elasticity of scale. This formulation assumes that for a cost-minimizing firm with a Cobb-Douglas production function, output elasticities of separate inputs are proportional to the shares of their respective costs in total costs. Since the exponents inside the parentheses in equation (2) add up to one, β_k can be calculated for every observation in the data set as $\frac{K_{it}}{K_{it} + V_{it}}$. Converting the production function into logs creates a linear function that can be estimated via linear regression

$$y_{it} = \beta_0 + \gamma input_{it} + year_t + sub_j + u_{it}, \tag{3}$$

where $input_{it} = \ln(K_{it}^{\beta_k} V_{it}^{1-\beta_k})$.

A distinct advantage of this method of estimating returns to scale over OLS is that it allows for heterogeneity in the exponents on inputs in the production function of individual firms. Allowing heterogeneity in input-output elasticities recognizes that even within the same industry firms may use different technologies and combinations of inputs to produce equivalent outputs. However, allowing heterogeneity comes at a price of the restrictive condition that output elasticities of capital and variable inputs are in direct proportion to their shares in total costs, which is a strong assumption.

Another advantage of Syverson's formulation is that standard errors are easy to

estimate for the elasticity of scale. In section 3.4, I discuss how I estimate standard errors in the approach where elasticities of variable and capital inputs are estimated separately.

3.3 Olley-Pakes Method

The two biases in the OLS estimation of a firm's production function are simultaneity and selection (Marschak and Andrews, 1944). The simultaneity bias has to do with the fact that firms choose the capital and variable inputs based on the information available to them at the time. If the firm faces a positive productivity shock, it selects to invest in more inputs. The naive OLS estimation of the production function does not account for the unobserved (to the econometrician) productivity shock. The productivity shock is therefore left in the error term biasing the coefficients on inputs. The selection bias results from neglecting to account for the fact that firms may respond to a negative productivity shock by exiting the market altogether.

Olley and Pakes (1996) devise a method to resolve these biases. Assuming Cobb-Douglas technology and taking logs results in equation (1), where u_{it} is the error term that contains the productivity shock ω_{it}. The OP methodology assumes that variable inputs are not dynamic and get chosen at time t. Unlike variable inputs, capital is a dynamic input and gets chosen at time $t-1$. When a positive productivity shock is perceived by the firm, the firm reacts to that shock by investing in more capital. Thus, capital investment is a function of the productivity shock and can be assumed to be positive and strictly increasing: $inv_t(\omega_{it}, k_{it})$. Inverting the investment function generates a function for the productivity shock

$$\omega_{it} = h_t(inv_{it}, k_{it}). \tag{4}$$

Subscript t in h_t allows for functional heterogeneity in different periods but assumes the same functional form for productivity for all firms within an industry. Pulling the productivity shock out of the error term of the production function (1) and substituting (4) for the productivity shock results in

$$y_{it} = \beta_0 + \beta_v v_{it} + \beta_k k_{it} + h_t(inv_{it}, k_{it}) + year_t + sub_j + e_{it}, \tag{5}$$

where e_{it} is an unbiased error term. We can combine terms with capital and investment into a composite term $\phi_{it} = \beta_k k_{it} + h_t(inv_{it}, k_{it})$ resulting in

$$y_{it} = \beta_0 + \beta_v v_{it} + \phi_{it} + year_t + sub_j + e_{it}. \tag{6}$$

Equation (6) can be estimated using OLS and approximating ϕ_{it} with a second-order polynomial series in capital and investment. This step provides an estimate of output elasticity of variable input, β_v. The coefficient β_v is unbiased because the error term no longer contains the productivity shock and thus is no longer correlated with the explanatory variables.

The next assumption in OP is that productivity follows a first-order Markov process. Productivity also depends on the probability of exit. These premises result in the following expression for the evolution of productivity shocks:

$$\omega_{it} = g_t(\omega_{it-1}, P_{it}) + \varepsilon_{it}, \tag{7}$$

where $g(\cdot)$ is an unknown function of the productivity shock in the previous period and the probability of exit in this period, and ε_{it} is an innovation term. Exit in turn is determined by capital, investment, and the productivity shock in the previous period: $P_{it}(inv_{it-1}, k_{it-1}, \omega_{it-1})$. From (4), it follows that $\omega_{it-1} = h_{t-1}(inv_{it-1}, k_{it-1})$. Thus, the probability of exit can be expressed as an unknown function $p(\cdot)$ of investment and capital in the previous period,

$$P_{it} = p_t(inv_{it-1}, k_{it-1}), \tag{8}$$

which I estimate via a probit regression using a second-order polynomial of investment and capital in the previous period.

Now I have all the necessary components to find the coefficient on capital β_k. I start with equation (6) and use the predicted impact of the variable input to isolate the impact of capital and the productivity shock, so the dependent variable is now $y_{it} - \hat{\beta}_v v_{it}$. From the definition of ϕ_{it}, it follows that $\phi_{it-1} = \beta_k k_{it-1} + h_{t-1}(inv_{it-1}, k_{it-1})$. Results from the same first-stage regression (6) provide predicted values of $\hat{\phi}_{it-1}$, while results from the probit regression based on equation (8) provide predicted values of $\hat{P}_{it}$. The coefficient on capital, β_k, can be recovered by fitting the following equation by nonlinear least squares:

$$y_{it} - \hat{\beta}_v v_{it} = \beta_0 + \beta_k k_{it} + g_t(\hat{\phi}_{it-1} - \beta_k k_{it-1}, \hat{P}_{it}) + year_t + sub_j + \varepsilon_{it} + e_{it}, \tag{9}$$

where I estimate the unknown function $g_t(\hat{\phi}_{it-1} - \beta_k k_{it-1}, \hat{P}_{it})$ using a second-order polynomial of $\hat{\phi}_{it-1} - \beta_k k_{it-1}$ and $\hat{P}_{it}$.

3.4 Ackerberg-Caves-Frazer Method

Ackerberg et al. (2015) rectify a potential weakness in the Olley and Pakes (1996) methodology, namely the assumption that the variable input is non-dynamic. If that assumption is violated, then the OP estimation has a functional dependence problem and the estimator of β_v is not consistent. Relaxing the assumption of the non-dynamic nature of variable inputs makes it impossible to estimate β_v in the first stage because now productivity is a function of both capital and variable inputs,

$$\omega_{it} = h_t(inv_{it}, v_{it}, k_{it}), \tag{10}$$

which results in the production function

$$y_{it} = \beta_0 + \beta_v v_{it} + \beta_k k_{it} + h_t(inv_{it}, v_{it}, k_{it}) + year_t + sub_j + e_{it}. \tag{11}$$

Now ϕ_{it} is defined as $\phi_{it} = \beta_v v_{it} + \beta_k k_{it} + h_t(inv_{it}, v_{it}, k_{it})$, rendering the following expression for the production function:

$$y_{it} = \beta_0 + \phi_{it} + year_t + sub_j + e_{it}, \tag{12}$$

which I estimate using a second-order polynomial of the variable input, capital, and investment. The predicted values $\hat{\phi}_{it-1}$ are used in the second stage in the expression of productivity

$$\omega_{it-1} = h_{t-1}(inv_{it-1}, v_{it-1}, k_{it-1}) = \hat{\phi}_{it-1} - \beta_v v_{it-1} - \beta_k k_{it-1}. \tag{13}$$

Since the ACF methodology does not estimate β_v in the first stage, it uses the generalized method of moments to estimate all production function parameters in the second stage. Thus, I arrive at the following second stage conditional moment:

$$\begin{aligned} E\Big[\varepsilon_{it} + e_{it} | \omega_{it-1}\Big] = E\Big[y_{it} - \beta_0 - \beta_v v_{it} - \beta_k k_{it} \\ - g_t\left(\hat{\phi}_{it-1} - \beta_v v_{it-1} - \beta_k k_{it-1}, \hat{P}_{it}\right) - year_t - sub_j | \omega_{it-1}\Big] = 0. \end{aligned} \tag{14}$$

Transforming this conditional moment into unconditional moments for estimation and assuming that variable inputs are chosen at time $t-1$, I choose the set of the following second-stage moment conditions to estimate the parameters of the

production function, $\beta_0, \beta_v, \beta_k$, and g:

$$E\left[\begin{array}{l} y_{it} - \beta_0 - \beta_v v_{it} - \beta_k k_{it} \\ \\ \quad - g_t(\hat{\phi}_{it-1} - \beta_v v_{it-1} - \beta_k k_{it-1}, \hat{P}_{it}) - year_t - sub_j \end{array} \otimes \begin{pmatrix} v_{it} \\ k_{it} \\ \hat{P}_{it} \\ \hat{\phi}_{it-1} \\ year_t \\ sub_j \end{pmatrix}\right] = 0. \quad (15)$$

One of the challenges with estimating the elasticity of scale by adding up the variable and capital input elasticities is the estimation of standard errors to evaluate the statistical significance of the findings. The ACF method allows for a simple solution to this challenge because it estimates the coefficients on the variable and the capital inputs simultaneously. I use bootstrap standard errors for the GMM procedure. In addition to the standard errors, I retrieve the covariances between the estimates of β_k and β_v, which in turn allows me to estimate the standard errors of the elasticity of scale following the conventional approach for calculating the standard error of a sum of two estimates.

4 Estimation Results

The estimates I obtain from the methods described in section 3 are by industry and by period. To aggregate them into economy-wide estimates, I use industry shares of total period sales as weights.

Figure 2 shows returns to scale over time using the four estimation approaches. The labels on the horizontal axes mark the first year of the five-year rolling period, so a data point marked 1990 is the elasticity of scale estimated for the period 1990 to 1994. Estimates using the OLS and Syverson's methods contain the most bias. Syverson's method forces the output elasticities of capital and variable costs to be proportional to their cost shares. OLS trendline matches Syverson's method trendline almost perfectly until around 1995 when the two lines diverge. This is an inter-

esting observation in itself, marking a potential shift in the role of capital vs. variable inputs, such as labor, in production, which coincides with the period of information technology innovations of the late 1990s and early 2000s.

The Olley-Pakes and Ackerberg-Caves-Frazer methods reduce selection and simultaneity biases. In terms of the trendline, however, the OP method produces estimates higher than those generated using OLS. Between the OP and ACF methods, ACF is flatter since the early 2000s and appears to be more affected by the 2008 recession. Henceforth, I will focus on the ACF estimates for three reasons. First, it allows for the calculation of standard errors, which is an important advantage in evaluating the significance of the empirical results. Second, the ACF method estimates the lowest elasticity of scale of the four methods used, which means I will be using the most conservative estimates. Third, the ACF method estimates production functions under the assumption that capital and non-capital inputs are similar in their dynamic nature, an assumption I discuss in detail in section 5.

The biggest takeaway from Figure 2 is that the elasticity of scale is above 1 and has been rising since the 1980s. It experienced a steep increase in the 1990s, which coincides with the beginning of the Internet revolution. The latest sharp increase happened as the country was recovering from the 2008 financial crisis. It can be explained by the fact that the crisis is likely to have disproportionately affected businesses with outmoded technologies and older processes, creating a type of selection mechanism for firms that could more effectively utilize the economies of scale. Using the Ackerberg-Caves-Frazer method results in the estimate of the aggregate elasticity of scale for the U.S. economy in the five-year period 2015-2019 of 1.1.

This estimate is higher than that found by De Loecker et al. (2020), who although focusing on the output elasticity of the variable input, do mention that the average elasticity of scale according to their estimates has increased from 1.03 to 1.08. However, the authors do not delve into how they arrive at these estimates, nor do they draw industry or macroeconomic implications from them.

Figure 3 shows the 95% confidence interval for the estimate of the elasticity of scale using the ACF method. The graph shows that the estimate is well outside of the null hypothesis of being equal to one.

It is instructive next to look at industry-specific elasticities to see which industries drive the aggregate number. Figure 4 shows the elasticity of scale for the 21 U.S. 2-digit NAICS industries (excluding finance). Panel 4a shows the top seven industries by total revenues. Manufacturing of wood and chemicals (NAICS code 32) stands out the most with the highest returns to scale, reaching as high as 1.18 at its

peak. The information industry (NAICS code 51) displays a pattern shown earlier with an expected steep increase in scale elasticity in the 1990s. Exploring the underlying technological and organizational reasons for the upward trends in the scale elasticity graphs by industry is outside the scope of this study. Yet, the graph suggests that although a few industries such as retail trade of durables (NAICS code 44) or wholesale trade (NAICS code 42) are hovering close to one, others show an upward trajectory distinct from one.

The top seven industries by revenues also have a relatively large number of firms compared to the smaller industries in panels 4b and 4c, so their trend lines are much less volatile. In panel 4b, for example, warehousing (NAICS code 49) is characterized by a small number of firms with a market share distribution highly skewed to the left, dominated by three large firms: UPS, USPS, and FedEx. Any changes within any of the top firms will have an effect on the elasticities estimated for the industry. For example, FedEx restructuring in 2000 accounts for much of the volatility in the industry trend during that period.

Industries in panel 4c show the highest volatility, and such volatility precludes me from drawing robust conclusions about the elasticity of scale for these industries. Yet, since the aggregate elasticity is calculated using the share of industry revenue in the economy as weights, the bottom industries have little effect on the aggregate estimate as the share of the revenue for the 7 bottom industries in the data in 2019 was 3%.

5 Markup Calculation

In the following sections I will use the elasticity of scale I estimated previously to calculate markups for the U.S. industries and for the U.S. economy as a whole over the last four decades. In a paper that received a lot of attention due to its claim of high and increasing markups in the U.S., De Loecker et al. (2020) use the following formula for the markup:

$$\mu = e_V \frac{PQ}{P^V V}, \tag{16}$$

where Q is output, P is the price of output, V is the variable input, P_V is the price of the variable input, and e_V is the output elasticity of the variable input. Expression (16) is derived from the cost minimization Lagrangian of the form

$$L(V,K,\lambda) = P^V V + rK - \lambda(Q(\Omega,V,K) - \overline{Q}), \tag{17}$$

where K is capital, r is the user cost of capital, $Q(\cdot)$ is the production function, Ω is productivity, and $\overline{Q}$ is the target output. In this formulation, λ is effectively the marginal cost. The first-order condition with respect to V results in

$$\frac{1}{\lambda} = \frac{1}{P^V}\frac{\partial Q(\cdot)}{\partial V}, \tag{18}$$

which simplifies to equation (16) once both sides are multiplied by P and the right side is multiplied by $\frac{V}{Q} \times \frac{Q}{V}$.

Economics textbooks, such as Varian (1992), Krugman and Wells (2018), Perloff (2022), feature a different formula for markup,

$$\mu = e_{scale}\frac{PQ}{TC}, \tag{19}$$

where e_{scale} is the elasticity of scale and TC stands for "total costs." Varian (1992) shows that the elasticity of scale is the percent change in output to the percent change in all inputs and is equal to $\frac{AC}{MC}$ (average cost divided by marginal cost). Syverson (2019) derives the same relationship between the markup and the elasticity of scale by simple algebraic steps

$$\mu = \frac{P}{MC} = \frac{P}{MC}\frac{AC}{AC}\frac{Q}{Q} = \frac{AC}{MC}\frac{PQ}{AC \times Q} = e_{scale}\frac{PQ}{TC}. \tag{20}$$

The difference between the two formulas for the markup lies in the short-term vs. long-term view of the firm. In the short term, the firm has fixed and variable inputs, and it can only change the variable inputs. Hence, in the cost-minimization problem in De Loecker et al. (2020), the first-order condition is taken with respect to variable and not capital inputs, because capital inputs are considered fixed. However, in the long term, the firm can vary all inputs, and this assumption would allow taking a first-order condition with respect to capital, which results in

$$\frac{1}{\lambda} = \frac{1}{r}\frac{\partial Q(\cdot)}{\partial K}. \tag{21}$$

Together with equation (18), the long-term cost minimization requires that an additional dollar of the variable cost cause the same change in output as an additional dollar of capital cost. This means that there is an optimal proportion of expenditures on variable and capital inputs for a cost-minimizing firm. It also follows that the markup in the long-term view can be calculated using both variable and capital inputs.

$$\mu = e_V\frac{PQ}{P^V V} = e_K\frac{PQ}{rK}. \tag{22}$$

Substituting $e_K = e_{scale} - e_V$ and $rK = TC - P^V V$ and performing simple algebraic manipulations, we arrive at the textbook formulation of markup shown in equation (20):

$$\begin{aligned} e_V \frac{PQ}{P^V V} &= (e_{scale} - e_V) \frac{PQ}{TC - P^V V} \\ e_V PQ(TC - P^V V) &= e_{scale} PQP^V V - e_V PQP^V V \\ e_V PQ\ TC &= e_{scale} PQP^V V \\ e_V \frac{PQ}{P^V V} &= e_{scale} \frac{PQ}{TC} = \mu. \end{aligned} \tag{23}$$

I see three lines of argumentation along which to debate the use of the short-term or the long-term formula for markups: time horizon, adjustment costs for capital inputs, and classification of fixed and variable inputs. The argument in favor of the short-term approach would highlight the fact that firms in a competitive market must be nimble and respond to market conditions relatively fast. The counter-argument would stress the fact that firms plan ahead, develop long-term strategies, and make investments for the future. Both points have merit, but I advocate for the long-term approach because the issues of industry concentration and market power focus on large firms, which have the luxury of long-term strategic planning, as opposed to small businesses, which are at least anecdotally more driven by short-term goals of survival. A large firm has the capital to make long-term investments and the operating funds to lower prices to drive out competition. At any given point, markups of a large firm may be low as part of a strategic decision to win market share, with a longer-term result of more market power and higher prices and markups. Therefore, it appears more insightful to look at long-term markups.

The second line of argumentation deals with adjustment costs of capital inputs. The reason the variable inputs alone are used for the markup calculations in economics papers such as De Loecker et al. (2020) is because it is assumed that adjustment costs for capital are positive while adjustment costs for variable costs, specifically labor, are zero. Positive adjustment costs would mean that the markup needs to be higher to cover them. If adjustment costs are not accounted for in the derivation of the markup, the markup estimate will be biased upward. Since data on adjustment costs are difficult to come by, using the markup calculation with just the variable input as opposed to the markup calculation with just the capital input seems like a straightforward choice. However, the concern over adjustment costs vanishes altogether when I employ the long-term markup formula, because it uses total costs, and total costs subsume adjustment costs.

The third line of argumentation concerns the classification of fixed and variable costs using the firms' financial statements. The lines between variable and fixed costs become increasingly blurry. Variable costs are supposed to change with the output quantity and be characterized by costless and instantaneous adjustment. However, finding and onboarding new employees is far from costless. Firms are not required to disclose their onboarding costs, but when third-party recruiters are involved, hiring firms frequently pay such recruiters as much as 20% of a year's salary for a new hire. Additionally, the amount of human capital investments firms make into their employees has been growing (Cascio, 2019). In certain sectors, such as education for example, where the employer has limited ability to fire personnel, labor acquires characteristics of a fixed cost. Eliminating this cost requires a concerted effort of a layoff not much different in economic terms from the effort to decommission unused capital.

Fixed costs are supposed to be stable within a range of output quantities and be characterized by positive adjustment costs. Traditionally, IT equipment such as servers and storage arrays have been considered fixed costs. Today, firms can outsource such costs to providers such as Amazon or Microsoft and increase or decrease their infrastructure on demand, bringing IT capacity expenditures much closer to variable costs. When a firm does own its server infrastructure, much of it gets fully replaced on a three- to five-year schedule – a significant shift from the theoretical concept of long-lasting capital investments with multi-decade useful lives.

The three relevant cost categories in financial statements are COGS, SG&A, and capital. I discussed in section 3 why not including SG&A in the variable costs creates bias and inconsistency over time. A similar argument may be applied to capital. Firms frequently change their strategies with regard to capital. When they own office buildings, they have depreciation expense and potentially interest expense. When they rent, they have rent expense, which is recorded in SG&A. By including SG&A and excluding capital-related expenses, one would introduce bias into the calculation of markup. If the output elasticity of the chosen input can reflect the differences between firms that own real estate vs. rent it, then no issue arises, but since elasticities are calculated by industry, artificial differences in markups will arise. With the same elasticity within the industry, the markup calculation excluding capital-related expenses will automatically estimate a higher markup for firms that own real estate than firms that lease real estate, because leasing firms will have a higher SG&A.

The longer-term planning horizon of large firms, the blurry lines between variable and fixed costs, and the bypassing of the adjustment costs calculation challenge

lead me to choose the long-term markup calculation ($\mu = e_{scale}\frac{PQ}{TC}$) over the short-term markup calculation ($\mu = e_V \frac{PQ}{P^V V}$).

Bond et al. (2021, Section 2) show that when revenue is used instead of output quantity, the markup calculation using revenue elasticities does not carry any information about the true markup. I partially shield my results from this criticism by deflating the revenues by 2- and 3-digit industry deflators. Thus, I implicitly treat the elasticity estimates I derived as output elasticities in the markup calculations. Although my markup estimates are potentially subject to some degree of the omitted price bias, the main argument with respect to markups is that my estimates are substantially lower than those derived by De Loecker et al. (2020), subject to the same bias, and propagated in much of recent literature (De Ridder, 2019; Döpper et al., 2022; Liu et al., 2022). My goal here is to focus on a key methodological difference between our approaches, unrelated to the omitted price bias.

6 Markup Results

I estimate the elasticity of scale by period by industry, while the sales-to-cost ratio ($\frac{PQ}{TC}$) is available by year by firm. The sales-to-cost ratio is easy to calculate because most of the necessary values are directly observable in the data, with the exception of the user cost of capital, which has been estimated. Thus, I can calculate the markup for each individual firm for each year. To arrive at the aggregate markup for the economy while maintaining consistency with the five-year rolling periods used throughout the paper, I multiply the individual firm-year-specific markups by the share of the firm-year sales in the total sales across all firms in the five-year rolling period. Figure 5 shows the estimate of the aggregate U.S. markup over time together with the two aggregate components that make up the markup: the elasticity of scale (ACF method) and the sales-to-cost ratio. The markup has grown from 1.03 in the 1980s to 1.17 in the last five-year period of 2015-2019, peaking at 1.20. These figures stand in stark contrast to those arrived at by De Loecker et al. (2020), and the trend line has been going down since 2010.

The graph shows that both components of the markup equation contribute to the upward trajectory of markups over time. The sales-to-cost ratio appears to have responded to the Internet revolution with a characteristic jump, but it appears to lag the elasticity of scale, which is logical given price-stickiness in the economy. Overall, the sales-to-cost ratio appears to be a more volatile component of the markup.

Figure 6 shows the evolution of markups for individual industries, again broken

into three groups of seven industries sorted by revenues. The information sector (NAICS code 51) in Panel 6a is the leader in markups among the top seven industries. A rich literature on the rising market power and concentration of the information industry includes papers such as Decker et al. (2014) and Shapiro (2019) among others.

The markup of the utilities industry (NAICS code 22) stands out as it is below one much of the time, but it is not surprising given the regulated character of the utilities industry. On the one hand, regulation pushes the prices down, but on the other hand, the government can reduce the cost of capital for utility companies by providing grants or low-interest loans. Since I applied the same user cost of capital to all industries to avoid making additional assumptions, the markup graph for the utilities industry is likely biased downward.

The mining industry (NAICS code 21) in Panel 6b includes oil and gas extraction so price volatility of oil and gas on the world market can drive the volatility of markups for this industry. Thus, the inverse U-shape of the graph in the period from the mid-1990s to the recent decade reflects the price behavior of oil and gas during the same time period. The markups for educational services (NAICS code 61) in Panel 6c likely reflect the increased demand for distant-learning services, while the trend line of the construction industry's markup (NAICS code 23) reflects the boom and bust in the housing market preceding and following the 2008 financial crisis.

The industries with the lowest markups (apart from utilities discussed above) are transportation and warehousing, retail and wholesale trade, and accommodation and food industries.

Finally, it is interesting to test the notion that markups calculated inclusive of the capital input would be biased upward compared to markups calculated without the capital input due to the adjustment costs of capital. Figure 7 shows the aggregate markups calculated using the two approaches. It may be surprising to see the markup using total costs to be lower than the markup using only COGS and SG&A. However, as discussed earlier, both capital and other inputs require adjustment costs. Not including capital costs in the output elasticity and the sales-to-cost ratio creates other sources of bias that outweigh the difference in adjustment costs between capital and the other inputs. If one prefers the calculation without the capital costs, one also must admit that the change in markups between the 1980s and today has not been dramatic (from 1.13 to 1.23).

7 Conclusion

There is a growing consensus among economists that the U.S. economy has been exhibiting increasing industry concentration across multiple industries. This trend is frequently seen as having a negative effect on the economy in terms of competitiveness, productivity, innovation, and labor share of income. As concentration increases, markups are expected to grow, which drives the labor share down. This paper evaluates recent calculations of markups and proposes a new explanation for industry concentration, namely increasing returns to scale. I estimate the elasticity of scale for different U.S. industries over the period 1980-2019 using data on publicly traded companies. Four estimation methods are employed: simple OLS, Syverson's method (Syverson, 2004), Olley and Pakes method (Olley and Pakes, 1996), and Ackerberg, Caves, and Frazer method (Ackerberg et al., 2015). I find that the aggregate elasticity of scale is above one and has been rising (from 1.02 to 1.10 in the period from 1980 to 2019). Increasing returns to scale in turn can explain, at least in part, the rising industry concentration and increases in markups for broad sectors of the economy.

Increasing returns to scale provide an explanation for a finding reported by De Loecker et al. (2020), whereby higher markups have shifted to the bigger firms over the last few decades. The authors claim that this shift is due to higher market power. They likely interpret it that way because the output elasticity of variable input they estimate and use in their markup calculation is either flat or going down. But the conclusion is quite different when we account for the increasing returns to scale instead. The markup is calculated as elasticity multiplied by the sales-to-cost ratio, but the elasticity both in this paper and in De Loecker et al. (2020) is assumed to be the same for all the firms in the industry for a given period. Thus, comparing markups between firms within an industry amounts to comparing their sales-to-cost ratios. If we allow for the increasing returns to scale, then the sales-to-cost ratio would automatically be higher for the bigger firms, because their average costs are lower due to their size. The additional reward reaped by bigger firms in the form of higher markups is the consequence and not the cause of the market structure.

My analysis shows that the aggregate markup has been around 1.2 over the last decade, which is in stark contrast to estimates of 1.6 generated by recent literature (De Loecker et al., 2020). The reason for the disparity in markup estimates stems from differences in methodological approach to markup estimation and the classification of accounting data into economic categories of fixed and variable costs. The present research uses total costs as opposed to COGS or COGS+SG&A alone in cal-

culating scale elasticities, which in turn are inputs in the calculation of markups. The U.S. Generally Accepted Accounting Principles require publicly traded companies to report their financial results in a consistent and transparent manner. However, for the reporting principles to be adhered to by firms from vastly different industries with vastly different revenue and cost structures, they must be general enough that much variation is inevitably present. This variation is most pronounced between firms from different industries but likely exists between firms within the same industry as well. The cross-sectional variation in reporting is exacerbated by time-varying changes in the Principles themselves as well as how they are applied. To focus on one category of costs and ignore the others in running aggregate estimation procedures is therefore prone to errors even if the procedure itself is valid. The current research attempts to safeguard against the errors resulting from the data reporting by focusing on total costs, a category far less susceptible to reporting inconsistency than any one sub-category within the total costs.

Increasing returns to scale have far-reaching implications for policymaking. The implications of Autor et al. (2020) model, whereby the randomly drawn total factor productivity z_i is the cause for the firm's cost advantage and subsequent growth, are different from the implications of the increasing returns to scale in important respects. Although both theories suggest that large firms tend to be more productive than small firms, the theory where productivity advantage stems from the randomly drawn TFP presupposes that if the same TFP was available to all firms, more and stronger competition could be achieved. Conversely, the theory where productivity advantage comes from size, breaking up a large firm into smaller competing firms would have the effect of destroying productivity. Understanding which effect dominates, the total factor productivity or returns to scale, is crucial for policy implications of growing industry concentration.

Production technology with increasing returns is by itself a wealth-increasing phenomenon: as inputs grow, the output grows at a faster rate. An individual firm faced with scale elasticity above one has the incentive to get larger, and the limits to the resulting tendency toward concentration are set by the consumers' preference for variety (Dixit and Stiglitz, 1977). The firm will still face competition from producers of substitute products or from firms that happen to experience a higher productivity shock in a given year. Antitrust policies guided by the idea that breaking up large firms should increase and strengthen competition and therefore be efficiency enhancing must face the reality that large size may be an important source of productivity for a firm. Large firms do not get larger only due to the luck of the draw of

the total factor productivity, as is often argued, but also because their productivity grows with size.

With regard to international trade, the existence of increasing returns challenges some of the standard arguments for market efficiency and suggests new defenses for old policies (Hudson, 2010). The infant industry argument for protectionism, for example, may have greater force in a world of increasing returns (Palley, 2008; Giammetti et al., 2022). However, as Krugman (1987, 1993, 1997) successfully argued, it is unrealistic to expect that in today's globally interdependent world, national governments can engage in strategic competition without creating unforeseen negative consequences.

References

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

Akcigit, U. and Ates, S. T. (2021). Ten facts on declining business dynamism and lessons from endogenous growth theory. *American Economic Journal: Macroeconomics*, 13(1):257–298.

Antweiler, W. and Trefler, D. (2002). Increasing returns and all that: A view from trade. *American Economic Review*, 92(1):93–119.

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

Bartelsman, E., Haltiwanger, J., and Scarpetta, S. (2013). Cross-country differences in productivity: The role of allocation and selection. *American Economic Review*, 103(1):305–334.

Basu, S. (2019). Are price-cost markups rising in the united states? A discussion of the evidence. *Journal of Economic Perspectives*, 33(3):3–22.

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

Berry, S., Gaynor, M., and Scott Morton, F. (2019). Do increasing markups matter? Lessons from empirical industrial organization. *Journal of Economic Perspectives*, 33(3):44–68.

Blanchard, O. (2019). Public debt and low interest rates. *American Economic Review*, 109(4):1197–1229.

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

Cascio, W. F. (2019). Training trends: Macro, micro, and policy issues. *Human Resource Management Review*, 29(2):284–297.

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

De Loecker, J. and Warzynski, F. (2012). Markups and firm-level export status. *American Economic Review*, 102(6):2437–2471.

De Ridder, M. (2019). Market power and innovation in the intangible economy. Working Paper, Cambridge Working Papers in Economics.

De Roux, N., Eslava, M., Franco, S., and Verhoogen, E. (2021). Estimating production functions in differentiated-product industries with quantity information and external instruments. Working Paper No. 28323, National Bureau of Economic Research.

Decker, R., Haltiwanger, J., Jarmin, R., and Miranda, J. (2014). The role of entrepreneurship in US job creation and economic dynamism. *Journal of Economic Perspectives*, 28(3):3–24.

Demirer, M. (2020). Production function estimation with factor-augmenting technology: An application to markups. Working Paper, Massachusetts Institute of Technology.

Dixit, A. K. and Stiglitz, J. E. (1977). Monopolistic competition and optimum product diversity. *American Economic Review*, 67(3):297–308.

Döpper, H., MacKay, A., Miller, N., and Stiebale, J. (2022). Rising markups and the role of consumer preferences. Working Paper No. 22-025, Harvard Business School Strategy Unit.

Fan, Y. and Liu, X. (2017). Misclassifying core expenses as special items: Cost of goods sold or selling, general, and administrative expenses? *Contemporary Accounting Research*, 34(1):400–426.

Fernandes, A. M. (2007). Trade policy, trade volumes and plant-level productivity in Colombian manufacturing industries. *Journal of International Economics*, 71(1):52–71.

Frank Jr, C. R. (1965). Entry in a Cournot market. *The Review of Economic Studies*, 32(3):245–250.

Giammetti, R., Papi, L., Teobaldelli, D., and Ticchi, D. (2022). The network effect of deglobalisation on European regions. *Cambridge Journal of Regions, Economy and Society*, 15(2):207–235.

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

Gutiérrez, G. and Philippon, T. (2017). Declining competition and investment in the US. Working Paper No. 23583, National Bureau of Economic Research.

Gutiérrez, G. and Piton, S. (2020). Revisiting the global decline of the (non-housing) labor share. *American Economic Review: Insights*, 2(3):321–338.

Hall, R. E. (2018). New evidence on the markup of prices over marginal costs and the role of mega-firms in the US economy. Working Paper No. 24574, National Bureau of Economic Research.

Hudson, M. (2010). *America's Protectionist Takeoff: 1815–1914. The Neglected American School of Political Economy*. Islet, Dresden, Germany.

Kaldor, N. (1978). *Further Essays on Economic Theory*. Duckworth, London, United Kingdom.

Kaldor, N. (1981). The role of increasing returns, technical progress and cumulative causation in the theory of international trade and economic growth. *Economie Appliquée: Archives de l'ISMEA*, 34(4):593–617.

Karabarbounis, L. and Neiman, B. (2014). The global decline of the labor share. *The Quarterly Journal of Economics*, 129(1):61–103.

Karabarbounis, L. and Neiman, B. (2019). Accounting for factorless income. *NBER Macroeconomics Annual*, 33(1):167–228.

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

Krugman, P. (1991). Increasing returns and economic geography. *Journal of Political Economy*, 99(3):483–499.

Krugman, P. (1997). *Pop Internationalism*. MIT Press Books, Cambridge, MA.

Krugman, P. and Wells, R. (2018). *Microeconomics*. Worth Publishers, New York, 3rd edition.

Krugman, P. R. (1979). Increasing returns, monopolistic competition, and international trade. *Journal of International Economics*, 9(4):469–479.

Krugman, P. R. (1987). Is free trade passé? *Journal of Economic Perspectives*, 1(2):131–144.

Krugman, P. R. (1993). The narrow and broad arguments for free trade. *American Economic Review*, 83(2):362–366.

Lawrence, R. Z. (2015). Recent declines in labor's share in US income: A preliminary neoclassical account. Working Paper No. 21296, National Bureau of Economic Research.

Liu, E., Mian, A., and Sufi, A. (2022). Low interest rates, market power, and productivity growth. *Econometrica*, 90(1):193–221.

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

Montagna, C. (1995). Monopolistic competition with firm-specific costs. *Oxford Economic Papers*, 47(2):318–328.

Oberfield, E. and Raval, D. (2021). Micro data and macro technology. *Econometrica*, 89(2):703–732.

Okuguchi, K. (1973). Quasi-competitiveness and Cournot oligopoly. *The Review of Economic Studies*, 40(1):145–148.

Olley, S. and Pakes, A. (1996). The dynamics of productivity in the telecommunications industry. *Econometrica*, 64:1263–1298.

Palley, T. I. (2008). Institutionalism and new trade theory: rethinking comparative advantage and trade policy. *Journal of Economic Issues*, 42(1):195–208.

Pavcnik, N. (2002). Trade liberalization, exit, and productivity improvements: Evidence from Chilean plants. *The Review of economic studies*, 69(1):245–276.

Perloff, J. M. (2022). *Microeconomics*. Pearson Education, Boston, 7th. edition.

Romer, P. M. (1986). Increasing returns and long-run growth. *Journal of Political Economy*, 94(5):1002–1037.

Shapiro, C. (2019). Protecting competition in the American economy: Merger control, tech titans, labor markets. *Journal of Economic Perspectives*, 33(3):69–93.

Smith, A. (1776). *The Wealth of Nations*, volume One. W. Strahan and T. Cadell, London.

Solow, R. M. (1957). Technical change and the aggregate production function. *Review of Economics and Statistics*, pages 312–320.

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

Syverson, C. (2019). Macroeconomics and market power: Context, implications, and open questions. *Journal of Economic Perspectives*, 33(3):23–43.

Traina, J. (2018). Is aggregate market power increasing? Production trends using financial statements. Working Paper No. 272, Stigler Center for the Study of the Economy and the State.

Uchiyama, T. (2018). Quasi-competitiveness in the Cournot model with heterogeneous firms. *Economics Letters*, 165:62–64.

United States Securities and Exchange Commission (2006). Rite Aid Corporation: Form 10-K for fiscal year ended February 26, 2005.

United States Securities and Exchange Commission (2007). Mccormick and Company, Incorporated: Form 10-K for fiscal year ended November 30, 2006.

Varian, H. R. (1992). *Microeconomic Analysis*, volume 3. Norton New York.

Verdoorn, P. J. (1949). On the factors determining the growth of labor productivity. *Italian Economic Papers*, 2:59–68.

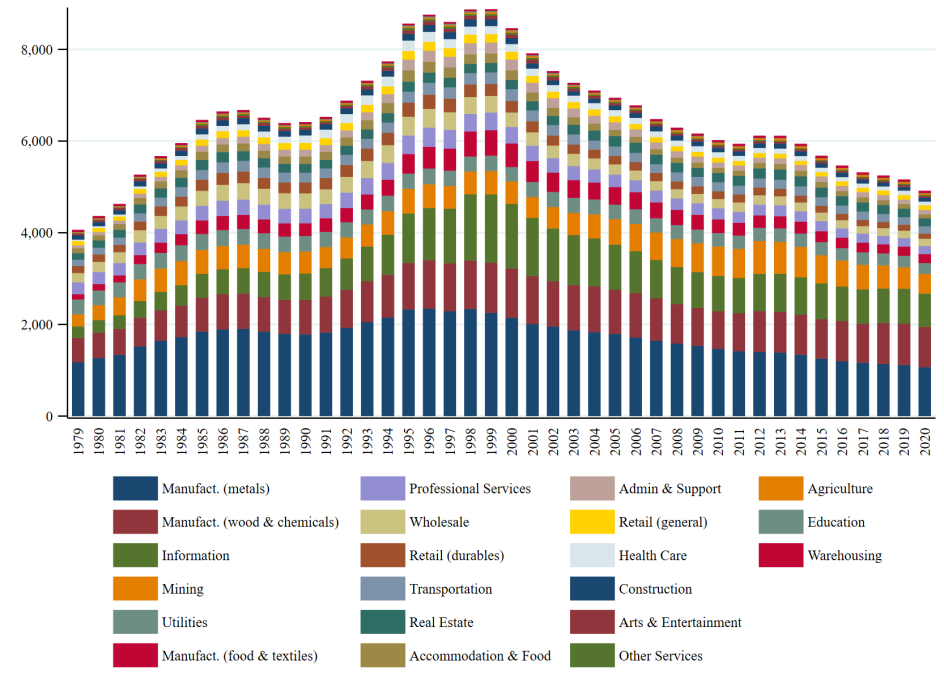
Young, A. (1928). Increasing returns and economic progress. *The Economic Journal*, 38:527–542.

Table 1: Summary statistics 1980-2019

	Acronym	Mean	Median	No. Obs
Sales	SALE	2623015	144853	266732
Cost of goods sold	COGS	1839702	88073	266732
Selling, general & admin.	XSG&A	353950	17043	266732
Capital stock	PPEGT	2662754	62840	266732

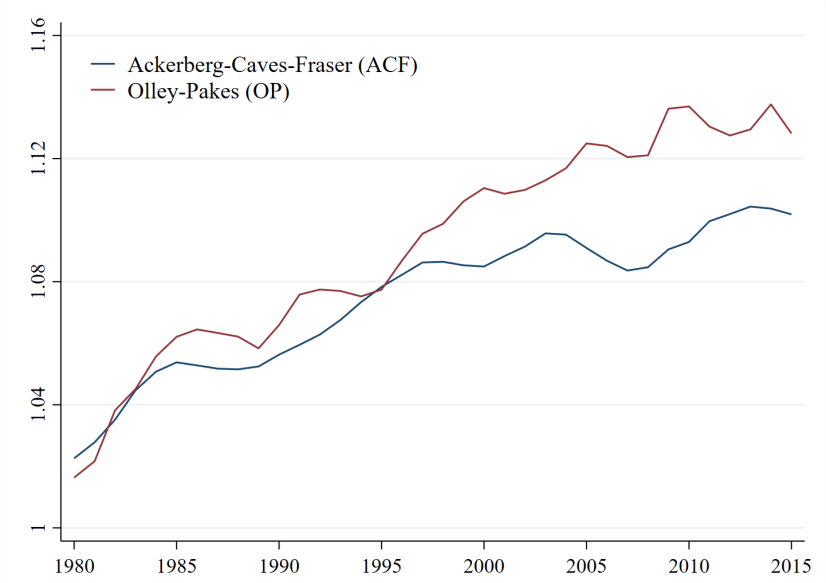
Notes: Thousands USD deflated using BEA's Chain-Type Price Indexes for Gross Output by Industry with base year 2012. The second column contains the Compustat acronym.

Figure 1: Number of firms in Compustat by year by industry

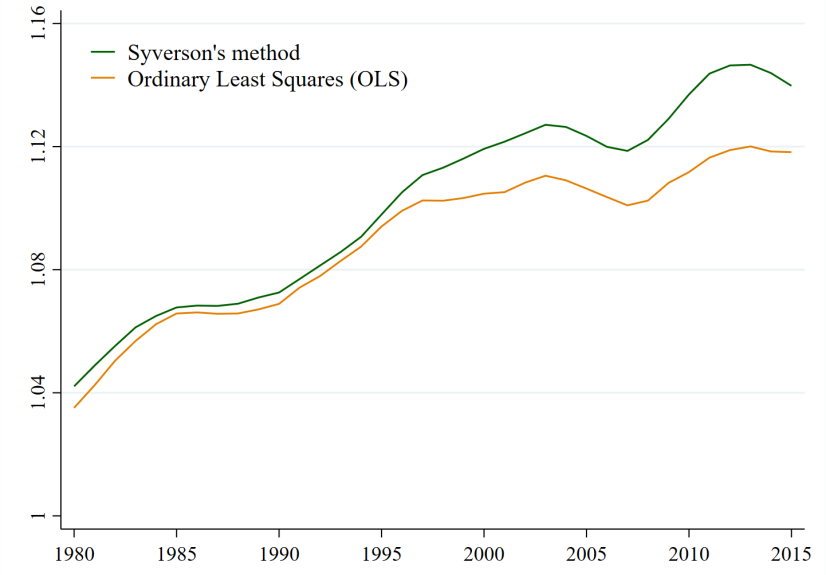


Note: The down and to the right order of the legend labels matches the bottom-up order of the bar colors.

Figure 2: Aggregate elasticity of scale estimates



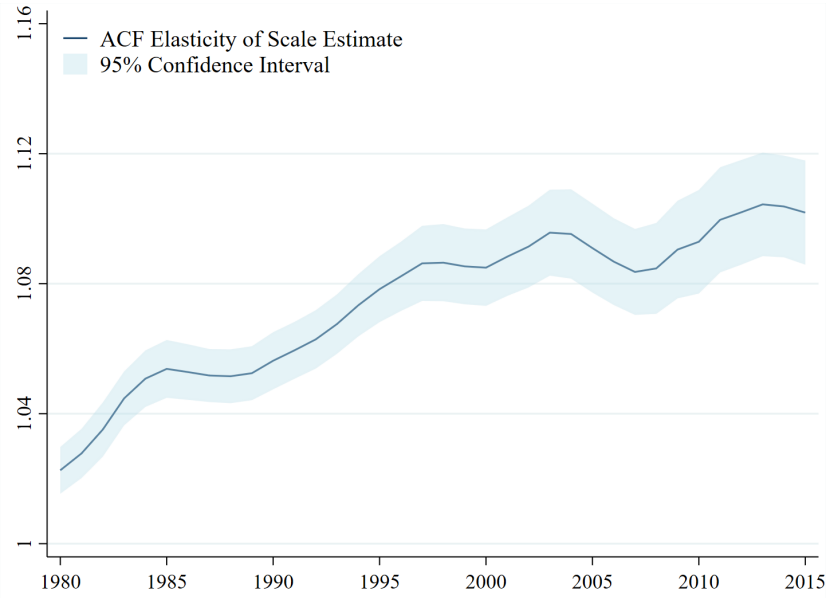
(a)



(b)

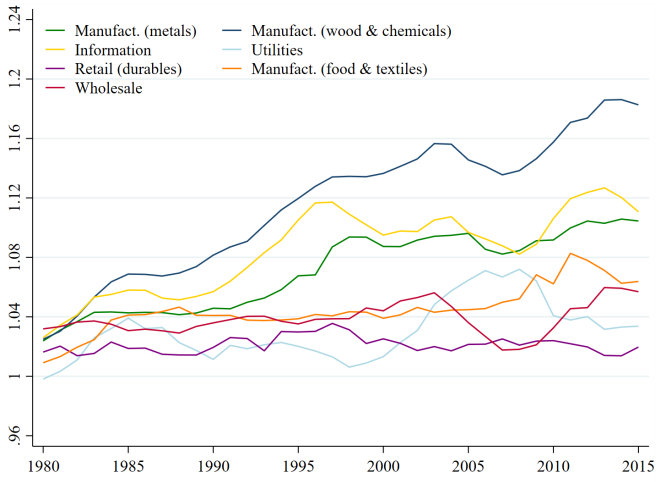
Notes: Panel (a) shows the estimated aggregate elasticity of scale using the ACF and OP methods; panel (b) shows the estimated aggregate elasticity of scale using the OLS and Syverson's methods. The labels on the horizontal axes mark the first year of the five-year rolling period, so a data point marked 1990 is the elasticity of scale estimated for the period 1990 to 1994.

Figure 3: Aggregate elasticity of scale estimate using the ACF method with 95% confidence interval

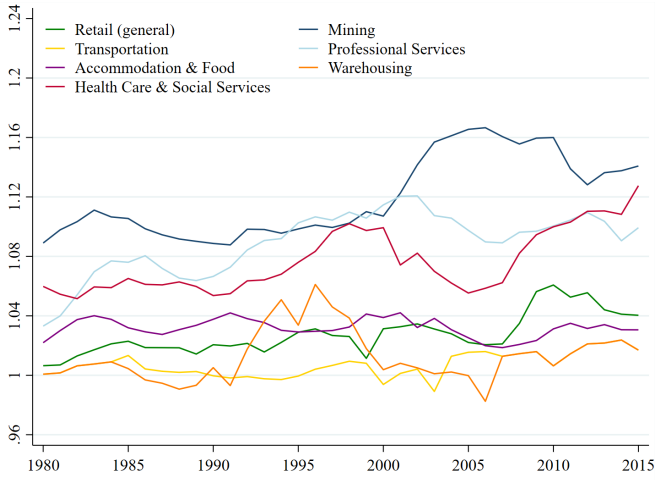


Note: The labels on the horizontal axes mark the first year of the five-year rolling period, so a data point marked 1990 is the elasticity of scale estimated for the period 1990 to 1994.

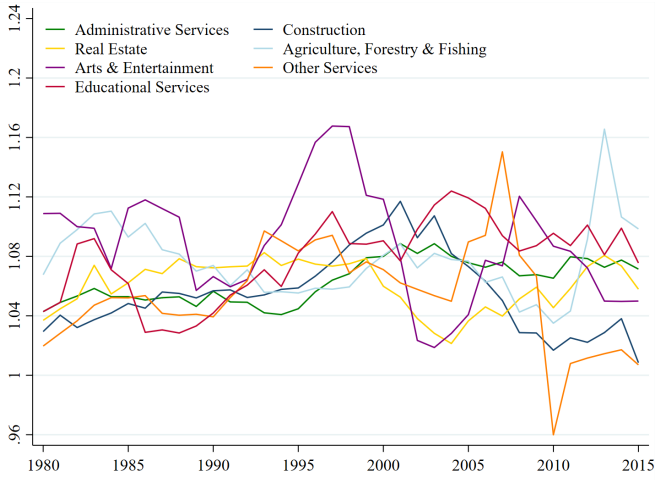
Figure 4: Industry-specific elasticity of scale estimates



(a)



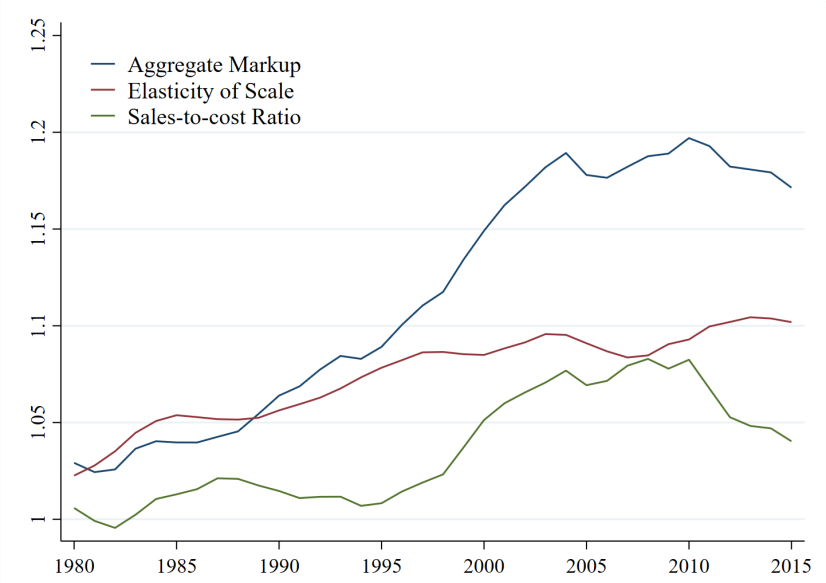
(b)



(c)

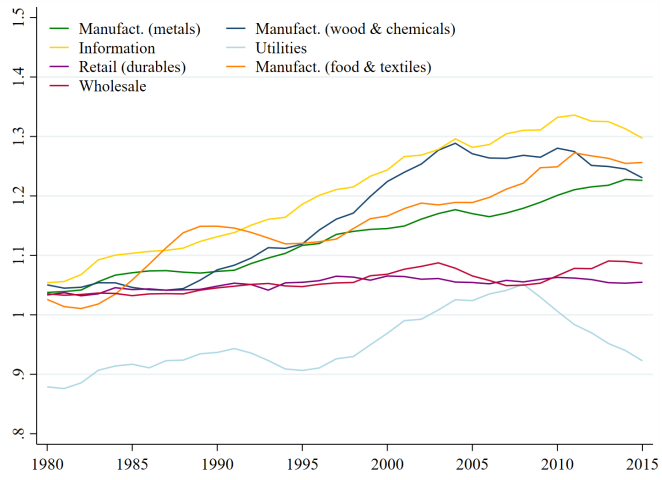
Notes: The figures show the estimated elasticities of scale using the ACF method by industry, with Panel (a) focusing on the top 7 industries, Panel (b) the next 7 industries, and Panel (c) the bottom 7 industries by total revenues. The labels on the horizontal axes mark the first year of the five-year rolling period, so a data point marked 1990 is the elasticity of scale estimated for the period 1990 to 1994.

Figure 5: Decomposition of the aggregate markup estimates

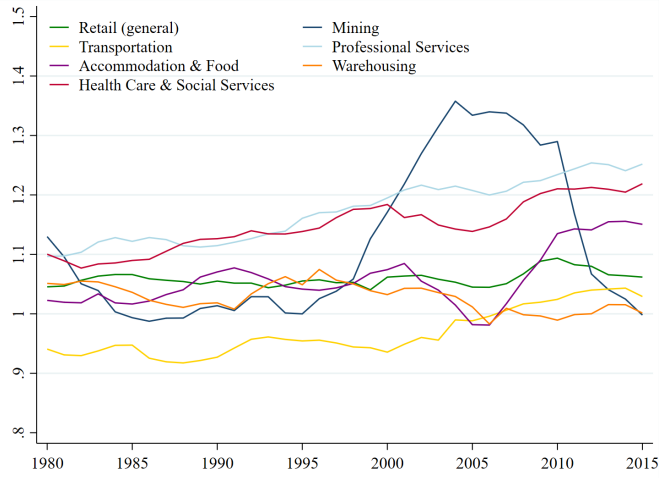


Note: The labels on the horizontal axes mark the first year of the five-year rolling period, so a data point marked 1990 is the elasticity of scale estimated for the period 1990 to 1994.

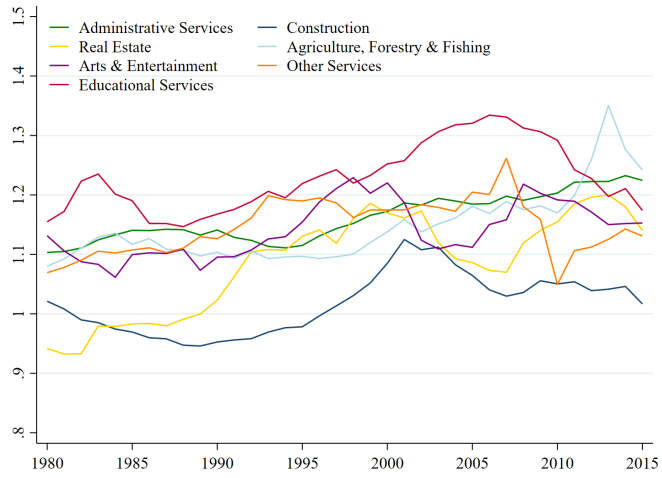
Figure 6: Industry-specific markup estimates



(a)



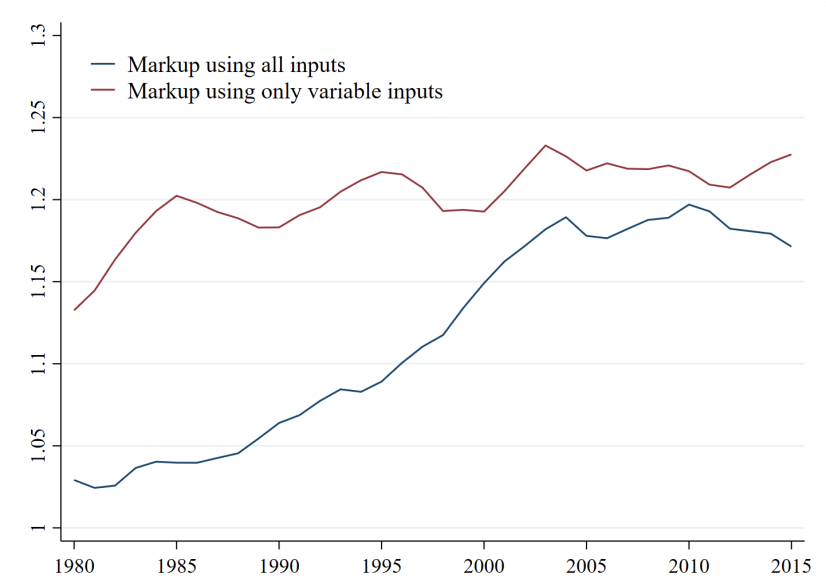
(b)



(c)

Notes: The figures show the estimated markups by industry, with Panel (a) focusing on the top 7 industries, Panel (b) the next 7 industries, and Panel (c) the bottom 7 industries by total revenues. The labels on the horizontal axes mark the first year of the five-year rolling period, so a data point marked 1990 is the elasticity of scale estimated for the period 1990 to 1994.

Figure 7: Aggregate markup estimate using only variable inputs vs. all inputs



Note: The labels on the horizontal axes mark the first year of the five-year rolling period, so a data point marked 1990 is the elasticity of scale estimated for the period 1990 to 1994.