Heterogeneous Markups Cyclicality and Monetary Policy*

Andrea Chiavari[†] Marta Morazzoni[‡] Danila Smirnov[§]

First draft: July 2021. This Version: August 2023.

Abstract

Firms' markups cyclicality is at the heart of monetary policy transmission in New Keynesian models. Using US Compustat data and employing local projection techniques, we uncover a novel fact: dominant firms have a countercyclical markup response after an unexpected contractionary monetary policy shock. Building a heterogeneous firms New Keynesian model with demand accumulation and endogenous markups that evolve over firms' life-cycle, we show that this can be due to the different demand elasticities faced by firms. Dominant firms face a more inelastic demand, which implies a lower pass-through rate from costs to prices. Therefore, after a contractionary monetary policy shock, dominant firms pass less the reduction in marginal costs to prices compared to competitors, and increase their markups by more, as documented empirically. After calibrating the model to US micro-level data, we also find that firms' heterogeneous demand elasticities can lead to the amplification of monetary policy shocks.

Keywords: Markups, Heterogeneous Firms, Firm Life-Cycle, Monetary Policy Shocks

JEL Codes: D4, E2, E52, L1, O4

*We are grateful to Isaac Baley, Andrea Caggese, James Cloyne, Davide Debortoli, Andrea Fabiani, Jordi Galí, Basile Grassi, Priit Jeenas, François Le Grand, Virgiliu Midrigan, Benjamin Moll, Emi Nakamura, Galo Nuño, Edouard Schaal, Jon Steinsson, Vincent Sterk, Gianluca Violante, and Thomas Winberry, as well as to the participants at CREi seminars, University of Barcelona, BGSE Jamboree 2021, PhD-EVS 2022, RES 2022, BSE Summer Forum 2022, and the 2022 conference on Individual Risks and the Macroeconomy at Sciences Po. The views expressed herein are those of the author and should not be attributed to the IMF, its Executive Board, or its management.

[†]University of Oxford. Email: andrea.chiavari@economics.ox.ac.uk.
[‡]University College London and Institute for Fiscal Studies. Email: m.morazzoni@ucl.ac.uk.
[§]International Monetary Fund. Email: dsmirnov@imf.org.

1 Introduction

While the cyclicality of aggregate markups has long been a subject of macroeconomic investigation, it continues to be a significant area of research due to markups' key role in transmitting business cycle fluctuations and monetary impulses. Simultaneously, recent studies have drawn attention to the macroeconomic relevance of heterogeneous market power across firms, a topic that is gaining prominence due to the increasing availability of firm-level datasets, which facilitate markup estimation from balance sheet information. Yet, empirical evidence concerning the heterogeneity in the response of firm-level markups following interest rate movements remains still limited.

This project addresses the question of how markups respond heterogeneously to monetary policy shocks, and how these responses can inform quantitative monetary models. Our study marks a first step towards closing this gap in the literature, and contributes to the body of previous works in two fundamental ways. First, it documents substantial differences in the response of markups to monetary policy shocks by firm age. Second, it evaluates the macroeconomic significance of these differences within a novel New Keynesian framework, enriched with firms' heterogeneity, demand accumulation dynamics, and the evolution of endogenous markups over firms' life cycles.

We begin by building a novel New Keynesian (NK) framework with nominal rigidities à la Rotemberg, which is enriched with firm heterogeneity, a process of exogenous demand accumulation, and endogenous markups. Moreover, the model allows for new firms to enter the market in each period, replacing those that exit exogenously, and it is thus able to replicate realistic firm and markup dynamics, while still maintaining quantitative tractability. To be more precise, our framework possesses two main desirable features. First, firms exhibit a realistic life-cycle profile, as they undergo a demand accumulation process characterized by some persistence and idiosyncratic shocks. This process is also characterized by a long-run mean, enabling the demand faced by companies to increase over time.[1] Second, firms display a realistic markup behavior, as we adopt the approach of Klenow and Willis (2016), wherein firms' goods are aggregated à la Kimball. This particular choice introduces endogenous markups into the model in a tractable manner, as the elasticity of substitution across intermediate goods decreases with their relative quantity.

We then discuss how to test the model insights using firm-level data. Our model predicts that firms' markups should respond differently to monetary policy (MP) shocks based on firms' relative size, that is, their dominance within their respective markets. However, typical firm-level data do not allow for a precise and satisfactory market definition. To overcome this data limitation, we propose the use of a readily available measure, namely firm age, as a proxy for market dominance.

[1] The observation that firms grow over their life-cycle mostly by accumulating demand is consistent with the findings of a growing body of empirical research started by Foster, Haltiwanger, and Syverson (2008).

We argue that firm age serves as an effective proxy for several reasons: (i) it captures similar economic forces that shape firms' relative size as well, offering a robust approximation of market dominance; (ii) it is an endogenous but predetermined variable to aggregate shocks; and (iii) it has been shown to be an important predictor of heterogeneous firm-level responsiveness to shocks.[2]

To empirically investigate the heterogeneous response of firm-level markups to MP shocks, we use Compustat, a dataset that offers two key advantages for our analysis. First, it has been already extensively exploited to measure firm-level markups, using balance sheet information and the methodology introduced by De Loecker and Warzynski (2012). Second, it is available at a quarterly frequency and for the 1990-2016 period, which allows us to study the dynamics of markups over the business cycle. We combine the sample from Compustat with the exogenously identified MP shock series from Jarociński and Karadi (2020), and base our empirical strategy on the local linear projection approach proposed by Jordà (2005). As discussed, we focus on firms' age as the main explanatory variable to capture heterogeneity. However, we conduct a series of robustness checks in order to show that resorting to alternative proxies for firm dominance, such as sales shares, and exploring potential covariates linked to the balance sheet channel do not affect our findings.

Our empirical findings yield several novel insights. To begin with, substantial heterogeneity emerges in the response of firms' markups to MP shocks, especially when distinguishing between older (above median age) and younger firms. Specifically, older firms exhibit a significant countercyclical response to contractionary MP shocks, while younger firms display a mildly procyclical response. Furthermore, our investigation into alternative explanatory variables, such as firms' relative size within their industry, leverage, liquidity, and assets, uncovers that these factors have minimal to no predictive power in explaining the observed heterogeneity in markup responses across firms. Note that, while it lies beyond the scope of this paper to present definitive evidence regarding the cyclical behavior of aggregate cost-weighted markups, it is worth mentioning that, as older firms represent approximately 75 percent of total costs in the data, our results hint towards a mildly countercyclical response of the aggregate markup in the US, at least to demand-like shocks.

The model is calibrated to the US economy to capture several critical aspects of the life-cycle of firms' markups. The validation analysis demonstrates that our model successfully reproduces various untargeted moments, including the presence of a fat right tail in the distribution of markups and the empirical relationships between firms' wage bill and their size. This latter moment, as emphasized recently by Edmond, Midrigan, and Xu (2018), is closely tied to the fact that firms wielding market power can bolster their profits by reducing quantities and increasing prices, leading to a reduction in labor demand and, consequently, in firms' labor income share. As additional untar-

[2]See, for instance, the argument and empirical strategies in Cloyne et al. (2023) and Colciago et al. (2019).

geted moments, the model replicates realistic steady-state distributions of firms and employment shares by firm age. Furthermore, it accurately matches the growth rates of sales and employment.

Importantly, the novel NK framework we have built speaks to the differential response of markups to MP shocks by firm age that we empirically document in Compustat. As previously mentioned, old firms in our model economy face a lower pass-through from costs to prices due to the presence of the Kimball aggregator. When hit by a contractionary MP shock that decreases wages and puts a downward pressure on prices, dominant firms can cut prices by relatively less compared to young ones. Since markups depend on the ratio between prices and marginal costs, this mechanism is in turn responsible for the countercyclical response in old firms' markups. In particular, we can match almost 30% of the empirically estimated relative difference in old and young firms' markups responses to a negative MP shock. In our analysis, we also show that the differential response of old firms in the model can be quantitatively decomposed to highlight the contribution of changes in other aggregate variables to the general equilibrium impact of MP shocks on markups. Specifically, we find that changes in the real wage generated by a negative MP shock are key in shaping the differential behavior of old (and hence dominant) firms' markups.

Finally, we conclude our quantitative analysis with an investigation of the shock amplification mechanisms at play in our framework, comparing our set up to a standard one-firm NK model with price rigidities. Both the presence of the Kimball aggregator and the heterogeneity of firms are shown to affect the way and extent to which MP shocks transmit in the economy, with output decreasing on average by roughly 20 percentage points (p.p.) more after a negative movement in the interest rate. Focusing on the role of the Kimball aggregator, since intermediate firms – especially old ones – temper their price drops after a negative MP shock due to the increase in their desired markup, MP shocks propagate more through quantities than through prices in our set up, as opposed to the standard constant elasticity NK framework. At the same time, the presence of the Kimball aggregator alone is not sufficient to generate the observed amplification of MP shocks, as its effects on the elasticity of demand faced by firms kick in when firms are indeed heterogeneous and have different pass-throughs from costs to prices. Hence, we conclude that firm heterogeneity is key in affecting and amplifying the response of macroeconomic aggregates to MP shocks.

Related Literature. Our paper contributes to multiple strands of the existing literature. First, it expands the body of work examining the empirical cyclicality of markups. Prior studies in this area have mostly focused on analyzing the cyclicality of aggregate markups, as done in Gali, Gertler, and Lopez-Salido (2007), Hall (1988), Bils, Klenow, and Malin (2018), and Nekarda and Ramey (2020). A parallel strand of research has instead delved into the cyclical behavior of firm-level markups, with notable contributions from Hong (2017), Alati (2020), Burstein et al. (2020), and Afrouzi and Caloi (2022). Compared to these papers, our question does not address markups comovements with respect to the business cycle, but rather in response to MP shocks, as in Meier and Reinelt (2022). While Meier and Reinelt (2022) investigate the response of markups dispersion to MP shocks through the lens of differential degrees of price stickiness across firms, we instead study the response of markups to MP shocks in a model where heterogeneous demand elasticities give rise to different pass-throughs across firms. After assessing the model-implied relation between a firm's age and demand elasticity, we also provide empirical estimates of the heterogeneous response of young and old firms' markups to MP shocks in the context of the US economy.

Second, we contribute to the theoretical and quantitative literature that has started incorporating firm-level heterogeneity into NK frameworks to understand its implications for the transmission of monetary policy. Most works in this field have developed monetary models with firm heterogeneity assuming a constant elasticity of substitution across firms, including Bilbiie, Ghironi, and Melitz (2007), Golosov and Lucas (2007), Bilbiie, Ghironi, and Melitz (2012), Jeenas (2019), Ottonello and Winberry (2020), Fang (2022), and Hamano and Zanetti (2022).[3] A recent research strand has instead ventured beyond the assumption of constant elasticity of substitution, by introducing realistic firm-level market power in monetary models. Papers in this literature include Gopinath and Itskhoki (2010), Klenow and Willis (2016), Mongey (2017), Baqaee, Farhi, and Sangani (2021), Wang and Werning (2022), Alvarez, Lippi, and Souganidis (2022).[4] We complement these studies by presenting a quantitative and tractable monetary model with firm dynamics, which incorporates a realistic notion of market power and is grounded in novel evidence on the cyclicality of markups. We then show how the presence of heterogeneous demand elasticities across firms alters and amplifies the propagation of MP shocks compared to standard NK models.

Finally, our paper is closely related to a group of empirical works – initially pioneered by Gertler and Gilchrist (1994) – that investigates how the effects of monetary policy can vary across firms with different characteristics. Much of this literature has generally centered around the behavior of firm investment, as exemplified by studies such as Jeenas (2019), Ottonello and Winberry (2020), Meier and Reinelt (2022), and Cloyne, Ferreira, Froemel, and Surico (2023). Notably, Fabiani, Falasconi, and Heineken (2020) employ a similar empirical strategy to ours, though their focus is on debt maturity. Our contribution builds upon the methodologies and insights of these studies. However, our primary innovation lies in documenting the heterogeneous responses of markups to monetary policy shocks, specifically focusing on the distinction between young and old firms.

[3]It is worth noting that we do not aim to provide an exhaustive survey of the extensive literature influenced by the seminal paper of Golosov and Lucas (2007), although our work certainly builds upon this entire line of research.

[4]Note that Höynck, Li, and Zhang (2023) study how the the rise in sectoral markups has lead to a strengthening of monetary policy transmission, while we focus on the role of within-sectors markups in affecting monetary transmission.

Outline. The paper is organized as follows: Section 2 presents the theoretical framework and explains our strategy to bring the model to the data. Section 3 discuss the dataset we use, the construction of the main variables of interest for the empirical analysis, and the measurement of markups. Section 4 highlights our empirical contribution and main results. Then, in Section 5, we illustrate the calibration and quantitative fit of the model, while in Section 6 we present steady state results and firm-level impulse responses to monetary policy shocks, while also assessing shocks amplification mechanisms. In Section 7, we finally conclude and discuss the way ahead.

2 Theoretical Framework

The following section presents the quantitative framework we build and use throughout the main analysis, and then explains the empirical strategy we adopt in order to bring it to available data. Appendix A contains additional derivations related to the model and discusses its solution method.

2.1 Household

Time is continuous. The model features a representative household that optimizes the discounted flow of utility from consumption and labor over an infinite lifetime horizon. We indicate the discount factor as $\rho \geq 0$. We also assume that the utility of the agent is strictly increasing and concave in consumption, and strictly decreasing and convex in the number of hours worked. Preferences are time-separable and the infinite stream of household's utility is hence given by:

$$\int_0^\infty e^{-\rho t} \left(\frac{C_t^{1-\nu}}{1-\nu} - \frac{L_t^{1+\gamma}}{1+\gamma} \right) dt, \tag{1}$$

where ν represents the risk aversion in the utility function over consumption, and γ is the inverse of the Frisch elasticity. Moreover, $L_t \in [0,1]$ are hours supplied as a fraction of the time endowment (normalized to 1), while C_t denotes the aggregate consumption good. In each period, the household can borrow in bonds B_t at the real interest rate r_t. Finally, the household owns all the firms in the economy, supplies labor, and chooses the paths of aggregate consumption and bond investment as a result of a standard maximization problem, given by the following expressions:

$$\begin{aligned} \mathcal{V} &= \max_{\{C_t, L_t, \dot{B}_t\}} \int_0^\infty e^{-\rho t} \left(\frac{C_t^{1-\nu}}{1-\nu} - \frac{L_t^{1+\gamma}}{1+\gamma} \right) dt, \\ \text{s.t.} \quad & C_t + \dot{B}_t = W_t L_t + r_t B_t + D_t; \end{aligned} \tag{2}$$

where we denote by D_t the dividends from the firms owned by the household, and by W_t the wage earned in real terms. Solving for the optimal value of consumption and labor, we get the following standard Euler and labor supply equations:

$$r_t = \rho + \nu \frac{\dot{C}_t}{C_t}, \tag{3}$$

$$W_t = L_t^{\gamma} C_t^{\nu}. \tag{4}$$

2.2 Final Good Producer

Turning to the supply side, a competitive representative final-good producer aggregates a continuum of intermediate inputs indexed by $i \in [0,1]$ according to the following expression:

$$\int_0^1 \mathcal{K}(x_{i,t}) di = 1, \tag{5}$$

and

$$x_{i,t} = a_{i,t} \frac{y_{i,t}}{Y_t}, \tag{6}$$

where $x_{i,t}$ is the firm-level *demand-adjusted relative size* and intermediate inputs $y_{i,t}$ are aggregated using the Kimball aggregator $\mathcal{K}(\cdot)$, with $\mathcal{K}'(\cdot) > 0$, $\mathcal{K}''(\cdot) < 0$, and $\mathcal{K}(1) = 1$. Notice that the CES aggregator obtains as a special case of the Kimball aggregator, and namely when $\mathcal{K}(x) = x^{\frac{\sigma-1}{\sigma}}$ for an elasticity of substitution $\sigma > 1$. Importantly, Y_t denotes aggregate output, while $a_{i,t}$ is a realization of a stochastic demand process that will be explained in due detail later.

Taking the relative prices $p_{i,t}$ of any intermediate input i as given, the final good producer minimizes production costs subject to Equation (5). The optimality condition of this problem gives rise to the *inverse demand* function for good i, given by:

$$p_{i,t} = \mathcal{K}'(x_{i,t}) a_{i,t} \mathcal{D}_t, \tag{7}$$

where:

$$\mathcal{D}_t = \left(\int_0^1 \mathcal{K}'(x_{i,t}) x_{i,t} di \right)^{-1} \tag{8}$$

is a *demand index*. In the CES case this index is a constant and equal to $\mathcal{D}_t = \frac{\sigma}{\sigma-1}$, and hence Equation (7) reduces to the familiar constant elasticity demand curve given by $p_{i,t} = x_{i,t}^{-1/\sigma}$. Following recent literature, we use the Klenow and Willis (2016) specification for $\mathcal{K}(x)$, which is given by:

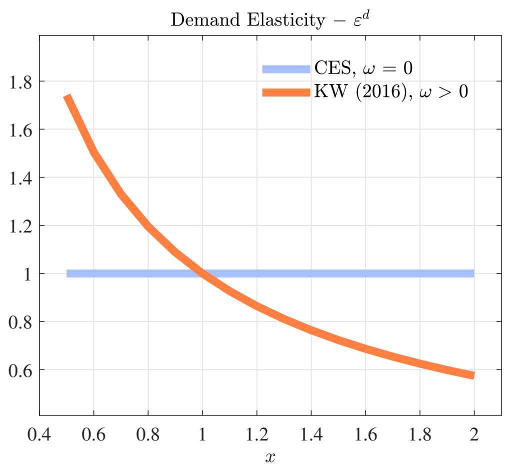
$$\mathcal{K}(x) = 1 + (\sigma - 1) \exp\left(\frac{1}{\omega}\right) \omega^{\frac{\sigma}{\omega}-1} \left[\Gamma\left(\frac{\sigma}{\omega}, \frac{1}{\omega}\right) - \Gamma\left(\frac{\sigma}{\omega}, \frac{x^{\omega/\sigma}}{\omega}\right) \right]. \tag{9}$$

Note that $\sigma > 1$, $\omega \geq 0$ and $\Gamma(s,x)$ is the upper incomplete Gamma function. In particular, this latter assumption satisfies $\Gamma(s,x) := \int_x^\infty t^{s-1}e^{-t}dt$. Moreover, the parameter ω denotes the *super elasticity*, which in the zero-limit turns the Kimball aggregator into the standard CES aggregator.

Finally, we can derive an analytical expression for the elasticity of demand faced by a producer of any good variety i as a function of the relative quantity of good i in the economy, given by:

$$\varepsilon_{i,t}^d = \sigma x_{i,t}^{-\frac{\omega}{\sigma}}. \tag{10}$$

Figure 1: CES and Kimball Demand Elasticity for Firms of Different Size



Note. Figure 1 shows the demand elasticity, for both the CES and the Kimball specifications, for firms of different demand-adjusted relative sizes, $x = ay/Y$.

Figure 1 shows the firm-level demand elasticity as a function of different demand-adjusted relative sizes, $x_{i,t} = a_{i,t}y_{i,t}/Y_t$, for the CES and the Kimball case. As already pointed out, the standard CES case is recovered when $\omega = 0$ and hence the elasticity of demand is constant across producers and given by $\varepsilon_{i,t}^d = \sigma$. In contrast, in the case of the Kimball aggregator, the elasticity of substitution is lower for firms with higher relative size, implying that larger firms face a more inelastic demand and hence can charge higher markups. This feature allows us to account for variable markups in a fully dynamic model, and represents a key advantage compared to static frameworks of imperfect competition, as the one in Atkeson and Burstein (2008) for instance.

2.3 Intermediate Good Producers

Each intermediate good i is produced by a monopolistically competitive firm, by means of using effective units of labor $\ell_{i,t}$ in the production process and according to a technology given by:

$$y_{i,t} = \ell_{i,t}^{1-\alpha}, \tag{11}$$

with $\alpha \in [0,1]$ governing the returns to scale in production and $\ell_{i,t}$ being the amount of labor hired by the firm at the wage rate W_t. Intermediate producers are monopolistic competitors on their respective variety, and face a demand function that can be written explicitly from Equation (7) as:

$$y_{i,t} = \left(1 - \omega \log\left(\frac{\sigma}{\sigma-1}\frac{1}{a_{i,t}}\frac{p_{i,t}}{\mathcal{D}_t}\right)\right)^{\sigma/\omega} \frac{Y_t}{a_{i,t}}. \tag{12}$$

In the model, firms exit at the Poisson arrival rate δ in each period and are replaced by new firms, such that the mass of firms operating in equilibrium is constant. Moreover, as anticipated above, each intermediate firm i is characterized by a process of demand accumulation given by:

$$d\log(a_{i,t}) = \rho_a(\bar{a} - \log(a_{i,t}))dt + \psi^a d\mathcal{W}_{i,t} \quad \text{and} \quad \log(a_{i,0}) \sim \mathcal{N}(\bar{a}_0, \psi_0^a), \tag{13}$$

where ρ_a is the persistence, $\bar{a}$ is the long run mean, $\log(a_{i,0})$ is the demand at entry, and $\psi^a d\mathcal{W}$ is the idiosyncratic risk component, with $d\mathcal{W}$ being a standard Wiener process. Throughout the paper, we work with the empirically relevant case in which $\bar{a} > \bar{a}_0$. As explained in more detail below, this implies that we assume firms' demand to grow on average over their life-cycle.

It is important to emphasize that our results do not depend on the specific assumption to model firm-level heterogeneity as stemming from a demand accumulation process over the life cycle. We could instead allow productivity shocks to characterize firm heterogeneity, or opt for a combination of the two (though the latter choice would involve an increase in the number of state variables in the model). Our choice is motivated by micro-level empirical evidence from Foster, Haltiwanger, and Syverson (2008), which shows that idiosyncratic demand differences explain most of the selection at the firm-level, and that firms grow over time by accumulating demand. Hence, in our model, firms enter with little demand and slowly accumulate it: while accumulating demand, they also expand and charge higher prices—hence, higher markups as well—which is precisely consistent with the findings in Foster, Haltiwanger, and Syverson (2008). Our demand process is purposely meant to be computationally tractable and may rationalize some underlying form of customer accumulation or consumer habit formation, as in Ravn, Schmitt-Grohé, and Uribe (2006).

Intermediate firms are subject to nominal rigidities, that is, adjustment costs when changing prices, which we assume to be proportional to their sales, paid in terms of the labor cost W_t, and quadratic. The joint distribution of firms over the idiosyncratic states is defined as $\lambda_t(da, dp)$. Finally, intermediate producers discount the future stream of profits at the rate $r_t + \delta$. Hence, we can summarize the problem of a given firm i as follows:

$$
\begin{aligned}
\mathcal{J}_{i,0} = & \max_{\{\dot{p}_{i,t},\ell_{i,t},y_{i,t}\}_{t\geq 0}} \mathbb{E}_0 \int_0^\infty e^{-\int_t^\infty (r_t+\delta)dt} \left\{ p_{i,t}y_{i,t} - W_t \left(\ell_{i,t} + \frac{\vartheta}{2}\left(\pi_t + \frac{\dot{p}_{i,t}}{p_{i,t}}\right)^2 p_{i,t}y_{i,t} \right) \right\} dt, \\
\text{s.t.} \quad y_{i,t} = & \left(1 - \omega \log\left(\frac{\sigma}{\sigma-1}\frac{1}{a_{i,t}}\frac{p_{i,t}}{\mathcal{D}_t}\right)\right)^{\sigma/\omega} \frac{Y_t}{a_{i,t}}, \\
y_{i,t} = & \ell_{i,t}^{1-\alpha}, \\
d\log(a_{i,t}) = & \rho_a(\bar{a} - \log(a_{i,t}))dt + \psi^a d\mathcal{W}_{i,t}, \\
\log(a_{i,0}) \sim & \mathcal{N}(\bar{a}_0, \psi_0^a), \\
p_{i,0} & \text{ given.}
\end{aligned}
\tag{14}
$$

We set the initial price of entrant firms $p_{i,0}$ such that it maximizes the value $\mathcal{J}_{i,0}$ for a given initial value of firm's productivity $a_{i,0}$. Intermediate firms take as given equilibrium paths for the real wage $\{W_t\}_{t\geq 0}$, aggregate output $\{Y_t\}_{t\geq 0}$, Kimball demand index $\{\mathcal{D}_t\}_{t\geq 0}$, inflation $\{\pi_t\}_{t\geq 0}$, and the interest rate $\{r_t\}_{t\geq 0}$. In steady state, the recursive solution to this problem consists of decision rules for labor $\ell(a, p; \mathcal{S})$ and output $y(a, p; \mathcal{S})$, with $\mathcal{S} := (r, W, Y, \mathcal{D}, \pi)$. In turn, these rules also imply optimal drifts for prices: together with a stochastic process for a, they also induce a stationary joint distribution of firms, which is given by $\lambda(da, dp; \mathcal{S})$ and is characterized by a standard Kolmogorov forward equation. Note that, out of the steady state, each of these objects is time-varying and depends on the time path of prices and policies: $\{\mathcal{S}_t\}_{t\geq 0} := \{r_t, W_t, Y_t, \mathcal{D}_t, \pi_t\}_{t\geq 0}$.

In Appendix A.1, we also report a decomposition of the Rotemberg nominal price adjustment cost into aggregate inflation and firm relative price change. As a final note, in the numerical solution of the model, we use a variable substitution method for which we replace firms' relative prices with a function of their market shares. The detailed explanation can be found in Appendix A.2.

2.4 Monetary Authority

Our model economy features a monetary authority that sets the nominal interest rate according to a standard simple interest rate rule given by:

$$
i_t = \rho + \phi_\pi(\pi_t - \pi^*) + \varepsilon_t^m,
$$

where ρ is the discount factor, $\pi^* = 0$, $\phi_\pi > 1$, and ε_t^m is the monetary policy shock that can be mapped directly into empirically identified monetary policy shocks series, such as, for example, the one from Jarociński and Karadi (2020) that we use in our empirical analysis in Section 4.

Note that $\varepsilon_t^m = 0$ in steady state, and one of the main quantitative exercises of the following sections will be precisely to study the economy's adjustment after an unexpected temporary monetary shock, namely after a change in ε_t^m. Finally, given inflation π_t and the nominal interest rate i_t, the real return on bonds r_t is determined by the Fisher equation, as in $r_t = i_t - \pi_t$.

2.5 Equilibrium Conditions

An equilibrium in this economy is defined as a set of paths for household's $\{C_t, L_t\}_{t\geq 0}$ and firms' decisions $\{\dot{p}_{i,t}, \ell_{i,t}, y_{i,t}\}_{t\geq 0}$, input prices $\{W_t\}_{t\geq 0}$, the return on bonds $\{r_t\}_{t\geq 0}$, the inflation rate $\{\pi_t\}_{t\geq 0}$, the distribution of firms $\{\lambda_t\}_{t\geq 0}$, the demand index $\{\mathcal{D}_t\}_{t\geq 0}$, and aggregate quantities such that, at every t: (i) the household and the firms maximize their objective functions taking as given equilibrium prices and aggregate quantities; (ii) the sequence of distributions satisfies aggregate consistency conditions; (iii) all markets clear. There are three markets in our economy: the bond, the labor, and the goods market. The bond market clears when the following holds:

$$B_t = 0, \tag{15}$$

The goods market clears according to the following condition:

$$C_t = Y_t,$$

Finally, the labor market clears when:

$$L_t = \int \left(\ell_t(a,p) + \frac{\vartheta}{2} \left(\pi_t + \frac{\dot{p}_t(a,p)}{p} \right)^2 p y_t(a,p) \right) d\lambda_t.$$

Unpacking the above expression, it is important to stress again that L_t is the aggregate labor demand, $\ell_{i,t}$ is the demand of the intermediate firm i for the labor that is used in production, and the second term inside the integral is the labor needed to cover the cost of price adjustments.

2.6 Mapping the Model to the Data

The objective of developing a quantitatively tractable but realistic monetary model with firms' market power is to study how markups respond to monetary policy shocks, and later assess the macroeconomic relevance of any firm-level heterogeneity in the pass-through of aggregate shocks. As a first step, in this subsection, we use a simplified version of our framework to show analytically how prices – and hence markups – respond to aggregate changes, and highlight which sufficient statistic can explain the heterogeneity in these responses at the firm-level. Second, we discuss a potential strategy to proxy this sufficient statistic in available US data, and then provide an empirical estimate of the heterogeneous response of firm-level markups to a monetary policy shock.

2.6.1 Intuition Through a Simplified Model

Abstracting from the presence of nominal rigidities, we can show analytically how prices respond to changes in aggregate variables. In our quantitative model, monetary policy shocks affect prices and markups mostly via their effect on aggregate variables: as such, this specific exercise can inform how to measure the heterogeneous response of markups to changes in the interest rate.

In particular, we look at the heterogeneous response of firm-level prices to changes in aggregate output Y and wage W, which we know to respond positively to positive monetary policy shocks. Also, note that both aggregates will be the main drivers of the response of markups, as explained in full detail in the quantitative section. The demand function in this environment is still given by:

$$y = \left(1 - \omega \log\left(\frac{\sigma}{\sigma-1}\frac{1}{a}\frac{p}{\mathcal{D}}\right)\right)^{\sigma/\omega}\frac{Y}{a}, \tag{16}$$

while the desired markup can be written as follows:

$$\frac{\alpha p}{W y^{\frac{1}{\alpha}-1}} = \frac{\sigma\,(x)^{-\omega/\sigma}}{\sigma\,(x)^{-\omega/\sigma}-1} \equiv \mu(x), \tag{17}$$

where $\mu(x)$ denotes the markup, which increases in the demand-adjusted relative size of the firm $x = ay/Y$. From these two equations, we can derive a set of expressions linking the change in firm prices (and similarly markups, see Appendix A.3) to changes in aggregates Y and W, given by:

$$\frac{\partial \log p}{\partial \log Y} = \frac{\frac{1}{\alpha}-1}{1+\left(\frac{1}{\alpha}-1\right)\frac{\mu(x)}{\mu(x)-1}+\frac{\omega}{\sigma}\mu(x)}, \tag{18}$$

$$\frac{\partial \log p}{\partial \log W} = \frac{1}{1+\left(\frac{1}{\alpha}-1\right)\frac{\mu(x)}{\mu(x)-1}+\frac{\omega}{\sigma}\mu(x)}; \tag{19}$$

where the standard CES equivalent can be obtained setting the Kimball superelasticity parameter $\omega = 0$. Notice that all derivatives are positive. At the same time, the second derivatives with respect to the demand-adjusted relative size of the firm are given by:

$$\frac{\partial^2 \log p}{\partial \log Y \partial x} < 0 \quad \text{and} \quad \frac{\partial^2 \log p}{\partial \log W \partial x} < 0. \tag{20}$$

The signs of these second derivatives imply that prices (and hence markups) of firms that have a higher demand-adjusted relative size decline in Y or W. This is because these firms are those that face a lower demand elasticity and hence a more inelastic demand, which also implies a lower cost pass-through. As an example of this mechanism, one can think of the case of a drop in wages: facing lower marginal costs, firms with a more inelastic demand will pass less of this decline in marginal costs onto their prices, and hence will lower their prices by less. Taken together, Equations (18) and (19) show that, for given parameters, the demand-adjusted relative size of the firms is sufficient to characterize the heterogeneous response of prices and markups to aggregate shocks in the model. A more detailed explanation of the derivations can also be found in Appendix A.3, while, next, we assess how we can empirically proxy this sufficient statistic using firm-level data.

2.6.2 A New Empirical Proxy for Firms' Dominance

We start by discussing the challenges of empirically measuring the demand-adjusted relative size of the firms, ay/Y, and how to realistically proxy it using observables. Specifically, there are three reasons why a direct measurement of firm-level demand-adjusted relative size may fail in the data.

First, we do not observe the precise market in which firms compete, but at most the industry, which we know to be a measure of production proximity but has little (if anything) to do with markets. The empirical challenge to measure firm-level demand-adjusted relative size stems exactly from the fact that the size of a market Y should be theoretically the sum of all the output of the competitors, and hence should be defined based on the unobservable market in which competition takes place.[5] The empirical challenge of the unobservability of market structures in standard firm-level data, that is the boundaries within which competition takes place, for theoretical models of market power has been also noticed and discussed in Deb, Eeckhout, Patel, and Warren (2020).

Second, firm-level demand-adjusted relative size depends on firm-level idiosyncratic demand a. To calculate it in the data, we should hence be able to measure idiosyncratic demand differences across firms, which are however unobservable. Finally, the relative size of the firm ay/Y should be theoretically constructed using firm-level output quantities. This is a difficult empirical task, as prices are not easily available in the data, and hence we cannot recover quantities from firm sales.

All these challenges taken together suggest that the effort of directly measuring firm-level demand-adjusted relative size in the data may be fruitless. Hence, we propose an empirically measurable proxy intended to capture a similar variation across firms compared to the one captured by the demand-adjusted relative size: the *age* of the firms. We choose firms' age as a proxy of demand-adjusted relative size as it has some desirable properties, such as: (i) it captures similar economic forces to demand-adjusted relative size; (ii) it is an endogenous but pre-determined variable; and (iii) it is an important predictor of heterogeneous firm-level responsiveness to shocks.

One advantage of an age proxy is that it captures similar economic forces to the ones captured by the demand-adjusted relative size of the firms. In the model, the demand-adjusted relative size is a sufficient statistic to understand the heterogeneous response of the firms to aggregate shocks,

[5] Technically, this means that one should calculate $Y = \sum_i y_i$ as the sum of all the output of the firms in a given market where the competition takes place.

because it captures the relative dominance of a given firm in its market. Dominant firms in their market face a more inelastic demand and hence have a lower cost pass-through and responsiveness to shocks in general. Similarly, age may capture a firm's dominance in its given market, as it relates to the firm's ability to survive the competition, establish itself and find a niche of customers with an inelastic demand. Following this line of reasoning, older firms should be more likely to be well-established in their market, have more (inelastically) attached customers, and thus a more muted response to shocks. Notice that focusing on age as a proxy for firm dominance is also useful insofar as it avoids looking at measures of size. In the data, many small firms are old (Fort, Haltiwanger, Jarmin, and Miranda, 2013), which means that they are likely at their desired size. In this sense, using a size proxy instead of an age proxy, we may interpret those firms as unestablished ones.

Another advantage of an age proxy is rank invariance, as pointed by Cloyne, Ferreira, Froemel, and Surico (2018). Normally, other firm-level characteristics – such as their sales or financial position – endogenously respond to shocks and vary over the business cycle, which can affect the ranking of firms in the distribution of these variables. This is the reason why it is often hard to empirically interpret any (ex-post) heterogeneity in response to economic shocks as being driven exclusively (or even partially) by ex-ante differences in these specific firm characteristics. Note that, a firm's starting date is instead fully pre-determined to any subsequent aggregate shock. Since firm age cannot vary as a result of changes in monetary policy, it is less subject to these concerns.

A final advantage of an age proxy is that it builds on a growing literature that shows that age is indeed a critical firm-level characteristic to understand firms' trajectories over the cycle, as well as the heterogeneous responsiveness of firm-level investment to monetary policy shocks (Fort, Haltiwanger, Jarmin, and Miranda, 2013; Cloyne, Ferreira, Froemel, and Surico, 2018; and Colciago, Lindenthal, and Trigari, 2019). This suggests that firms' age differences can indeed capture fundamental economic forces that influence firm-level behavior in response to aggregate changes.

We conclude this section emphasizing a potential empirical challenge associated with using corporate age as a proxy for firm dominance, which could arise if age was to mainly capture a mechanism that is different from firm dominance but can nonetheless move markups as well. In this respect, the biggest concern is that there might be a correlation between firm age and its financial position. As highlighted by Cloyne, Ferreira, Froemel, and Surico (2018), young firms, especially those not paying dividends, are more likely to face financial constraints compared to older ones, as they are in the process of accumulating their net worth. At the same time, as discussed in Gilchrist, Schoenle, Sim, and Zakrajšek (2017) and Meinen and Soares (2022), firms more exposed to liquidity risks tend to raise markups in response to negative demand or financial shocks, while less exposed firms generally reduce them. If correct, this interpretation of the balance sheet channel that predicts more countercyclical markups for relatively younger firms would anyway dampen the significance of firm age as a proxy for dominance, and – if anything – will bias our estimates downwards. Yet, mindful of this identification challenge, Section 4.3.1 elaborates on our empirical approach to make the case that, thanks to the control variables in our set up, corporate age should indeed capture a firm's dominance in its market rather than any underlying financial frictions.

3 Data and Measurement

In this section, we first describe the dataset we work with and explain the construction of the main variables employed in our empirical analysis, focusing on those that are unrelated to firm-level markups. Second, we present and discuss the strategy we adopt to measure firm-level markups, emphasizing the critical challenges we face, and the virtues and drawbacks of the method we use.

3.1 Data

In our empirical analysis, we make use of firm-level data from the quarterly version of Compustat, which contains balance sheet information for North-American listed companies between 1975 and 2016. We nonetheless focus on the period comprised between 1990q1 and 2016q4, for which empirically identified monetary policy shocks series are also available. In what follows, we briefly review the strengths and limitations of this dataset. Note that we provide more details on the data cleaning process and the construction of the sample we use in the analysis in Appendix B.1.

The advantage of using Compustat is twofold: first, it provides high-frequency firm-level data, that is, quarterly data, which is crucial to study questions related to the business cycle, as in our case. Second, it provides detailed firm-level and sector-level information for a large number of firms and sectors in the US economy. Specifically, it contains information on firms' age and firm-level financial statements, including measures of sales, input expenditures, capital stock and liabilities, as well as a detailed industry activity classification. These characteristics make Compustat a potentially good source of firm-level data to study the response of markups to monetary policy shocks. Finally, it is important to stress that, although publicly traded firms are few relative to the total number of firms, they tend to be the largest firms in the economy, and account for roughly 30% of US employment (see Davis, Haltiwanger, Jarmin, Miranda, Foote, and Nagypal, 2006).

To measure firm-level production, we exploit information on firms' sales (`SALEQ`), while we use the cost of goods sold (`COGSQ`) to measure variable inputs used in production, and gross capital (`PPEGTQ`) to measure tangible capital. In line with the literature, we use selling, general and administrative expenses (`XSGAQ`) as a measure of overhead costs, whereas we take (`ATQ`) as a measure of total assets. Finally, we use (`CHEQ/ATQ`) as a measure of liquidity, and ($\texttt{DLCQ/ATQ} + \texttt{DLTTQ/ATQ}$) as a measure of leverage. Summary statistics for these variables are further reported in Appendix B.1.

In addition to that, we measure firm-level relative size as its sale share, by computing firm-level sales over total sales in a given sector. Note that we do so at the 4-digit NAICS level, which is the lower level of disaggregation allowed by the data. As a measure of firms' age, we use corporate age, which is readily available for all firms in the sample. However, whenever using this variable, we corroborate our results employing the "true" firm's age reported in Jay Ritter's database on the founding years of several US businesses, which is available for a subset of the firms in Compustat.

As monetary policy shocks, we exploit the series provided by Jarociński and Karadi (2020), which is constructed using interest rate surprises based on the percentage change in the FED Funds Futures rate in 30-minute windows around policy announcements. To take care of potential confounding effects stemming from the release of information around the time of announcements, the authors clean the series removing from the estimation any component attributed to the provision of FED information on the state of the economy to private agents through policy announcements.

Finally, after merging Compustat data with this series of monetary policy shocks, we complement our dataset with general indicators of economic activity at quarterly level. In particular, we include the GDP growth rate, the Consumer Price Index (CPI) growth rate, the Excess Bond Premium (EBP), and the 1-Year Treasury rate change, all taken from the Federal Reserve of St.Louis.

3.2 Markups Measurement

A key object in our analysis are firm-level markups, which measure the ability of firms to price above marginal costs. Measuring markups is often difficult, as neither prices nor marginal costs tend to be directly observable in the data. To overcome this challenge, we follow recent work by De Loecker and Warzynski (2012) and De Loecker, Eeckhout, and Unger (2020), which propose a method to measure markups based on the production function approach pioneered by Hall (1988).

Their estimation strategy is grounded on firm's optimizing behavior with respect to production costs-minimization, and delivers an estimate of markups at the firm-level without specifying an explicit demand system. For that, consider a firm i employing a production technology given by:

$$Q_{i,t} = F_{i,t}(\boldsymbol{X}_{i,t}, K_{i,t}, \omega_{i,t}), \tag{21}$$

where $\boldsymbol{X}$ is a vector of variable inputs, K is the predetermined input, and ω is firm-specific productivity. The cost minimization problem for each producer can hence be expressed as follows:

$$\min_{\{\boldsymbol{X}_{i,t}, K_{i,t}\}} \{\boldsymbol{P}'_{i,t}\boldsymbol{X}_{i,t} + R_t K_{i,t} + \lambda_{i,t}(Q_{i,t} - Q(\cdot)\}, \tag{22}$$

where $\boldsymbol{P}_{i,t}$ is the vector of prices for variable inputs, R_t is the price of the predetermined input, and $\lambda_{i,t}$ is the Lagrangian multiplier associated to the firm's cost minimization problem. One can then compute the first order condition (FOC) for a generic variable input $X^v \in \boldsymbol{X}$, which is given by:

$$\frac{\partial \mathcal{L}(\cdot)}{\partial X^v_{i,t}} = P^v_{i,t} - \lambda_{i,t} \frac{\partial Q(\cdot)}{\partial X^v_{i,t}} = 0. \tag{23}$$

Notice that the Lagrangian multiplier $\lambda_{i,t}$ can be also interpreted as the marginal cost of producing at a given level of output. Moving one step forward, Equation (23) can be further rearranged as:

$$\frac{\partial Q(\cdot)}{\partial X^v_{i,t}} \frac{X^v_{i,t}}{Q_{i,t}} = \frac{1}{\lambda_{i,t}} \frac{P^v_{i,t} X^v_{i,t}}{Q_{i,t}}. \tag{24}$$

Finally, defining the markup as price over marginal costs, $\mu_{i,t} \equiv \frac{P_{i,t}}{\lambda_{i,t}}$, it is possible to rearrange the FOC expressed in Equation (23) for a generic variable input $X^v \in \boldsymbol{X}$, so that it yields:

$$\mu_{i,t} = \theta^v_{s,t} \frac{P_{i,t} Q_{i,t}}{P^v_{i,t} X^v_{i,t}}, \tag{25}$$

where $\theta^v_{s,t}$ is the elasticity of output with respect to the variable input X^v. Therefore, to measure firm-level markups one just needs firms' sales and variable input expenditure, which are readily available from the data, and the elasticity of output with respect to the variable input, which requires the estimation of the firm-level production function. As pointed out by Bond, Hashemi, Kaplan, and Zoch (2021), the precise estimation of firm-level markups depends crucially on the correct identification of the elasticity of output with respect to the variable input. For our analysis, and to increase comparability with previous estimates, we use as a benchmark the sector-level elasticities estimated at yearly level by De Loecker, Eeckhout, and Unger (2020). However, we double-check our results with quarterly level estimates under different production function specifications. As a final remark, we emphasize that markups changes – the focus of our analysis – are correctly identified even if the elasticity is not, as shown in De Ridder, Grassi, and Morzenti (2021).

4 Empirical Analysis

In this section, we test the predictions of the theoretical model by empirically studying the heterogeneous cyclicality of firm-level markups conditional on MP shocks. We begin by presenting the empirical strategy and our main results. Then, we discuss the robustness of our findings.

4.1 Empirical Specification

In order to investigate any cross-sectional difference in the relationship between monetary policy and firm age, we employ a panel version of the local projections (LP) method proposed by Jordà (2005). In practice, we estimate by ordinary least squares (OLS) the following set of equations:

$$\begin{aligned}\Delta_h \log \mu_{i,t+h} = \textstyle\sum_{x\in\mathcal{X}} \left(\alpha_{x,h} + \beta_{x,h}\Delta Y_{t-1} + \sum_{k=-\kappa}^{h} \gamma_{x,h}^{k}\varepsilon_{t+k}^{m}\right) \times \mathbb{1}_{i\in\mathcal{I}^x} \\ + \textstyle\sum_{\ell=1}^{L} \boldsymbol{\delta}_h' \boldsymbol{X}_{i,t-\ell} + \varphi_{i,h} + \varphi_{s,t,h} + \boldsymbol{\vartheta}_h \boldsymbol{t} + u_{i,t+h},\end{aligned} \tag{26}$$

with horizons given by $h = 0, 1, \dots, H$. The dependent variable on the left-hand side is the cumulative change in markups for any firm i at horizon h, and is defined by the following expression:

$$\Delta_h \log \mu_{i,t+h} \equiv \log \mu_{i,t+h} - \log \mu_{i,t-1}. \tag{27}$$

We include the interaction between the MP shock ε_t^m from Jarociński and Karadi (2020) and $\mathbb{1}_{i\in\mathcal{I}^x}$, which is an indicator that takes a value of 1 if $i \in \mathcal{I}^x$, namely if firm i is above the *median* for a given set of variables and during the previous year. In the baseline specification, we assume that $\mathcal{X} = \{age\}$ and do not allow for additional covariates, only comparing firms above and below median age. Our main coefficient of interest is hence $\gamma_{age,h}^0$, which captures the relative response of old companies (compared to young ones) to a variation in the monetary policy rate. As a benchmark measure, we use corporate age, which is available for the entire sample. However, our robustness analysis shows that using the "true" founding age for the firms for which this information is available does not alter the results. Moreover, in Section 4.3.1, we not only account for heterogeneity in age, but also horse race age with other equally relevant and competing covariates that are often associated with heterogeneous firm performance (e.g. indicators of potential financial frictions). These factors are summarized in the vector $\mathcal{X} = \{age, sales\ shares, leverage, liquidity, assets\}$. [6]

In our preferred specification, we adopt a non-parametric estimation approach by employing dummies instead of linear interactions, which is line with the strategy in Cloyne, Ferreira, Froemel, and Surico (2018). We show later in our robustness checks that results do not qualitatively change if we use a parametric specification with a linear interaction term. We also control for past and future monetary policy shocks, which is done for two reasons. First, to control for potential serial correlation in the identified MP shock. Second, to control for the attenuation bias likely present in our LP model, due to the short time horizon of our panel dataset (Teulings and Zubanov, 2014).

[6] Another natural dimension to study would be heterogeneity in the degree of price stickiness. However, existing measures of this are at the sector level and would be absorbed by the presence of sector time fixed effects in our analysis.

Following Ottonello and Winberry (2020), we also interact $\mathbb{1}_{i \in \mathcal{I}^x}$ with ΔY_{t-1}, which is GDP growth over the previous quarter, to control for potentially different sensitivities of firm markups to the business cycle. Moreover, the vector of controls $\boldsymbol{X_{i,t}}$ includes both firm-level variables, such as sales growth and overhead costs to sales, and macro-level controls like GDP and CPI growth, 1-year treasury rate change, and the EBP. It also contains fiscal quarter dummies to account for seasonality coming from accounting practices, as argued in Ottonello and Winberry (2020). Following standard strategies in the literature, we include lagged MP shocks and control variables ($\kappa, L = 4$).

We allow for firm $\varphi_{i,h}$ and sector-time $\varphi_{s,t,h}$ fixed effects (FE) to control for any unobserved time-invariant heterogeneity at the firm-level, and to absorb time-varying shocks that are common to all firms in a given industry. Saturating the regression in Equation (26) with these FE implies that, first, our coefficients of interest are identified by within-firm variation over time, namely by changes in the markup response of an otherwise identical firm when it is old compared to when it was young. Secondly, our estimation fully exploits the cross-sectional variation across firms in a given industry. We also include a linear and quadratic trend $\boldsymbol{\vartheta_h t}$ to flexibly account for the growth of markups documented by De Loecker, Eeckhout, and Unger (2020). Finally, we cluster the standard errors $u_{i,t}$ at the firm and quarter level to account for correlation in the error term.[7]

The adjustment of relatively older firms estimated through Equation (26) does not allow us to understand the overall individual response of both old and young firms. In fact, Equation (26) is saturated with sector-time fixed effects, which span out time-series variation common across firms. Hence, we additionally estimate the following separate model for firms in different age categories:

$$\Delta_h \log \mu_{i,t+h} = \sum_{k=-\kappa}^{h} \gamma_h^k \varepsilon_{t+k}^m + \sum_{\ell=1}^{L} \boldsymbol{\delta_h' X_{i,t-\ell}} + \varphi_{i,h} + \boldsymbol{\vartheta_h t} + u_{i,t+h}, \tag{28}$$

with horizons given by $h = 0, 1, \dots, H$. Note that the dependent variable is still the cumulative change in markups for any firm i at horizon h, which has been previously defined in Equation (27).

In this second specification, we therefore exploit time-variation and look at the absolute change in markups after a change in the identified MP shock for firms of different age categories. The coefficient of interest γ_h is therefore estimated for firms in each age group separately. Also, note that ε_t^m is still the series of MP shocks from Jarociński and Karadi (2020). Moreover, $\boldsymbol{X_{i,t}}$ is a vector of controls that include: (i) firm-level variables, such as sales growth, overhead costs to sales, leverage, liquidity and assets; and (ii) macro-level controls, such as GDP growth, CPI growth, 1-year treasury rate change, the EBP, and fiscal quarter dummies. We again include lagged MP

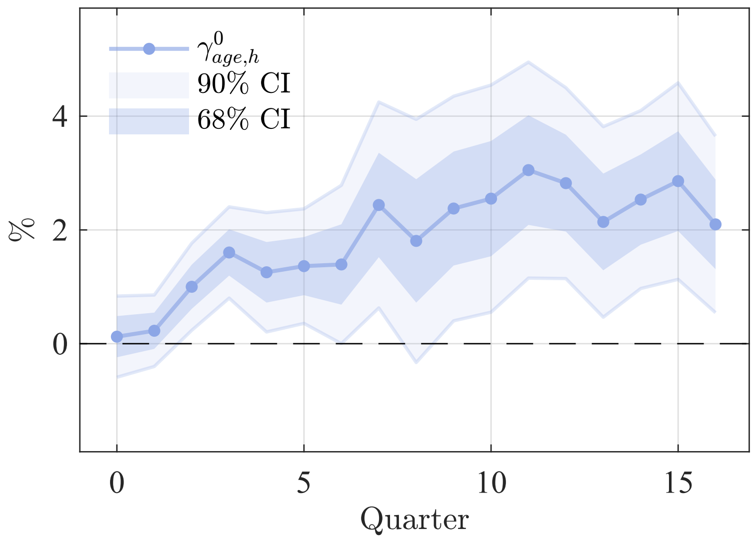
[7]Clustering at the firm level allows for a fully flexible dependence in the error terms across time within each company. Clustering by time is necessary whenever firm-level shocks are correlated within a quarter and if this effect may go potentially above the co-movement caused by industry-level shocks already captured by the sector-quarter dummies. We note that the confidence intervals on estimates would be significantly lower without clustering at the quarter level.

shocks and control variables (κ, $L = 4$). We also allow for firm FE $\varphi_{i,h}$ to account for time-invariant firm-heterogeneity, and add a linear and quadratic trend $\boldsymbol{\vartheta_h t}$ to flexibly account for the growth of markups documented by De Loecker, Eeckhout, and Unger (2020). Finally, we keep clustering our robust standard errors $u_{i,t}$ at the firm and quarter level to account for correlation in the error term.

4.2 Results

Figure 2 shows the impulse response function (IRF) obtained from the OLS estimation of $\gamma_{age,h}$ in Equation (26), along with confidence intervals around the point estimates. The coefficient associated with firm age is statistically significant and precisely estimated. This finding supports the argument developed in the theoretical section and motivates the focus of our analysis on firms' age as an important predictor of firm-level markups' heterogeneous response to MP shocks.

Figure 2: Relative Response of Old Firms' Markups to a Monetary Policy Shock

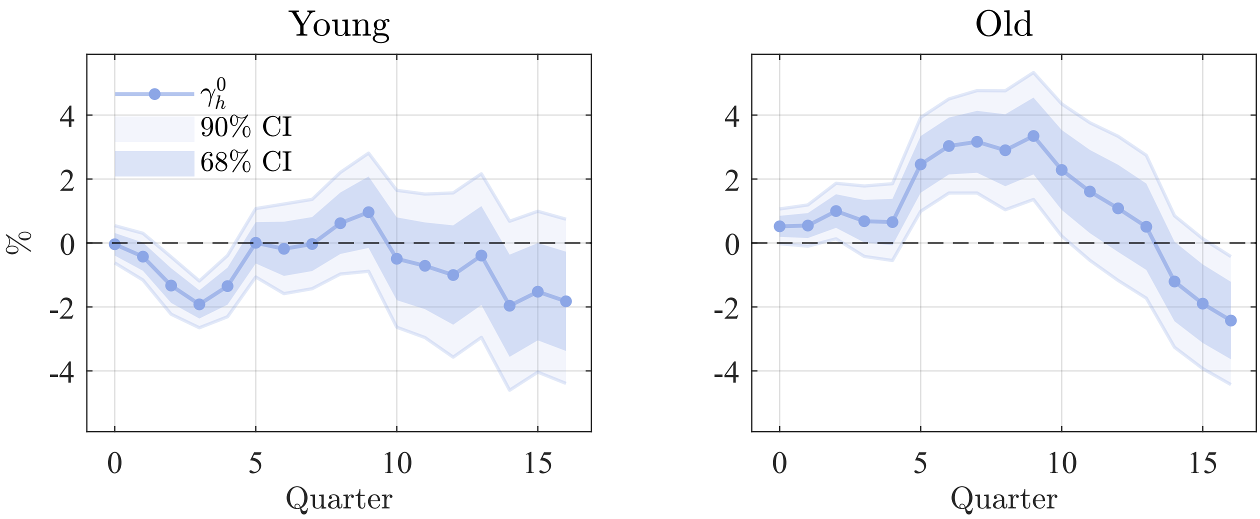


Note. Figure 2 shows the relative response of old firms' markups to a MP shock (compared to young firms). It reports the evolution of the coefficient $\gamma^0_{age,h}$ from the estimation of Equation (26) for $h = 1, ..., 16$. The figure is normalized to a 25 basis points contractionary MP shock. The solid dark blue line with circles reports the point estimates of $\gamma^0_{age,h}$. The dark and light blue areas report the 68% and the 90% confidence intervals around our estimates.

The magnitude of the $\gamma_{age,h}$ coefficient suggests that being above the median age – which amounts to being an old firm according to our definition – before a contractionary MP shock of 25 basis points implies approximately a 2% statistically significant difference in the subsequent response of markups. Put differently, old firms' markups shows an excess countercyclicality in response to MP shocks, in line with the theoretical predictions we have outlined in Section 2.

In order to understand the absolute response of markups at the firm-level, we estimate Equation (28) within different age groups, namely for firms above and below median age. Figure 3 reports the resulting IRFs and suggests that old firms increase markups by up to 3% after a 25 basis points contractionary MP shock, while young firms reduce them by 2%. This implies that firms of different age groups have qualitatively different responses. In particular, old firms have countercyclical markups conditional on a demand shock, while young firms have procyclical ones.

Figure 3: Markups Response to a Monetary Policy Shock by Firm Age Group



Note. Figure 3 shows the response of firms below and above median age to a 25 basis points contractionary MP shock. It reports the evolution of the coefficient γ_h^0 from the estimation of Equation (28) for $h = 1, ..., 16$. The solid dark blue line with circles reports the point estimates of γ_h^0. The dark and light blue areas report the 68% and the 90% confidence intervals around our estimates.

Taking stock, we find that old and young firms' markups respond differently to MP shocks, with old firms having countercyclical markups and young firms having procyclical markups conditional to a demand shock. We believe that the significance of firms' age in our analysis supports the reasoning presented in the theoretical section, and strengthen the case of using age as a useful predetermined variable to capture how dominant a firm is in its (unobservable) product market.

Finally, we conclude with a brief note about the implication of our findings for the cyclicality of the aggregate markup. Taking as a benchmark the observation that the aggregate markup is the cost-weighted average of firm-level markups, as demonstrated by Grassi (2017) and Edmond, Midrigan, and Xu (2018), we notice that in our dataset old firms incur 75% of total costs. Therefore, under the admittedly strong assumption that the cost shares of the different age groups are approximately stable conditional on demand shocks, this suggests a mildly countercyclical response of the aggregate markup to MP shocks, which should be driven by the response of old firms in the data.

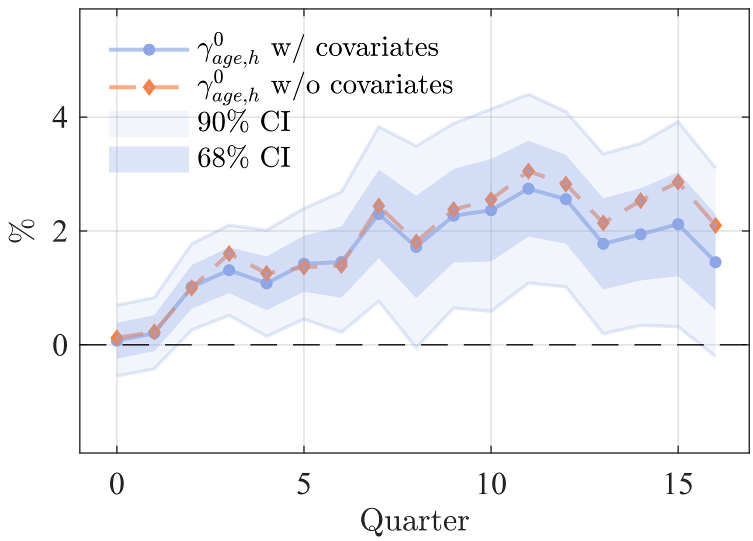
4.3 Accounting for Financial Frictions and Other Robustness Exercises

This section presents additional exercises aimed at validating the robustness of our findings. We delve into the role played by financial frictions compared to firms' market dominance in explaining markups' heterogeneous responses, and then follow with a battery of general robustness exercises.

4.3.1 Financial Frictions

In this section, we continue the discussion of Section 2.6.2 and assess if and how any potential correlation between firms' age and their underlying financial frictions could impact our findings. To this end, we take the model estimated in Equation (26) and augment it to include as regressors other covariates commonly employed by the literature to measure financial frictions. Specifically, we conduct a horse race between firm age – our proxy variable for firms' dominance – and assets, sales shares, leverage and liquidity. Moreover, we draw on recent insights by Cloyne, Ferreira, Froemel, and Surico (2018) to further confirm that financial frictions are unlikely to be the primary driver of the heterogeneous response of markups to MP shocks across old and young firms.

Figure 4: Old Firms' Markups Response to a Monetary Policy Shock, with and without Covariates



Note. Figure 4 shows the relative response of old firms' markups to a MP shock (compared to young firms) with and without additional covariates. It reports the evolution of the coefficient $\gamma^0_{age,h}$ from the estimation of Equation (26) for $h = 1, ..., 16$. The figure is normalized to a 25 basis points contractionary MP shock. The solid dark blue line with circles reports the point estimates of $\gamma^0_{age,h}$ with additional covariates. The dashed orange line with diamonds reports the point estimates of $\gamma^0_{age,h}$ without additional covariates, that is, the point estimates of the baseline specification from Figure 2. The dark and light blue areas report the 68% and the 90% confidence intervals around our estimates with covariates.

Figure 4 shows the IRF obtained from the estimation of $\gamma_{age,h}$ in this augmented version of Equation (26). The solid dark blue line with circles represents the IRF with additional covariates (summarized by the vector $\mathcal{X} = \{age, \textit{sales shares}, \textit{leverage}, \textit{liquidity}, \textit{assets}\}$), while the dashed orange line with diamonds depicts the IRF without covariates from the baseline specification presented in Section 4.2. Confidence intervals around the point estimates are also provided. This alternative regression with covariates entails a horse race between our primary regressor, corporate age, and other dimensions of firm heterogeneity that might be crucial explanatory factors for the cross-sectional response of markups, including sales shares, leverage, liquidity, and assets.

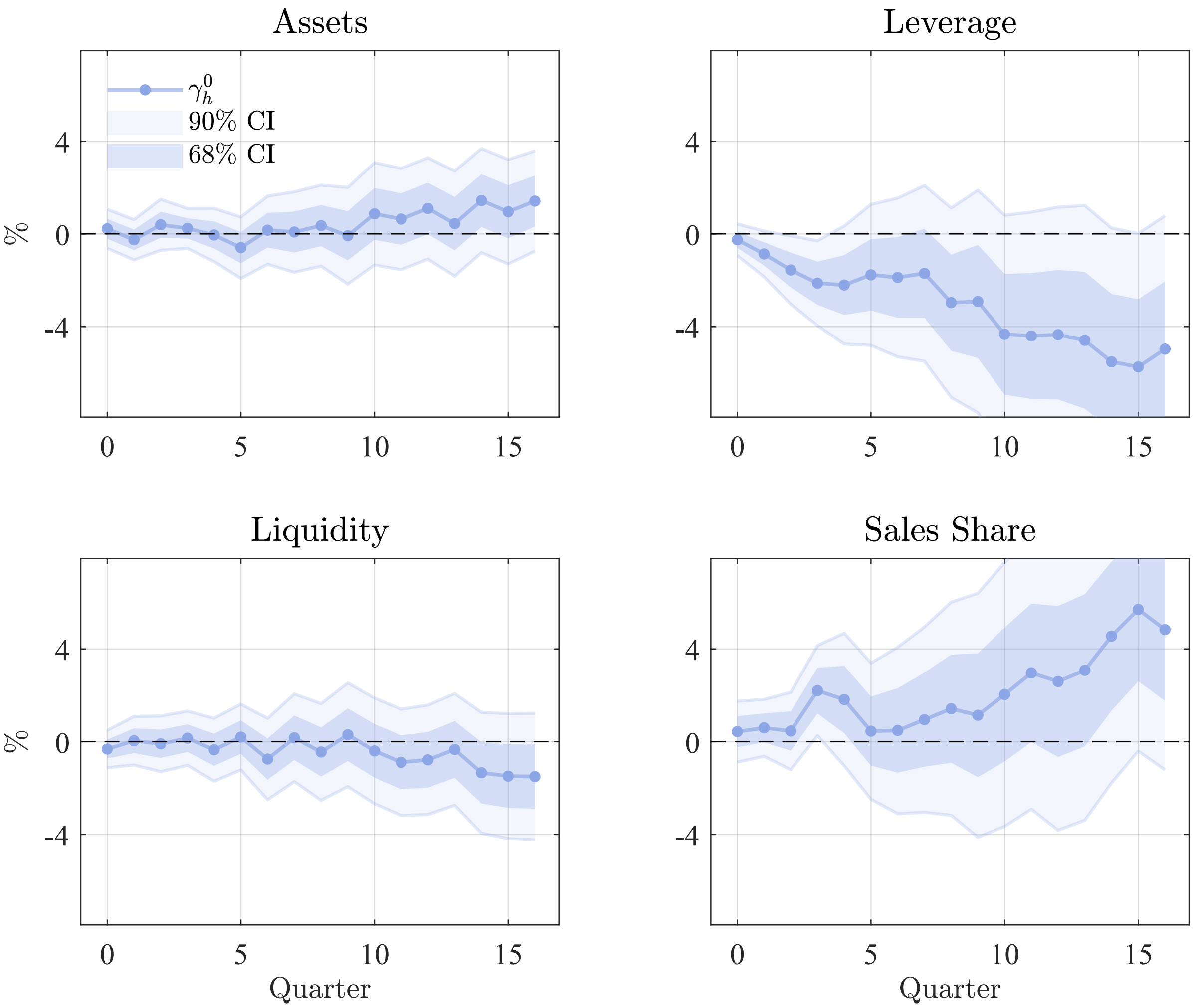
Sales share could be an alternative variable to age for the sake of proxying demand-adjusted relative size. Simultaneously, leverage, liquidity, and assets, are heavily popularized proxies for financial frictions influencing firms' responses to MP shocks (Ottonello and Winberry, 2020; Jeenas, 2019; Fabiani, Falasconi, and Heineken, 2020). The key insight from Figure 4 is that firm age seems to influence the heterogeneous markup response beyond its correlation with firms' financial positions. The robustness of our point estimates in the face of controlling for proxies of financial frictions supports this interpretation. If firm age was to merely mirror financial conditions, we would have observed a substantial impact on our estimates when controlling for these proxies.

While age emerges from our analysis as an important predictor of heterogeneity in the response of markups to monetary policy – extending beyond the influence of financial frictions –, it does not exclude that financial frictions may also contribute to shaping firms' markup responses. Figure B.1 shows the relative response of markups to a MP shock for firms above the median in previous year's assets, leverage, liquidity, sales shares, and reports the evolution of the coefficient $\gamma^0_{x,h}$ for $x \in \mathcal{X} = \{\textit{sales share}, \textit{leverage}, \textit{liquidity}, \textit{assets}\}$ from the estimation of Equation (26) for $h = 1, ..., 16$.

While the precision of these interaction estimates varies, it's noteworthy that two of them—leverage and sales share—exhibit statistically significant deviations from zero in specific quarters. First, the fact that more leveraged firms keep markups relatively higher upon the arrival of a contractionary MP shock is consistent with Gilchrist, Schoenle, Sim, and Zakrajšek (2017), who point out that financial distortions create an incentive for firms to preserve internal liquidity and raise prices in response to adverse financial or demand shocks. Second, the fact that sales shares in an industry are less predictive than age (though responding in the same direction) when it comes to the heterogeneous responses of firms' markups to MP shocks suggests that, at least in our data, sales shares are less indicative of a firm's dominance than firm age. This observation can relate to the fact that sales shares are at the industry level, and an industry rarely if never coincides with the actual boundaries of the market where a firm is active. This further supports the suitability of age as a better proxy for firm dominance. We then conclude that both leverage, reflecting financial frictions, and sales shares, representing a firm's relative size in an industry, contribute to the heterogeneous responses of firms' markups to MP shocks, although to a lesser degree than age.

While this exercise confirms the significance of age as a proxy for firms' dominance, in Appendix B.2 we perform an additional test following the logic outlined in Cloyne, Ferreira, Froemel, and Surico (2018). They posit that firms likely to be financially constrained are young ones that do not pay dividends. This suggests that splitting young firms into two groups – those that pay dividends and those that do not – could help further disentangle the role of financial frictions and that of firm dominance when using age. However, this exercise imposes a burden on the data, as dividend payments are sparse in Compustat (a limitation noted in Cloyne, Ferreira, Froemel, and

Figure 5: Relative Response of Markups to a MP Shock for Firms of Different Categories



Note. Figure B.1 shows the relative response of the markups to a monetary policy shock for firms above median in assets, leverage, liquidity, sales share. In particular, it reports the evolution of the coefficient $\gamma^0_{x,h}$ for $x \in \mathcal{X} = \{sales\ share, leverage, liquidity, assets\}$ from the estimation of Equation (26) for $h = 1, ..., 16$. The figure is normalized to a 25 basis points contractionary monetary policy shock. The solid dark blue line with circles reports the point estimates of $\gamma^0_{x,h}$. The dark blue and light blue areas report the 68% and the 90% confidence intervals of our estimates.

Surico, 2018). Despite substantially noisier estimates, results in Appendix B.2 confirm the validity of age as a proxy for firms' dominance, while indicating some role for financial frictions as well.

This discussion serves to confirm that firm age is a valuable proxy for firms' dominance and has significant and sizeable implications for the heterogeneous responses of firms' markups to monetary policy. Moreover, it also establishes that age is not the sole factor contributing to the heterogeneity of this response; financial frictions and firms' relative size within a broad industry also play a role. However, we observe that the role of age goes above and beyond these channels and, at least in our data, appears to be quantitatively stronger. Hence, in the remainder of the paper, we focus on understanding its aggregate implications for monetary policy in isolation.

4.3.2 Robustness

In what follows, we discuss the robustness of our results regarding the heterogeneous response of markups by firms of different age groups conditional on MP shocks. In particular, we consider the following alternative specifications of Equation (26): (i) excluding the zero lower bound (ZLB) period; (ii) excluding future MP shocks; (iii) adopting a linear parametric interaction; (iv) using the MP shock series from Gürkaynak, Sack, and Swanson (2005); (v) generating firm age groups by sector and quarter; (vi) using the founding age by Jay Ritter; (vii) using alternative elasticities; (viii) using an alternative markup measure. We focus our discussion on the robustness of Equation (26) because it is our main result; however, all these checks go through for Equation (28) as well.

(i) Excluding the ZLB. Under this alternative specification, we estimate Equation (26) excluding the ZLB period, thereby focusing on the years between 1990q1 and 2008q4. We do so because during the ZLB period the interest rate was stuck at zero, and there is virtually no variation in the MP shock measure used in our analysis. In so doing, we want to ensure that our results are not driven by this period, when conventional monetary policy was not in place. Results are reported in Figure B.2 in Appendix B.3 and show that excluding the ZLB period does not affect our conclusions.

(ii) Excluding Future MP Shocks. In this case, we estimate Equation (26) without including future MP shocks as controls. Results are shown in Figure B.3 in Appendix B.3 and highlight that excluding future MP shocks does not qualitatively affect our results, although it reduces their magnitude. This comes with no surprise, given that future MP shocks were introduced, as explained above, to address the attenuation bias in LP models based on panel datasets with a short horizon span.

(iii) Linear Parametric Interaction. Here, we estimate Equation (26) interacting the MP shock with the age variable instead of with an age group dummy, thereby using a more parametric approach compared to our benchmark specification, as done in Ottonello and Winberry (2020). Results are reported in Figure B.4 in Appendix B.3, and do not pose a challenge to our overall conclusions.

(iv) MP Shocks from Gürkaynak, Sack, and Swanson (2005). We also estimate Equation (26) using the identified MP shocks series from Gürkaynak, Sack, and Swanson (2005). This shock series is similar to the one by Jarociński and Karadi (2020), which we use as a benchmark, but does not net out from the series the information channel of monetary policy. Results are shown in Figure B.5 in Appendix B.3, and highlight that using this alternative series does not affect our conclusions.

(v) Grouping Firms by Sector and Quarter. Here, we estimate Equation (26) using an alternative definition of the variable $\mathbb{1}_{i \in \mathcal{I}^x}$. In particular, instead of defining firms' categories by being above or below the median of variables contained in $\mathcal{X} = \{age, sales\ shares, leverage, liquidity, assets\}$ and over the entire sample – as done in the main analysis –, we define firms' groups by being above or below the median in a given sector and quarter. Results are shown in Figure B.6 in Appendix B.3.

Overall, we see that using this alternative definition of the dummies does not alter our conclusions.

(vi) Using Firm Founding Age by Jay Ritter. In this robustness check, we estimate Equation (26) using the true founding age of firms instead of corporate age. Data on firms' founding age is made available by Jay Ritter for a subset of firms in Compustat. Results are shown in Figure B.7 in Appendix B.3. Overall, using this alternative measure of firms' age does not alter our conclusions.

(vii) Using Alternative Elasticities. Here, we estimate Equation (26) using markups calculated with different production function elasticities. In particular, instead of using the elasticities provided by De Loecker, Eeckhout, and Unger (2020), which are common for all the quarters within a year, we calculate ourselves the production function elasticities allowing them to vary at a quarterly level. In particular, we estimate (i) a Cobb-Douglas production function which varies at quarter and sector levels, and (ii) a Translog production function that also varies at quarterly and sector levels. This second specification allows us to have production function elasticities that are heterogeneous across firms within sectors and quarters. Results are reported in Figure B.8 in Appendix B.3: using alternative production function elasticities does not challenge our main conclusions.

(viii) Using an Alternative Markup Measure. In this last robustness check, we adopt a different methodology to calculate markups at the firm-level, and use this alternative markup measure to estimate Equation (26). In particular, we follow the strategy proposed by Gutiérrez and Philippon (2016) and Baqaee and Farhi (2020), which we describe in full detail in Appendix B.3. Results are shown in Figure B.9 in Appendix B.3. Overall, using this alternative measure of market power does not affect the qualitative nature of our conclusions, although it dampens slightly the magnitude of the main coefficient of interest compared to our benchmark estimates from previous paragraphs.

5 Quantitative Analysis

In what follows, we proceed to explain the quantification of the model, including the calibration strategy and the overall fit of both targeted and untargeted moments computed from US data. In particular, we discuss the ability of our framework to replicate salient features of the markups and firms' distribution, which is a crucial property needed to provide a link with the empirical analysis of the previous section. Once quantified, the model is then used in Section 6 to study and analytically decompose the impulse response functions of firms' markups after a negative monetary policy shock. Moreover, Section 6 also compares the amplification mechanism at work in our framework to the one implied by a standard representative-firm New Keynesian model.

5.1 Calibration

The full list of both fixed and fitted parameter, as well as the targeted moments, is presented in Table 1. A model period in one quarter. Of the 14 parameters we need to calibrate, 8 are fixed outside of the model, for which we pick common values used in the literature. In particular, we set the risk aversion $\nu = 2$ and the disutility of labor $\gamma = 2$, while the discount factor $\rho = 0.012$ is specified to deliver a yearly interest rate of 5% in equilibrium. With respect to the parameters related to firms' life-cycle, technology and pricing behavior, we fix the quarter exit rate $\delta = 0.024$ to imply that 10% of the firms exit each year, and the returns to scale $\alpha = 0.33$ so that the labor share is around 0.6 in equilibrium. Moreover, we normalize to 1 the mean demand $\bar{a}_0$ faced by entrant intermediate firms, while the demand dispersion ψ_0^a at entry is set to be equal to the dispersion of the demand process faced by incumbents.[8] Finally, we follow similar strategies as in Taylor (1999) and Galí (2015) to set the monetary policy coefficient $\phi_\pi = 1.5$ in the simple interest rate rule.

Table 1: Estimated Parameters and Targeted Moments

Fixed	Value	Description			
ρ	0.012	Discount factor			
ν	1	Risk aversion			
γ	2	Inverse Frisch elasticity			
α	0.33	Production function curvature			
δ	0.024	Exit rate			
$\bar{a}_0$	0	Mean demand entrants			
ψ_0^a	0.11	Demand dispersion entrants			
ϕ_π	1.5	Taylor rule coefficient			
Fitted	**Value**	**Description**	**Moments**	**Model**	**Data**
θ	20	Price adjustment cost	Avg. cost change prices over sales	0.11	0.09
σ	4	Elasticity of demand	Avg. markup	1.68	1.68
ω	5.1	Superelasticity of demand	Elasticity markups to sale shares	0.11	0.10
$\bar{a}$	2	Mean demand	Median markup	1.37	1.30
ψ^a	0.11	Demand dispersion	Markups standard deviation	1.23	1.22
ρ_a	0.02	Demand mean reversion	Markups growth between age 0-5	0.24	0.22

Note: Empirical estimates for fitted parameters from Compustat data (1990q1-2016q4). For fixed parameters, see text.

In addition to that, we need to endogenously assign values to the remaining 6 parameters, for which we match as many salient moments computed from US data. To begin with, we set the price adjustment cost factor $\theta = 20$ such that the average ratio between the cost paid by firms to change

[8] Our results do not depend on this choice, which is just a simplification for the sake of the estimation procedure.

prices and their sales is the same in the model and in the data.[9] As standard in the literature, we set the elasticity of demand $\sigma = 4$ to match an average markup of 1.68 computed in the sample of Compustat firms:[10] this parameter determines the level of substitutability across the output of different producers in the model, and hence influences the average market power in the economy. Moreover, the superelasticity of demand ω is fitted such that the elasticity of markups to sale shares in the model is the same as in the data. Our choice is motivated by the fact that the parameter ω in the Kimball aggregator is tightly linked to the relationship between the relative size of firms and their markups. Specificallyt, if ω was 0, such relationship would be null because all firms would have the same markup independently of their size. On the contrary, for $\omega > 0$, the higher the ω the higher the dependence of markups on sales shares. To this end, using Compustat firm-level data, we empirically estimate the elasticity of (log) markups to (log) sales shares according to:

$$\log \mu_{i,t} = \beta * \log(\textit{sales shares})_{i,t} + \varphi_{s,t} + \varepsilon_{i,t} \tag{29}$$

where $\varphi_{s,t}$ are sector-time FE and the coefficient β precisely informs by how much markups are linked to firms' sales shares. In mapping this regression specification to the quantitative model, we use the theoretical definitions of markups and sales shares discussed in previous sections.

Finally, turning to the parameters related to the demand accumulation process, the mean demand is set to match the median markup in the US economy, as $\bar{a}$ identifies the distance between the average demand faced by entrants and incumbents, and hence relates to the skewness of the markup distribution. Furthermore, the dispersion in the demand process faced by incumbent firms ψ^a is identified from the standard deviation of markups, while the mean reversion in the demand process ρ_a is picked to match the growth of markups for firms between age 0 to 5. This choice is motivated by the fact that a higher mean reversion in the demand accumulation process impacts how fast firms grow, and therefore relates to the trajectory of markups over the firm's life-cycle.

5.2 Quantitative Fit

In what follows, we present and discuss our main validation exercises, which provide an overview of the quantitative fit of our framework with respect to specific features of the data that have not been targeted in the calibration. In particular, we first discuss the cross-sectional and life-cycle characteristics of firms in our model, and how they compare to their empirical counterparts from Compustat. Secondly, we dig into the properties of the markup distribution and analyze markups

[9]Proposed estimates for the adjustment cost factor (θ in our model) vary between 0.04 for physical costs and 0.09 for customer costs, see for example Levy, Bergen, Dutta, and Venable (1997) and Zbaracki, Ritson, Levy, Dutta, and Bergen (2004). As in Golosov and Lucas (2007) and Baley and Blanco (2019), we choose a value within this range of estimates.

[10]We use Compustat Data between 1990q1 and 2016q4. For the empirical definition of markups, see Section 4.

dynamics over firms' life-cycle. Finally, we conclude with a note on the model and data-implied elasticity of wages to sales, and relate it to the behavior of markups under the Kimball aggregator case and in imperfect competition, following similar lines as in Edmond, Midrigan, and Xu (2018).

5.2.1 Implications for Markups in Steady State

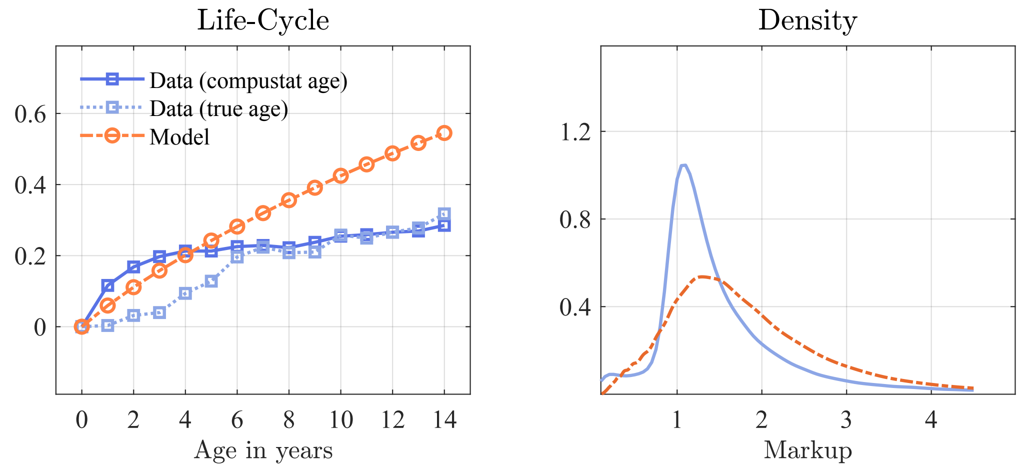
One of the key validation exercises we perform is to look at the properties of markups in the data and compare them with the ones implied by our quantitative framework. In particular, there is rich evidence showing that markups increase with firms' age in Compustat (as well as in other datasets). For instance, Alati (2020) argues that such behavior may be due to the fact that, as businesses advance along their life-cycle, they are also able to establish their position in their respective markets and progressively accumulate demand for their products. In turn, this allows producers to progressively charge higher prices and set higher markups. To illustrate how this mechanism plays out in our framework, the left panel in Figure 6 reports the pattern of markups over firms' life-cycle both in the model and in the data. Note that, to ensure comparability, the empirical series is computed using Compustat data between 1990q1 and 2016q4 netting out sector and time FE.

On the one hand, the model underestimates the rapid increase of markups in the first 5 years of a firm's life, whereas it tends to modestly overestimate their subsequent growth in the next years.[11] On the other hand, our calibrated framework can qualitatively replicate the growth of markups over firm age, and matches more than half of the quantitative features of the relationship between markups and the life-cycle of producers. Importantly, it needs to be stressed that the ability of the model to imply life-cycle markups' properties consistent with the empirical observations will prove crucial when assessing the differential response of firms to MP shocks. In fact, as documented in Section 4, old firms' markups show a more countercyclical response after a negative MP shock: absent the fit of the life-cycle profile of markups, our model would then not be able to replicate the heterogeneous response of markups to a MP shock according to firms' relative age.

Secondly, as illustrated in the right panel of Figure 6, the model reproduces reasonably well the distribution of markups estimated from Compustat. While the mean, median and standard deviation of markups are targeted in the calibration, the model itself delivers a fat right tail in the distribution of markups consistent with our empirical observation and with the analysis of De Loecker, Eeckhout, and Unger (2020). As reported in Table 2, our quantitative framework implies that the bottom 25% firms in the distribution have an average markup of 1.15, against a value of 1.03 computed in the data. Moreover, a similar fit holds for the top 75% firms as well. Matching the correct shares of high and low-markup firms' will prove crucial when comparing the

[11] The fit is very precise during the first years of business operations which is due to the fact that, in our calibration, we target the mean reversion in the demand process ρ_a to match the growth of markups for firms aged 0 to 5.

Figure 6: Markups Steady State Properties



response of markups to a MP shock across firms that are below and above the median age.

Table 2: Distributional Properties of Markups

	Model	Data
Bottom 25% Firms	1.15	1.03
Top 75% Firms	1.79	1.86

5.2.2 The Link between Wages and Sales

In quantifying the model, we have calibrated the superelasticity ω in the Kimball aggregator to match the elasticity of markups to sales shares. Our framework has also a testable prediction with respect to the link between the wage bill and the sales of firms, which we can match as an untargeted dimension. In particular, recall that markups are a measure of whether firms can set prices above their marginal costs. Similarly to Edmond, Midrigan, and Xu (2018), in our theoretical setup the salaries paid by firm i depend on its sales and markup according to the following expression:

$$\text{wage bill} \propto \frac{\text{sales}}{\text{markup}}$$

If the superelasticity ω in the Kimball aggregator was equal to zero, as in a standard NK model, markups would not increase with firm sales and, in turn, labor income shares would increase one-for-one with sales shares. But when ω is strictly positive, as in our framework, markups do increase with firm sales, implying that the wage bill increases less than one-for-one with firm sales. In this sense, both empirically and quantitatively, the extent to which the labor income share of firms increases with their sales shares can therefore be linked to the extent to which markups increase with producers' size. A small caveat to keep in mind is that Compustat does not report a precise measure for firms' wage bills but only a balance sheet item related to the cost of goods sold. This variable comprises the cost of all variable inputs used in production, including (but not exclusively) labor. Nevertheless, we exploit the available data to run the following regression:

$$\log(\textit{labor income shares})_{i,t} = \beta * \log(\textit{sales shares})_{i,t} + \varphi_{s,t} + \varepsilon_{i,t} \tag{30}$$

where $\varphi_{s,t}$ are sector-time FE and the coefficient β precisely informs by how much variable input costs are linked to firm sales shares. The results of the empirical estimation and quantitative fit are reported in Table 3. Importantly, a value of the elasticity $\beta < 1$ confirms the fact that, absent perfect competition – as in our model –, firms increase sales by increasing prices, thereby suppressing produced quantities. In turn, this mechanism implies that growing firms demand less employment, which is the reason why labor income shares do not move one-for-one with sales shares.

Table 3: Estimated Relationship between Wages and Sales

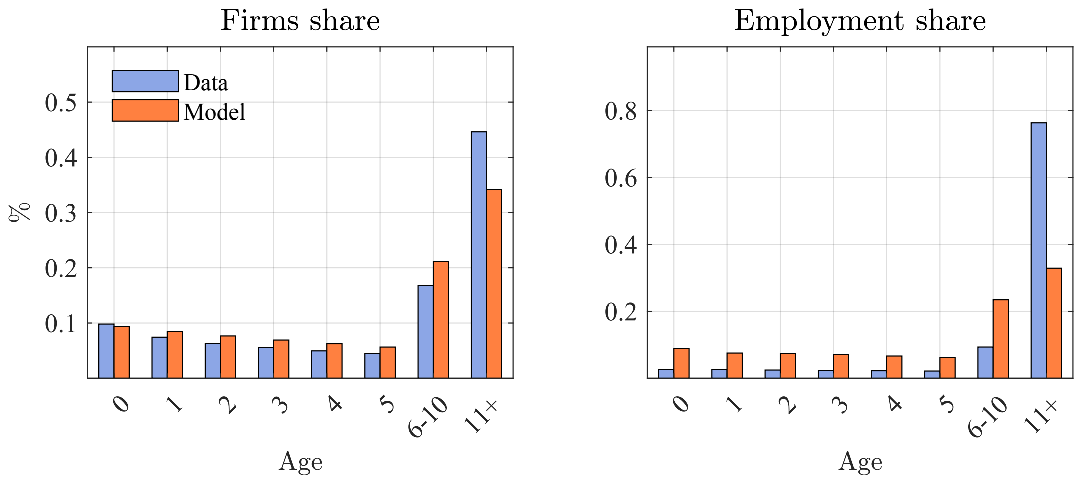
	Model	Data
Elasticity of Labor Income Shares to Sales Shares	0.87	0.88

5.2.3 Other Cross-sectional and Life-Cycle Properties

In the last exercise of this section, we analyze the distribution of firms by age, and the life-cycle profile of employment and sales growth rates in our model economy. In Figure 7, we report the distribution of firms and employment shares by age, comparing the empirical ones from Compustat (1990q1-2016q4) to the ones obtained in our quantified framework. Note that neither of these distributions was targeted in the calibration, and hence both comparisons are to be considered as a pure validation exercise. First, focusing on the left panel of Figure 7, one can conclude that our framework overall succeeds in replicating the distribution of firms over age, albeit it partially underestimates the share of businesses that are 11+ years old. Since most of our analysis is focused on markups' properties over the life-cycle of firms, it is important that we are able to capture the correct number of firms across age bins. In particular, the share of companies in each age bin influences the heterogeneous response of markups' to monetary policy shocks, and hence is relevant to get a correct quantitative fit of the empirically estimated dynamics of markups by firms' age.

Secondly, the right panel of Figure 7 plots the distribution of employment shares over firm age, comparing the empirical ones with their model-implied counterparts. Clearly, our framework

Figure 7: Distributions of Firms and Employment Shares by Age

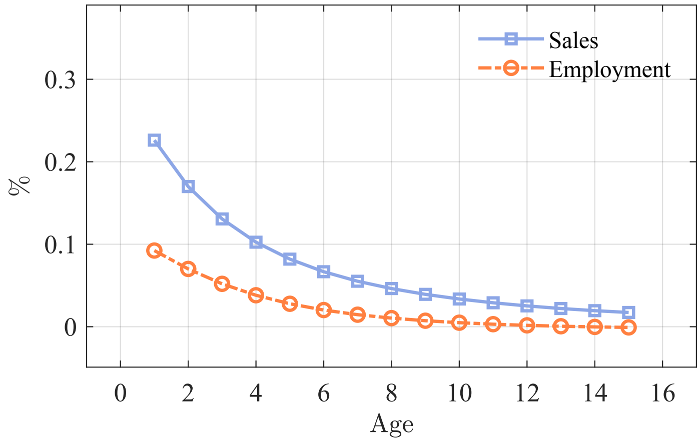


is able to match only up to half of the right tail in the employment share distribution. This is precisely due to the fact that, in our model economy, big firms (and hence old firms), find optimal to increase sales by increasing prices, thereby suppressing produced quantities and employment demand. This mechanism is a key characteristic of our set up, in which companies operate in an environment with imperfect competition, and it is hence responsible for the fact that 11+ years old firms in the model generate a lower employment share compared to their empirical counterpart.[12]

As a final note, in Figure 8 we plot the average employment and sales growth rates over the life-cycle of firms. Understandably, both measures decrease over time, as companies become old and hence slow down in their growth processes: this means that growth rates are unconditionally negatively correlated with age, as empirically noted in Dunne et al. (1989). However, sales grow relatively more than employment, which is consistent with the early discussion related to the employment share distribution depicted in the right panel of Figure 7. In particular, as argued in the previous paragraphs and due to the presence of the Kimball aggregator, markups do increase with firms sales, implying that the wage bill increases less than one-for-one with sales, depressing the labor demand by firms and resulting in lower employment growth rates compared to the growth rate of firm sales. In other words, due to market power, companies can increase sales by raising prices and decreasing output, which lowers their demand of labor and hence employment growth.

[12] A possible way to attain a better fit in this dimension would be to allow firms to be hit by other shocks over the life-cycle, which may for instance expand permanently their production capacity. This could generate "superstar" firms in the model, with larger than average employment shares and higher markups. At the firm level, this extra element would further strengthen the link between firm age and cost shares. At the aggregate level, average pass-through could be lower than what is implied by our baseline framework. Since the current version of our model misses this ingredient, we hence consider the quantitative claims Section 6, regarding shocks amplification, to be conservative estimates.

Figure 8: Employment and Sales Growth Rates



6 Results

In the following section, we begin by analyzing the response of firm markups to interest rate shocks, and compare the relative response of old and young firms in the model with the ones obtained in the data and reported in Section 4. Secondly, having assessed how much of the heterogeneity in the response of markups to MP shocks by firm age our model is able to replicate, we illustrate how changes in aggregate variables after a MP shock contribute to the differential response of markups of old firms with respect to young ones. Finally, we conclude with a note on the amplification of shocks at work in our framework compared to a standard one-firm NK model.

6.1 Response of Markups to Monetary Policy Shocks

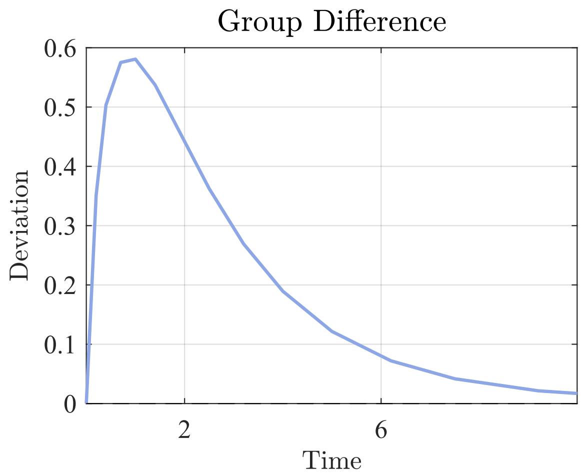
We proceed to illustrate the dynamics of the economy after the arrival of a negative MP shock and compare the response of old and young firms' markups to the one obtained from the data and discussed in Section 4. As standard in frameworks characterized by nominal rigidities, a negative MP shock features an increase in the nominal interest rate and implies a downward pressure on the labor cost W. Both employment, consumption, and output decrease on impact and slowly recover as the shock fades away, while the downward pressure on prices engineers a deflationary episode. Moreover, the aggregate markup increases as a result of decreasing labor costs, and hence shows a countercyclical behavior in response to negative shocks to the nominal interest rate. The aggregate response of our calibrated economy resembles qualitatively the one of a standard NK textbook model, as in Galí (2015).[13] However, the aggregate pattern of markups masks a noticeable degree of heterogeneity at the firm-level, which we further explore in the following paragraphs.

To obtain a comparable setup to the empirical analysis, we first categorize firms in the model

[13]We report the aggregate impulse response functions of our model economy in Appendix C.

economy by their age decile, and classify businesses above the median age as "old" and below the median age as "young". We then simulate the hit of a negative MP shock, defined as an exogenous increase in the nominal interest rate, and compute the differential response of markups to a MP shock across firms above and below the median age. Similarly to the empirical results presented in Section 4, we plot the differential response of markups by firm age over a horizon of several quarters and in deviation from the mean response. To be more precise, we take the markup response of old firms and subtract the one of young firms, which represents the group difference between the two sets of firms. As shown in Figure 9, firms above the median age show an excess countercyclicality after a negative MP shock, consistent with our evidence from Compustat data.

Figure 9: Markups IRFs After a Negative MP Shock



Note that the differential response of old firms' markups upon a negative MP shock peaks at a value of 0.6% in the model, while empirically it goes up to more than 2%. Our quantitative framework is hence able to qualitatively reproduce the heterogeneity in the response of markups by firm age that we have documented in Compustat data. Moreover, it quantitatively replicates around 30% of the excess countercyclicality of old firms' markups in response to MP shocks.

To conclude, we stress that the overall difference in markups we observe in the model is attributed to differences across firms regarding both the desired markup change after a MP shock – which is governed by the Kimball aggregator in the demand function – and price adjustment costs. In our model, the latter component is proportional to firm sales, which means that our set up does not bias the cost of price adjustment towards big or small firms. It is important to note that the relatively modest difference in markups response by firm age can increase in a calibration that features higher superelasticity ω, and relatively larger price adjustment costs for bigger firms.

6.2 Decomposing the Differential Response of Markups

In Section 2.6.1, we have shown analytically that MP shocks affect prices and markups mostly via their effect on aggregate variables. Moreover, to get analytical solutions, we have derived expressions for the changes in prices (or markups) in response to changes in output or the labor cost. The same decomposition can be carried out quantitatively in the calibrated model with nominal rigidities. Specifically, we compute numerically the general equilibrium (GE) response of the economy to a negative MP shock. Then, taking as given the equilibrium paths for the aggregate variables $Y, W, \mathcal{D}, r, \pi$ we look at the partial (and relative) responses of old firms' markups to the changes in each aggregate variable separately. Note that, standard decomposition exercises typically illustrate the direct and indirect contributors to the response of aggregate output following a MP shock. Instead, here we focus on identifying the drivers of heterogeneity in the response of markups across firms. For this reason, the direct and indirect contributors to the response of markups and aggregate output respectively should not necessarily behave the same or carry the same weight.

Before discussing the results, we follow Kaplan et al. (2018) and provide intuition for the channels at play in our economy with heterogeneous firms and endogenous markups. Let us first write the difference between the average markups of old firms and the average markups of young firms as a function of equilibrium prices, quantities, and inflation. We collect these terms in the vector $\{\mathcal{S}_t\}_{t\geq 0}$, with $\mathcal{S}_t = \{r_t, W_t, Y_t, \mathcal{D}_t, \pi_t\}$, and define the above-mentioned difference $\widehat{\mathcal{M}}(\{\mathcal{S}_t\}_{t\geq 0})$ induced by the path of the MP shock $\{\varepsilon_t\}_{t\geq 0}$ from its initial hit until it fully reverts to zero as:

$$\widehat{\mathcal{M}}(\{\mathcal{S}_t\}_{t\geq 0}) := \int \mu_t(p,a;\{\mathcal{S}_t\}_{t\geq 0})\mathbb{1}\{g_t(p,a) \geq \overline{\xi}\}d\lambda_t - \int \mu_t(p,a;\{\mathcal{S}_t\}_{t\geq 0})\mathbb{1}\{g_t(p,a) < \overline{\xi}\}d\lambda_t. \tag{31}$$

where $\mu_t(p,a;\{\mathcal{S}_t\}_{t\geq 0})$ is the firm markup, $g_t(p,a)$ is a mapping between firm's states and its age, $\overline{\xi}$ is the median firm age, and $d\lambda_t(p,a;\{\mathcal{S}_t\}_{t\geq 0})$ is the joint distribution of firm prices and idiosyncratic demand. Totally differentiating Equation 31, we can then decompose the difference in the response of the average markup between old and young firms at time $t=\tau$ as:

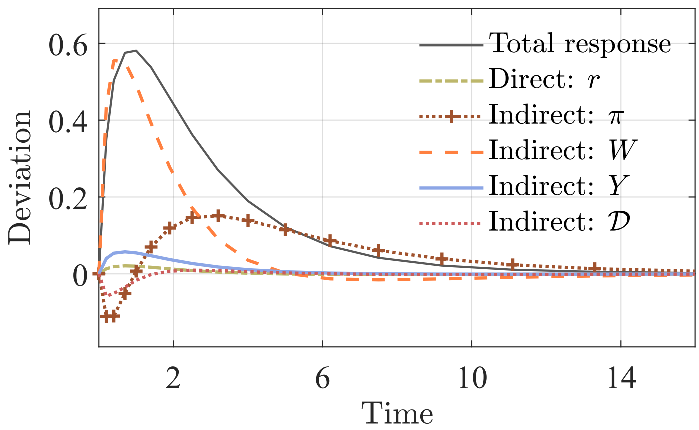
$$d\widehat{\mathcal{M}}_\tau = \underbrace{\int_\tau^\infty \frac{\partial \widehat{\mathcal{M}}_\tau}{\partial r_t} dr_t dt}_{\text{direct effect}} + \underbrace{\int_\tau^\infty \left(\frac{\partial \widehat{\mathcal{M}}_\tau}{\partial W_t} dW_t + \frac{\partial \widehat{\mathcal{M}}_\tau}{\partial Y_t} dY_t + \frac{\partial \widehat{\mathcal{M}}_\tau}{\partial \mathcal{D}_t} d\mathcal{D}_t + \frac{\partial \widehat{\mathcal{M}}_\tau}{\partial \pi_t} d\pi_t \right) dt}_{\text{indirect effect}} \tag{32}$$

where the first term reflects the direct effect of a change in the interest rate, which enters the Euler equation of the agents, holding the other variables of interest constant. The remaining terms in the decomposition reflect the indirect effects of changes in inflation, the real wage, real output and the demand index that arise in general equilibrium after the hit of the MP shock. In practice, we need to compute each of these components numerically. For example, the formal definition of the first term in Equation 32, which is the direct effect of changes in the real interest rate $\{r_t\}_{t\geq 0}$, is:

$$\int_{\tau}^{\infty} \frac{\partial \widehat{\mathcal{M}}_{\tau}}{\partial r_t} dr_t dt = \int_{\tau}^{\infty} \frac{\partial \widehat{\mathcal{M}}(\{r_t, \overline{W}, \overline{Y}, \overline{\mathcal{D}}, \overline{\pi}\}_{t\geq 0})}{\partial r_t} dr_t dt. \tag{33}$$

This term represents the *partial-equilibrium* response of the difference in the average markups between old and young firm that face a time-varying real interest rate path $\{r_t\}_{t\geq 0}$, but holding the paths for the real wage $\overline{W}$, the real output $\overline{Y}$, the demand index $\overline{\mathcal{D}}$, and nominal inflation rate $\overline{\pi}$ constant at their steady-state values. We calculate this term from the model by feeding the time paths for these variables into the firms' (and household's) optimization problem, computing the policy function and markups for each firm, and aggregating across firms using the corresponding distribution. Note that the other terms in the decomposition are computed in a similar fashion.

Figure 10: Decomposing the Differential Response of Markups



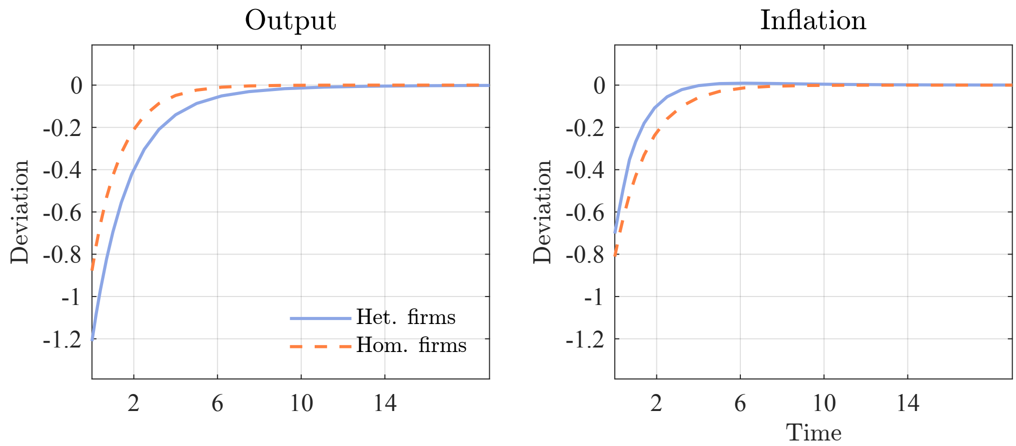
The results of the decomposition exercise are shown in Figure 10, and all the effects are to be intended as percentage deviations from the mean response across firms in the economy. The outer dark line represents the total GE effect of a negative MP shock on the differential impulse response of markups for old firms compared to young ones. It is important to note that the GE effect is not a direct sum of the partial effects due to non-linearities in aggregation. Clearly, most of the resulting effect on the differential response of old firms' markups is to be attributed to changes in the aggregate W, hence to changes in the cost of labor after a negative shock to the interest rate.

We want to stress again that both demand heterogeneity and the presence of a Kimball aggregator are key to deliver the result in Figure 10. As theoretically shown in Section 2.6.1, since our model features heterogeneous and endogenous markups in the presence of a Kimball aggregator, old firms have a lower pass-through from production costs to prices. As such, a negative MP shock in the economy puts a downward pressure on the labor input cost W, but old firms' sales react less than proportionally, as dominant companies do not decrease prices as much. Since markups are the ratio between business sales and costs, the resulting effect on markups is positive, leading to the observed stronger countercyclical response of old firms' markups to a negative MP shock.

6.3 Amplification Mechanism

In what follows, we conclude our quantitative analysis by studying the extent of MP shock amplification at work in the economy, and comparing our calibrated framework with a standard one-firm NK model. As pointed out in an early contribution by Ball and Romer (1990), and recently stressed by Mongey (2017), shocks have a strong propagation through quantities in economies where real rigidities are present, and this is what we set to verify in our case as well. Moreover, we further explain to which extent both firm heterogeneity and the presence of a Kimball aggregator can be conducive to greater movements in macroeconomic aggregates in response to MP shocks.

Figure 11: Comparing Output and Inflation Responses



To ensure comparability across economies, we calibrate the one-firm NK economy to have the same size as in our heterogeneous firms framework (hereafter: FDNK) in terms of overall output. Moreover, since the standard NK model features a constant elasticity of demand, we set the elasticity of substitution σ in its CES aggregator to match an aggregate steady state markup of 1.68, which is the value targeted in the FDNK model under the Kimball aggregator. With the two models at hand, we simulate a negative MP shock and solve for the response of the main macroeconomic aggregates in the two economies. In particular, we analyse the trajectories of inflation π and output Y over an 16-quarters period, and hence compare the relative percentage deviation from steady state values of both prices and quantities. The results of this exercise are depicted in Figure 11.

Focusing on the response of output and inflation across the two models, it is clear that a negative MP shock produces a bigger drop in output and a milder decline in prices in our FDNK set up compared to a standard one-firm NK model. The negative change in the interest rate decreases output by on average 20 p.p. more in the economy characterised by heterogeneous firms and endogenous markups, with the effect lasting for more than 10 quarters after the shock. At the same time, prices and hence inflation drop by relatively more in the one-firm NK model, which implies that the presence of the Kimball aggregator and the differential pass-through that characterize our model economy mitigate the downward pressure exerted by the negative MP shock on firm prices.

On the one hand, as argued in Klenow and Willis (2016), the presence of the Kimball aggregator adds a source of real frictions in the NK model, represented by a higher degree of concavity in the firm's profit function with respect to its relative price. Under the Kimball aggregator, sellers face a price elasticity of demand that is increasing in their good's relative price. For instance, a repricing firm facing lower labor costs after a negative MP shock will temper its price drop because of the endogenous increase in its desired markup, and this effect would be stronger the lower the elasticity of demand faced by the producer. Since the presence of a real rigidity makes firms more reluctant to change prices, firms do not pass marginal cost shocks as fully onto their prices as they would in a standard NK model with a CES aggregator. Hence, in our FDNK set up, MP shocks propagate more through quantities than prices, and decrease aggregate output by relatively more.

On the other hand, without heterogeneity on the firm's side, the presence of the Kimball aggregator alone does not automatically lead to the amplification of shocks. In fact, the effects of the real rigidity introduced by the Kimball aggregator kick in only when firms are indeed heterogeneous and characterized by different pass-throughs from costs to prices with respect to one another. If all firms were to be equal (as in the one-firm NK model), they would also be equal to the average firm in the economy and have identical sales shares.[14] Specifically, focusing on Equation 7, the elasticity of demand faced by producers would not vary across firm, and their response to MP shocks would be identical. On the contrary, in our set up, since big firms (hence old firms) respond more countercyclically than small ones and decrease their prices by less, the propagation of a negative shock gets strengthened. The heterogeneity of firms, combined with the real rigidity introduced by the Kimball aggregator, therefore delivers the amplification mechanism at work in the model.

As a concluding remark, it is important to stress that our quantitative framework does not allow for endogenous entry and exit of firms. As shown in Jaimovich and Floetotto (2008), changes to the number of operating firms alone could lead to countercyclical markups variations even in homogeneous firms NK models with constant demand elasticity. Yet, as discussed in Section 2, exit in our set up is modelled as a Poisson process with given arrival rate, and exiters are replaced by new entrants so that the mass of firms stay constant. Hence, since we do not allow for endogenous

[14] In Appendix C, we show the comparison between the differential response of old and young firms' markups to a MP shock between our baseline calibration and one where firms' demand heterogeneity is reduced. In this illustrative case, the steady state heterogeneity of markups is much lower, and thus, the heterogeneity in responses to MP shock is almost zero. Notice that the assumption of Rotemberg pricing plays crucial role in this argument: in our setting, homogeneous firms make similar pricing decisions. Instead, under Calvo pricing, pricing would have differ even across firms that are homogeneous in the level of demand they face, thus resulting in heterogeneity of markups as well. We leave the assessment of the performance of the model under different pricing assumptions for future investigation.

business formation, we can conclude that the excess countercyclicality of old firms' markups in our model economy cannot be attributed to changes in the composition of operating firms.

7 Conclusion

In this paper, we have taken a theoretical and empirical approach to the study of firm-level heterogeneity in the response of markups to MP shocks. First, we have built a novel NK model, augmented with heterogeneous firms and a process of demand accumulation, in which markups arise endogenously and evolve over the life-cycle of the firms. We have highlighted how firm prices respond to aggregate changes, which sufficient statistic explains the heterogeneity in these responses, and how to empirically proxy it. In mapping such sufficient statistic to the data, we have merged exogenously-identified MP shocks series with a rich quarterly dataset comprising publicly-listed companies based in the US between 1990q1 and 2016q4. Next, we have documented that old firms' markups tend to increase after a monetary policy tightening, while young firms' markups show a mildly procyclical behavior after a negative interest rate shock. Connecting to existing literature, we suggest that the differential response of markups by firm age could be related to a process of accumulation of customers and demand over time, which may enable older firms to change by relatively less their prices after a MP shock, thanks to an established position in their markets.

We have further shown that our calibrated framework can replicate the life-cycle profile of firms' markups, as well as their growth rates of sales and employment, and the distribution of companies and employment shares by firm age. Moreover, the model explains almost a third of the empirically estimated excess countercyclicality in the markups of firms above the median age after a negative MP shock. Finally, we have shown that both firms' heterogeneity and endogenous markups generate amplification in the response of aggregate quantities to contractionary interest rate movements, which distinguishes our set up from standard frameworks with nominal rigidities. In the future, we aim to study further optimal monetary policies in the presence of imperfect competition, demand accumulation, and heterogeneity in the pass-through from costs to prices.

References

AFROUZI, H. AND L. CALOI (2022): "Endogenous firm competition and the cyclicality of markups," *The Review of Economics and Statistics*, 1–45.

ALATI, A. (2020): "Essays on firms heterogeneity and business cycles," Ph.D. thesis, The London School of Economics and Political Science (LSE).

ALVAREZ, F. E., F. LIPPI, AND T. SOUGANIDIS (2022): "Price setting with strategic complementarities as a mean field game," Tech. rep., National Bureau of Economic Research.

ATKESON, A. AND A. BURSTEIN (2008): "Pricing-to-market, trade costs, and international relative prices," *American Economic Review*, 98, 1998–2031.

BALEY, I. AND A. BLANCO (2019): "Firm uncertainty cycles and the propagation of nominal shocks," *American Economic Journal: Macroeconomics*, 11, 276–337.

BALL, L. AND D. ROMER (1990): "Real rigidities and the non-neutrality of money," *The Review of Economic Studies*, 57, 183–203.

BAQAEE, D., E. FARHI, AND K. SANGANI (2021): "The Supply-Side Effects of Monetary Policy," Tech. rep., National Bureau of Economic Research.

BAQAEE, D. R. AND E. FARHI (2020): "Productivity and misallocation in general equilibrium," *The Quarterly Journal of Economics*, 135, 105–163.

BILBIIE, F. O., F. GHIRONI, AND M. J. MELITZ (2007): "Monetary policy and business cycles with endogenous entry and product variety [with comments and discussion]," *NBER Macroeconomics Annual*, 22, 299–379.

——— (2012): "Endogenous entry, product variety, and business cycles," *Journal of Political Economy*, 120, 304–345.

BILS, M., P. J. KLENOW, AND B. A. MALIN (2018): "Resurrecting the role of the product market wedge in recessions," *American Economic Review*, 108, 1118–46.

BOND, S., A. HASHEMI, G. KAPLAN, AND P. ZOCH (2021): "Some unpleasant markup arithmetic: Production function elasticities and their estimation from production data," *Journal of Monetary Economics*, 121, 1–14.

BURSTEIN, A., V. M. CARVALHO, AND B. GRASSI (2020): "Bottom-up markup fluctuations," Tech. rep., National Bureau of Economic Research.

CLOYNE, J., C. FERREIRA, M. FROEMEL, AND P. SURICO (2018): "Monetary policy, corporate finance and investment," Tech. rep., National Bureau of Economic Research.

——— (2023): "Monetary policy, corporate finance, and investment," *Journal of the European Economic Association*, jvad009.

COLCIAGO, A., V. LINDENTHAL, AND A. TRIGARI (2019): "Who Creates and Destroys Jobs over the Business Cycle?" .

DAVIS, S. J., J. HALTIWANGER, R. JARMIN, J. MIRANDA, C. FOOTE, AND E. NAGYPAL (2006): "Volatility and dispersion in business growth rates: Publicly traded versus privately held firms [with comments and discussion]," *NBER macroeconomics annual*, 21, 107–179.

DE LOECKER, J., J. EECKHOUT, AND G. UNGER (2020): "The rise of market power and the macroeconomic implications," *The Quarterly Journal of Economics*, 135, 561–644.

DE LOECKER, J. AND F. WARZYNSKI (2012): "Markups and firm-level export status," *American economic review*, 102, 2437–71.

DE RIDDER, M., B. GRASSI, AND G. MORZENTI (2021): "The hitchhikerâ€™s guide to markup estimation," Tech. rep.

DEB, S., J. EECKHOUT, A. PATEL, AND L. WARREN (2020): "What drives Wage Stagnation: Monopoly or Monopsony?" Tech. rep., UPF mimeo.

DUNNE, T., M. J. ROBERTS, AND L. SAMUELSON (1989): "The growth and failure of US manufacturing plants," *The Quarterly Journal of Economics*, 104, 671–698.

EDMOND, C., V. MIDRIGAN, AND D. Y. XU (2018): "How costly are markups?" Tech. rep., National Bureau of Economic Research.

FABIANI, A., L. FALASCONI, AND J. HEINEKEN (2020): "Monetary Policy and Corporate Debt Maturity," .

FANG, M. (2022): "Lumpy investment, fluctuations in volatility and monetary policy," in *Lumpy Investment, Fluctuations in Volatility and Monetary Policy: Fang, Min*, [Sl]: SSRN.

FORT, T. C., J. HALTIWANGER, R. S. JARMIN, AND J. MIRANDA (2013): "How firms respond to business cycles: The role of firm age and firm size," *IMF Economic Review*, 61, 520–559.

FOSTER, L., J. HALTIWANGER, AND C. SYVERSON (2008): "Reallocation, firm turnover, and efficiency: Selection on productivity or profitability?" *American Economic Review*, 98, 394–425.

GALÍ, J. (2015): *Monetary policy, inflation, and the business cycle: an introduction to the new Keynesian framework and its applications*, Princeton University Press.

GALI, J., M. GERTLER, AND J. D. LOPEZ-SALIDO (2007): "Markups, gaps, and the welfare costs of business fluctuations," *The review of economics and statistics*, 89, 44–59.

GERTLER, M. AND S. GILCHRIST (1994): "Monetary policy, business cycles, and the behavior of small manufacturing firms," *The Quarterly Journal of Economics*, 109, 309–340.

GILCHRIST, S., R. SCHOENLE, J. SIM, AND E. ZAKRAJŠEK (2017): "Inflation dynamics during the financial crisis," *American Economic Review*, 107, 785–823.

GOLOSOV, M. AND R. E. LUCAS (2007): "Menu costs and Phillips curves," *Journal of Political Economy*, 115, 171–199.

GOPINATH, G. AND O. ITSKHOKI (2010): "Frequency of price adjustment and pass-through," *The Quarterly Journal of Economics*, 125, 675–727.

GRASSI, B. (2017): "IO in IO: Competition and volatility in input-output networks," *Unpublished Manuscript, Bocconi University*.

GÜRKAYNAK, R. S., B. SACK, AND E. SWANSON (2005): "The sensitivity of long-term interest rates to economic news: Evidence and implications for macroeconomic models," *American economic review*, 95, 425–436.

GUTIÉRREZ, G. AND T. PHILIPPON (2016): "Investment-less growth: An empirical investigation," Tech. rep., National Bureau of Economic Research.

HALL, R. E. (1988): "The relation between price and marginal cost in US industry," *Journal of political Econ-*

omy, 96, 921–947.

HAMANO, M. AND F. ZANETTI (2022): "Monetary policy, firm heterogeneity, and product variety," *European Economic Review*, 144, 104089.

HONG, S. (2017): "Customer capital, markup cyclicality, and amplification," *FRB St. Louis Working Paper*.

HÖYNCK, C., M. LI, AND D. ZHANG (2023): "The (Unequal) Rise of Market Power, Nominal Rigidities, and Monetary Non-Neutrality," .

JAIMOVICH, N. AND M. FLOETOTTO (2008): "Firm dynamics, markup variations, and the business cycle," *Journal of monetary Economics*, 55, 1238–1252.

JAROCIŃSKI, M. AND P. KARADI (2020): "Deconstructing monetary policy surprisesâ€"the role of information shocks," *American Economic Journal: Macroeconomics*, 12, 1–43.

JEENAS, P. (2019): "Firm Balance Sheet Liquidity, Monetary Policy Shocks, and Investment Dynamics," .

JORDÀ, Ò. (2005): "Estimation and inference of impulse responses by local projections," *American economic review*, 95, 161–182.

KAPLAN, G., B. MOLL, AND G. L. VIOLANTE (2018): "Monetary policy according to HANK," *American Economic Review*, 108, 697–743.

KLENOW, P. J. AND J. L. WILLIS (2016): "Real rigidities and nominal price changes," *Economica*, 83, 443–472.

LEVY, D., M. BERGEN, S. DUTTA, AND R. VENABLE (1997): "The magnitude of menu costs: direct evidence from large US supermarket chains," *The Quarterly Journal of Economics*, 112, 791–824.

MEIER, M. AND T. REINELT (2022): "Monetary policy, markup dispersion, and aggregate tfp," *Review of Economics and Statistics*, 1–45.

MEINEN, P. AND A. C. SOARES (2022): "Markups and Financial Shocks," *The Economic Journal*, 132, 2471–2499.

MONGEY, S. (2017): "Market structure and monetary non-neutrality," *Job market paper. http://www. simon-mongey. com/uploads/6/5/6/6/65665741/mongey_market_structure_monetary_nonneutrality_draft.pdf*.

NEKARDA, C. J. AND V. A. RAMEY (2020): "The Cyclical Behavior of the Price-Cost Markup," *Journal of Money, Credit and Banking*, 52, 319–353.

OTTONELLO, P. AND T. WINBERRY (2020): "Financial heterogeneity and the investment channel of monetary policy," *Econometrica*, 88, 2473–2502.

RAVN, M., S. SCHMITT-GROHÉ, AND M. URIBE (2006): "Deep habits," *The Review of Economic Studies*, 73, 195–218.

TAYLOR, J. B. (1999): "The robustness and efficiency of monetary policy rules as guidelines for interest rate setting by the European Central Bank," *Journal of Monetary Economics*, 43, 655–679.

TEULINGS, C. N. AND N. ZUBANOV (2014): "Is economic recovery a myth? Robust estimation of impulse responses," *Journal of Applied Econometrics*, 29, 497–514.

WANG, O. AND I. WERNING (2022): "Dynamic oligopoly and price stickiness," *American Economic Review*, 112, 2815–2849.

ZBARACKI, M. J., M. RITSON, D. LEVY, S. DUTTA, AND M. BERGEN (2004): "Managerial and customer costs of price adjustment: direct evidence from industrial markets," *Review of Economics and statistics*, 86, 514–533.

A Analytical Appendix

In what follows, we first report the derivations related to how, in the model, we can decompose firm nominal price changes into their relative price changes and aggregate inflation. This helps visualize more clearly the notion of inflation in our framework. Second, we explain in full detail the numerical strategy we adopt to solve our quantitative model, which makes use of a variable substitution method. Finally, we derive the analytical expressions for the changes of prices and markups conditional on changes in aggregate variables, such as output and the cost of labor, which illustrates the main channels through which MP shocks propagate and affect firms' market power.

A.1 Price and inflation

The demand of a given firm on the market depends on its relative price, so we use the relative firm's prices $p_{i,t}$ in the equations in the main text. At the same time, Rotemberg price adjustment costs depend on the square of the nominal price change of the firm. Below we show how the firm's nominal price change can be decomposed into relative price change and aggregate inflation. Note first that the relative price is defined as:

$$p_{i,t} = \frac{P_{i,t}}{P_t} \tag{34}$$

Then, the firms' nominal price change can be expressed as following:

$$\frac{\dot{P}_{i,t}}{P_{i,t}} = \frac{\dot{P}_t p_{i,t} + P_t \dot{p}_{i,t}}{P_t p_{i,t}} = \frac{\dot{P}_t}{P_t} + \frac{\dot{p}_{i,t}}{p_{i,t}} = \pi_t + \frac{\dot{p}_{i,t}}{p_{i,t}} \tag{35}$$

A.2 Variable substitution

Here, we show the analytical reasoning behind the variable substitution method we apply to the firm's problem in order to simplify the numerical solution. The main idea for this substitution is that the state variable $p_{i,t}$ used to solve the firm's optimization problem is not convenient for the numerical solution. The reason lies is the demand function from the Kimball aggregator, given by:

$$y = \left(1 - \omega \log\left(\frac{\sigma}{\sigma - 1}\frac{1}{\xi(a)}\frac{p}{\mathcal{D}}\right)\right)^{\sigma/\omega} \frac{Y}{\xi(a)} \tag{36}$$

This demand function intersects the axis at different levels of price for the different values of a. This means that if one has a grid with an orthogonal basis, which is necessary for exploiting the continuous-time approach with a finite difference solution method, some of the points will not satisfy the non-negative values of demand. Moreover, a high precision in computing the solution is needed close to the boundary of zero demand. This is particularly crucial because the significant level of firm heterogeneity we embed in our framework will require a lot of points on the grid.

As mentioned above, the method we propose to tackle this problem is a variable substitution. Despite the seeming simplicity of the variable substitution concept and wide use in the discreet time approach, to the best of our knowledge, this work is the first one to provide an example of such substitution in a continuous-time approach. This includes the treatment of the associated analytical challenges and the characterization of the numerical solution method, associated with the new problem with the substituted variable. Our derivation reconciles the problem with the standard finite-difference approach, with minor changes, due to the substitution. Specifically, the substitution concerning firms' prices that we propose in this problem is the following one:

$$s \equiv \log\left(\frac{1}{\xi(a)}\frac{p}{\mathcal{D}}\right) \tag{37}$$

With this substitution, instead of the previous control variable p, we have to deal with a new variable that depends on the control variable p, stochastic process of demand accumulation a, and the aggregate Kimball demand index $\mathcal{D}$. This also requires a careful treatment of the stochastic differential equation associated with the evolution of the new variable s. Notice that for the demand process specification assumed in the paper ($\xi(a) = exp(a)$) the following simplification applies:

$$s = \log(p) - \log(\mathcal{D}) - a \tag{38}$$

which implies the following differential equation for the evolution of our new variable s:

$$ds = d\log(p) - d\log(\mathcal{D}) - \rho_a(\bar{a} - a)dt - \sigma_a d\mathcal{W} \tag{39}$$

Combining the above equation with the evolution of the state variable a, we get the following stochastic process that determines the evolution of the two state variables in the matrix notation:

$$d\begin{pmatrix} s \\ a \end{pmatrix} = \mu dt + GdB \tag{40}$$

where μ is the vector determining the drift of the two entries in the stare variables vector given by:

$$\mu = \begin{pmatrix} d\log(p)/dt - d\log(\mathcal{D})/dt - \rho_a(\bar{a} - a) \\ \rho_a(\bar{a} - a) \end{pmatrix} \tag{41}$$

and GdB is the matrix multiplying the Wiener process determining the stochastic component:

$$G = \begin{pmatrix} -\sigma_a & 0 \\ \sigma_a & 0 \end{pmatrix}, \quad dB = \begin{pmatrix} d\mathcal{W} \\ 0 \end{pmatrix} \tag{42}$$

Now, to determine the change in the evolution of the new value function – defined over the new set of state variables –, we use Ito's lemma for multivariate stochastic processes, which gives us:

$$d\mathcal{J}(s,a,t) = \left\{ \frac{\partial \mathcal{J}}{\partial t} + (\nabla_{s,a}\mathcal{J})^T \mu + \frac{1}{2} Tr[G^T(H_{s,a}\mathcal{J})G] \right\} dt + (\nabla_{s,a}\mathcal{J})^T GdB \tag{43}$$

Substituting for the derivatives and hessian of the value function we get the following expression:

$$d\mathcal{J}(s,a,t) = \left\{ (d\log(p)/dt - d\log(\mathcal{D})/dt - \rho_a(\bar{a}-a)) \frac{\partial \mathcal{J}}{\partial s} + \frac{\sigma_a^2}{2}\frac{\partial^2 \mathcal{J}}{\partial s^2} + \right. \tag{44}$$

$$\left. + \rho_a(\bar{a}-a)\frac{\partial \mathcal{J}}{\partial a} + \frac{\sigma_a^2}{2}\frac{\partial^2 \mathcal{J}}{\partial a^2} - \sigma_a^2 \frac{\partial^2 \mathcal{J}}{\partial s \partial a} + \frac{\partial \mathcal{J}}{\partial t} \right\} dt + (\nabla_{s,a}\mathcal{J})^T GdB \tag{45}$$

Notice that, because of the same stochastic term affecting both state variables, the crucial component that has to be accounted for is the cross-derivative of the value function. Finally, we can rewrite the HJB equation in terms of the new value function defined on the state space with the variable substitution:

$$\rho\mathcal{J}(s,a,t) = p_{i,t}y_{i,t} - W_t\ell_{i,t} - W_t\frac{\vartheta}{2}\left(\pi_t + \frac{\dot{p}_{i,t}}{p_{i,t}}\right)^2 p_{i,t}y_{i,t} + \tag{46}$$

$$+ \left(\frac{d\log(p_{i,t})}{dt} - \frac{d\log(\mathcal{D}_t)}{dt} - \rho_a(\bar{a}-a_{i,t})\right)\frac{\partial \mathcal{J}}{\partial s} + \frac{\sigma_a^2}{2}\frac{\partial^2 \mathcal{J}}{\partial s^2} + \tag{47}$$

$$+ \rho_a(\bar{a}-a)\frac{\partial \mathcal{J}}{\partial a} + \frac{\sigma_a^2}{2}\frac{\partial^2 \mathcal{J}}{\partial a^2} - \sigma_a^2\frac{\partial^2 \mathcal{J}}{\partial s \partial a} + \frac{\partial \mathcal{J}}{\partial t} \tag{48}$$

$$\text{s.t.} \quad y_{i,t} = \left(1 - \omega\left(\log\left(\frac{\sigma}{\sigma-1}\right) + s_{i,t}\right)\right)^{\sigma/\omega} \frac{Y_t}{\xi(a_{i,t})} \tag{49}$$

$$y_{i,t} = \ell_{i,t}^{1-\alpha} \tag{50}$$

$$da_{i,t} = \rho_a(\overline{a} - a_{i,t})dt + \sigma_a d\mathcal{W}_{i,t} \tag{51}$$

$$ds_{i,t} = d\log(p_{i,t}) - d\log(\mathcal{D}) - \rho_a(\bar{a} - a_{i,t})dt - \sigma_a d\mathcal{W}_{i,t} \tag{52}$$

$$s_{i,0} \text{ and } a_{i,0} \text{ given} \tag{53}$$

The first order condition with respect to $d\log(p)/dt$ has the same form as before, namely:

$$W_t\theta\left(\pi_t + \frac{\dot{p}_{i,t}}{p_{i,t}}\right)p_{i,t}y_{i,t} = \frac{\partial \mathcal{J}}{\partial s} \tag{54}$$

Moreover, the equations above define firms' optimality conditions.

To implement the solution of this problem numerically, we build the grids for the state variables s and a and use the standard continuous-time finite difference method with the non-equally spaced grid for s. The additional part that we have to consider are the components attributed to $\frac{\sigma_a^2}{2}\frac{\partial^2 \mathcal{J}}{\partial s^2}$ and $\sigma_a^2 \frac{\partial^2 \mathcal{J}}{\partial s \partial a}$, where the first one is calculated in the same way as the one for a and given by:

$$\frac{\partial^2 \mathcal{J}}{\partial s^2} = \frac{\mathcal{J}(s_{n+1}, a_m) - 2\mathcal{J}(s_n, a_m) + \mathcal{J}(s_{n-1}, a_m)}{\Delta s^2} \tag{55}$$

And the second one, namely the cross-derivative, is computed instead as follows:

$$\frac{\partial^2 \mathcal{J}}{\partial s \partial a} = \frac{\mathcal{J}(s_{n+1}, a_{m+1}) - \mathcal{J}(s_{n-1}, a_{m+1}) - \mathcal{J}(s_{n+1}, a_{m-1}) + \mathcal{J}(s_{n-1}, a_{m-1})}{4\Delta s \Delta a} \tag{56}$$

A.3 Flexible prices

In what follows, we show the derivations for the case of the partial equilibrium response of firms' prices and markups to changes affecting aggregate state variables. In particular, we are interested in the behavior of the markup gap between high and low-demand firms in response to the cost-push shock (increase in W). Recall that the demand function in our model is given by:

$$y = \left(1 - \omega \log\left(\frac{\sigma}{\sigma - 1}\frac{1}{\xi(a)}\frac{p}{\mathcal{D}}\right)\right)^{\sigma/\omega} \frac{Y}{\xi(a)}$$

And the desired markup can be calculated to be equal to the following expression:

$$\mu \equiv \frac{\alpha p}{W y^{\frac{1}{\alpha}-1}} = \frac{\sigma\left(\frac{y}{Y}\xi(a)\right)^{-\omega/\sigma}}{\sigma\left(\frac{y}{Y}\xi(a)\right)^{-\omega/\sigma} - 1}$$

Taking the total derivative of the demand function, after the variable substitution of prices gives:

$$d\log(y) = d\log(Y) - d\log(\xi(a)) + \frac{\sigma}{\omega}\left(\frac{y}{Y}\xi(a)\right)^{-\frac{\omega}{\sigma}} d\left(1 - \omega\log\left(\frac{\sigma}{\sigma-1}\frac{\mu W y^{\frac{1}{\alpha}-1}}{\alpha\xi(a)\mathcal{D}}\right)\right) \tag{57}$$

where:

$$d\left(1 - \omega\log\left(\frac{\sigma}{\sigma-1}\frac{\mu W y^{\frac{1}{\alpha}-1}}{\alpha\xi(a)\mathcal{D}}\right)\right) = \tag{58}$$

$$-\omega\left(d\log(\mu) + d\log(W) + \left(\frac{1}{\alpha} - 1\right) d\log(y) - d\log(\xi(a)) - d\log(\mathcal{D})\right) \tag{59}$$

Taking instead the total derivative of firm markup leads to the following result:

$$d\log(\mu) = \frac{d\log\left(\sigma\left(\frac{y}{Y}\xi(a)\right)^{-\omega/\sigma}\right)}{\sigma\left(\frac{y}{Y}\xi(a)\right)^{-\omega/\sigma}-1} = -\frac{\omega}{\sigma}\frac{d\log(y)+d\log(\xi(a))-d\log(Y)}{\sigma\left(\frac{y}{Y}\xi(a)\right)^{-\omega/\sigma}-1} \tag{60}$$

which also implies:

$$d\log(y)+d\log(\xi(a))-d\log(Y) = -\frac{\sigma}{\omega}\left(\sigma\left(\frac{y}{Y}\xi(a)\right)^{-\omega/\sigma}-1\right)d\log(\mu) \tag{61}$$

Substituting this latter expression into the total derivative of the demand function we then obtain:

$$d\log(\mu)\left(\frac{\sigma}{\omega\mu}-1\right) = d\log(W)+\left(\frac{1}{\alpha}-1\right)d\log(y)-d\log(\xi(a))-d\log(\mathcal{D}) \tag{62}$$

To keep the algebra tractable here and for the sake of illustrating the channels at play in a much clearer way, we further make the simplifying assumption that $\alpha = 1$ (not in the quantitative solution though). We can then conclude that markups depend on the demand in the following way:

$$\frac{\partial\log(\mu)}{\partial\log(\xi(a))} = \frac{\frac{\omega}{\sigma}\mu}{\frac{\omega}{\sigma}\mu-1} \tag{63}$$

The dependence of the markup progression on the cost-push shock is instead given by:

$$d\log\left(\frac{\partial\log(\mu)}{\partial\log(\xi(a))}\right) = \frac{d\log(\mu)}{\frac{\omega}{\sigma}\mu-1} = -\frac{d\log(W)}{\frac{\omega}{\sigma}\mu-1}\frac{\frac{\omega}{\sigma}\mu}{\frac{\omega}{\sigma}\mu-1} \tag{64}$$

which can be rearranged to obtain the following expression:

$$\frac{\partial\log\left(\frac{\partial\log(\mu)}{\partial\log(\xi(a))}\right)}{\partial\log(W)} = -\frac{\frac{\omega}{\sigma}\mu}{\left(\frac{\omega}{\sigma}\mu-1\right)^2} \tag{65}$$

Finally the last derivative with respect to the measure of superelasticity $\frac{\omega}{\sigma}$ gives us:

$$\frac{d}{d\frac{\omega}{\sigma}}\frac{\partial\log\left(\frac{\partial\log(\mu)}{\partial\log(\xi(a))}\right)}{\partial\log(W)} = -\frac{d}{d\frac{\omega}{\sigma}}\frac{\frac{\omega}{\sigma}\mu}{\left(\frac{\omega}{\sigma}\mu-1\right)^2} = \mu\frac{\left(\frac{\omega}{\sigma}\mu\right)^2-1}{\left(\frac{\omega}{\sigma}\mu-1\right)^4} \tag{66}$$

We conclude by highlighting the intuition behind these results. Considering the case when $\frac{\omega}{\sigma} > 1$:
1) Markups increase with demand:

$$\frac{\partial\log(\mu)}{\partial\log(\xi(a))} = \frac{\frac{\omega}{\sigma}\mu}{\frac{\omega}{\sigma}\mu-1} > 0 \tag{67}$$

2) A reduction in production costs (contractionary MP shock) increases the dispersion of markups:

$$\frac{\partial \log\left(\frac{\partial \log(\mu)}{\partial \log(\xi(a))}\right)}{\partial \log(W)} = -\frac{\frac{\omega}{\sigma}\mu}{\left(\frac{\omega}{\sigma}\mu - 1\right)^2} < 0 \tag{68}$$

3) The increase in dispersion due to the contractionary MP shock decreases with the superelasticity:

$$\frac{d}{d\frac{\omega}{\sigma}} \frac{\partial \log\left(\frac{\partial \log(\mu)}{\partial \log(\xi(a))}\right)}{\partial \log(W)} = \mu \frac{\left(\frac{\omega}{\sigma}\mu\right)^2 - 1}{\left(\frac{\omega}{\sigma}\mu - 1\right)^4} > 0 \tag{69}$$

In other words, the higher is the fraction $\frac{\omega}{\sigma}$, the less negative is the second derivative, so the smaller the difference in response to a contractionary MP shock.

B Empirical Appendix

B.1 Data Appendix

Following standard practice in the literature and to ensure that the firms in the Compustat sample have as the benchmark interest rate the interest rate set by the FED, we restrict our attention to firms that are incorporated in the US. We exclude utilities (SIC codes between 4900 and 4999) because they are heavily regulated on prices. We also exclude financial firms (SIC codes between 6000 and 6999) because their balance sheets are extremely different from other firms.

Furthermore, we drop all the observation with missing industry classification, as well as all those observations with negative or missing sales (`SALEQ`) and cost of goods sold (`COGSQ`). Whenever applicable, we deflate variables using a GDP deflator from the NIPA tables. Table B.1 reports summary statistics for the variables of interest.

Table B.1: Summary Statistics

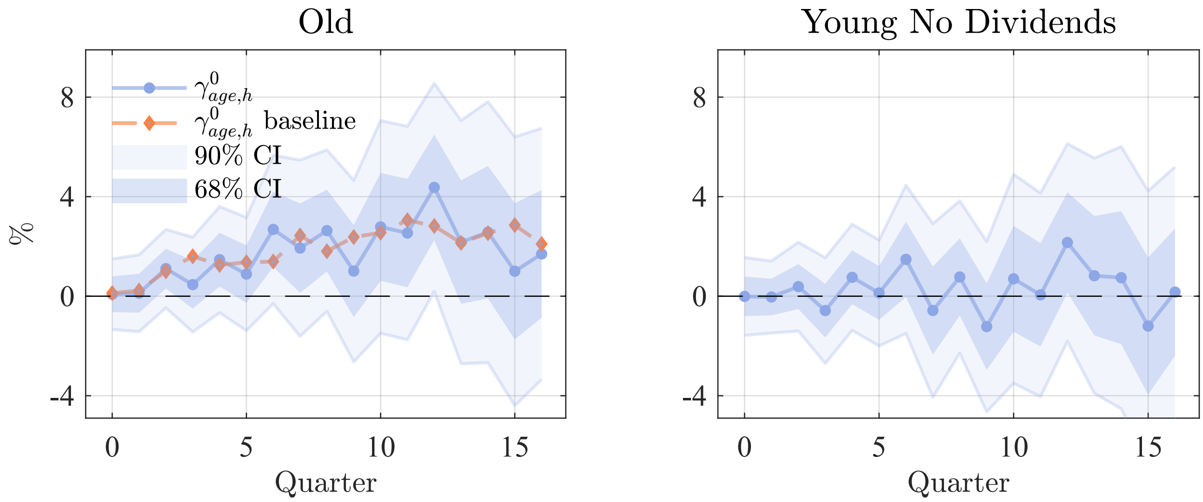
	Sales	Cogs	Assets	Leverage	Liquidity	Age
mean	447.69	303.17	4919.69	0.45	0.17	9.46
p25	6.06	3.31	37.83	0.04	0.02	4
p50	31.01	17.18	229.50	0.18	0.07	8
p75	164.58	100.60	1118.33	0.39	0.22	14
N	715,874	715,874	685,784	641,316	683,696	715,874

Note. Summary statistics of cleaned quarterly Compustat dataset between 1990q1 and 2016q4. Sales, Cogs, and Assets are measured in millions of real 2012 US$, while Leverage and Liquidity are ratios and Age is measured in years.

B.2 Financial Frictions Appendix

This section builds on the insights from Cloyne, Ferreira, Froemel, and Surico (2018), suggesting that young firms are financially constrained if they do not pay dividends. The underlying rationale is that a firm paying dividends should not face financial constraints; otherwise, it would first cut these payments to finance its operations. Following this logic, we split the group of young firms into those that pay dividends and those that do not. The subset of young firms paying dividends is expected to identify financially unconstrained ones, while the others represent financially constrained firms. While this approach seems straightforward, it places a heavy burden on the data, as dividend payments are full of missing observations in Compustat. Additionally, the group of young firms paying dividends is quite small given the rarity of positive dividend payments in the data, making the estimation procedure demanding with a limited number of observations.

Figure B.1: Relative Response of Markups to a MP Shock for Firms of Different Categories



Note. Figure B.1 illustrates the relative response of markups to a monetary policy shock for old firms and young firms that do not pay dividends compared to young firms that do pay dividends. The estimates are derived from the estimation of Equation (26) for $h = 1, ..., 16$, normalized to a 25 basis points contractionary monetary policy shock. The solid dark blue line with circles represents the point estimates of $\gamma^0_{x,h}$, while the dark blue and light blue areas depict the 68% and 90% confidence intervals, respectively. The dashed orange line with diamonds represents the point estimates of $\gamma^0_{x,h}$ from the baseline specification in Section 2.

Figure B.1 depicts the relative response of markups to a MP shock for old and young firms that do not pay dividends, compared to young firms that do pay dividends. It illustrates the evolution of the coefficient from the estimation of Equation (26) for $h = 1, ..., 16$. The baseline for comparison is the group of young firms that do pay dividends, representing financially unconstrained young firms. This group serves as a benchmark to assess whether age plays a role beyond financial frictions – when such group is compared to old firms –, or if financial frictions are the main driver of the observed heterogeneity – when the group is compared to financially constrained young firms.

Several observations are noteworthy. First, the IRFs are less precisely estimated, which can be attributed to the sparsity of dividends in the data, as previously discussed. Second, there is a statistically significant difference (both at the 90% CI and particularly at the 68% CI) between the markup responses of old firms and young firms paying dividends. This difference aligns in the same direction and exhibits a magnitude similar to our benchmark results, indicating that age, as a proxy for firms' dominance, plays a role independently, without being solely mediated by financial frictions. Third, we find no statistically significant difference (both at the 90% CI and at the 68% CI) between young firms that do or do not pay dividends, suggesting that variations in the degree of financial frictions do not play a major role in their markup response to monetary policy shocks.

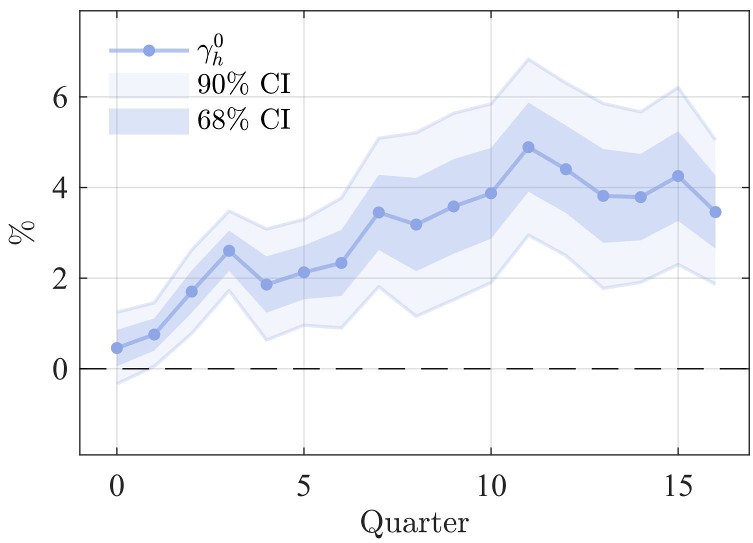
While these results lend support to using firm age as a proxy for dominance, we observe that the IRFs are somewhat more muted than in the baseline specification. Therefore, while we can conclude that our mechanism is supported, we cannot dismiss the possibility that the noisier estimates in this section are, in part, due to the role played by financial frictions, not just data sparsity. Acknowledging that financial frictions may indeed play a role, we emphasize that the role of age goes above and beyond these channels and, at least in our data, appears to be quantitatively stronger.

B.3 Robustness Appendix

B.3.1 Excluding the ZLB

Here, we estimate Equation (26) over the period between 1990q1 and 2008q4, which excludes the ZLB. We do this for several reasons: (i) over the ZLB period, the measure of MP shocks from Jarociński and Karadi (2020) exhibits nearly no variation and (ii) over the ZLB period, conventional monetary policy went in second place among central banks in favor of new unconventional monetary policy practices. Both reasons suggest that the ZLB period could be atypical and we want to ensure that this is not what is driving our main conclusions. Figure B.2 shows the result. Overall, we notice that excluding the ZLB from the period of analysis does not affect our conclusions.

Figure B.2: Relative Response of Old Firms' Markups to a Monetary Policy Shock—No ZLB



Note. Figure B.2 shows the relative response of old firms' markups to a monetary policy shock (as compared to young firms). In particular, it reports the evolution of the coefficient $\gamma^0_{age,h}$ from the estimation of Equation (26) for $h = 1, ..., 16$. The figure is normalized to a 25 basis points contractionary monetary policy shock. The solid dark blue line with circles reports the point estimates of $\gamma^0_{age,h}$. The dark blue and light blue areas report the 68% and the 90% confidence intervals of our estimates.

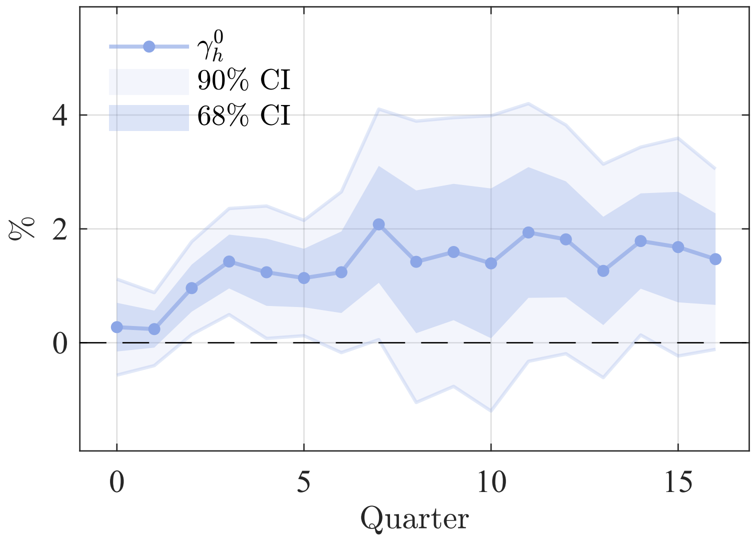
B.3.2 Excluding Future Shocks

Here, we estimate an alternative specification compared to Equation (26), given by:

$$\Delta_h \log \mu_{i,t+h} = \sum\nolimits_{x\in\mathcal{X}} \left(\alpha_{x,h} + \beta_{x,h}\Delta Y_{t-1} + \sum\nolimits_{k=0}^{\kappa} \gamma^k_{x,h}\varepsilon^m_{t-k} \right) \times \mathbb{1}_{i\in\mathcal{I}^x} \\ + \sum\nolimits_{\ell=1}^{L} \boldsymbol{\delta}'_h \boldsymbol{X}_{i,t-\ell} + \varphi_{i,h} + \varphi_{s,t,h} + \boldsymbol{\vartheta}_h \boldsymbol{t} + u_{i,t+h}, \tag{70}$$

with horizons $h = 0, 1, \ldots, H$. Equation (70) is similar in spirit to Equation (26) but it does not include future monetary policy shocks. Figure B.3 shows the result. Overall, we notice that excluding future monetary policy shocks from Equation (26) does not affect our results but it only increases the downward bias present in LP models, as explained by Teulings and Zubanov (2014).

Figure B.3: Relative Response of Old Firms' Markups to a MP Shock—No Future Shocks



Note. Figure B.3 shows the relative response of old firms' markups to a monetary policy shock (as compared to young firms). In particular, it reports the evolution of the coefficient $\gamma^0_{age,h}$ from the estimation of Equation (70) for $h = 1, \ldots, 16$. The figure is normalized to a 25 basis points contractionary monetary policy shock. The solid dark blue line with circles reports the point estimates of $\gamma^0_{age,h}$. The dark blue and light blue areas report the 68% and the 90% confidence intervals of our estimates.

B.3.3 Linear Parametric Interaction

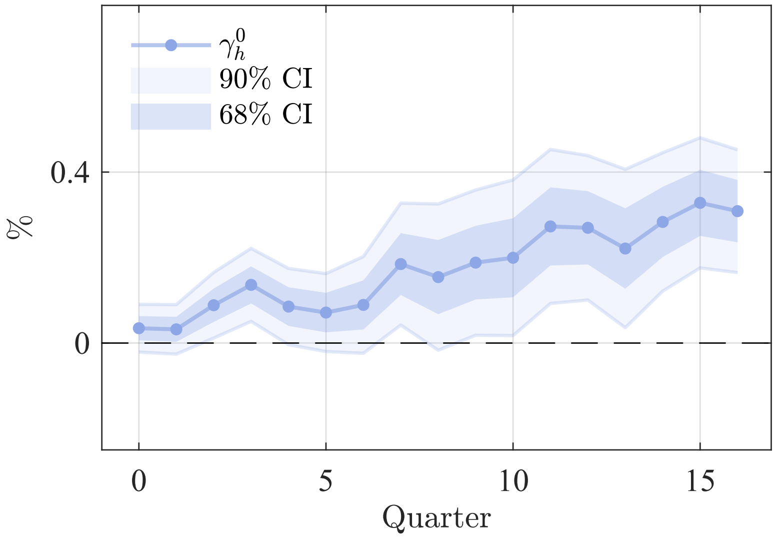
Here, we estimate an alternative specification compared to Equation (26), given by:

$$\Delta_h \log \mu_{i,t+h} = \sum_{x \in \mathcal{X}} \left(\alpha_{x,h} + \beta_{x,h} \Delta Y_{t-1} + \sum_{k=-\kappa}^{h} \gamma^k_{x,h} \varepsilon^m_{t+k} \right) \times x_t + \sum_{\ell=1}^{L} \delta'_h \boldsymbol{X}_{i,t-\ell} + \varphi_{i,h} + \varphi_{s,t,h} + \boldsymbol{\vartheta}_h \boldsymbol{t} + u_{i,t+h}, \tag{71}$$

with horizons $h = 0, 1, \ldots, H$. Equation (71) is similar in spirit to Equation (26) but it uses a linear parametric approach in the variables capturing age-driven heterogeneity, as done by Ottonello and Winberry (2020).

Figure B.4 shows the result. Overall, we notice that Equation (71) estimates a positive linear interaction between the monetary policy shock and the age variable, suggesting that the older a firm is, the more its markup responds to monetary policy shocks, as in our benchmark specification.

Figure B.4: Relative Response of Old Firms' Markups to a MP Shock—Linear Interaction

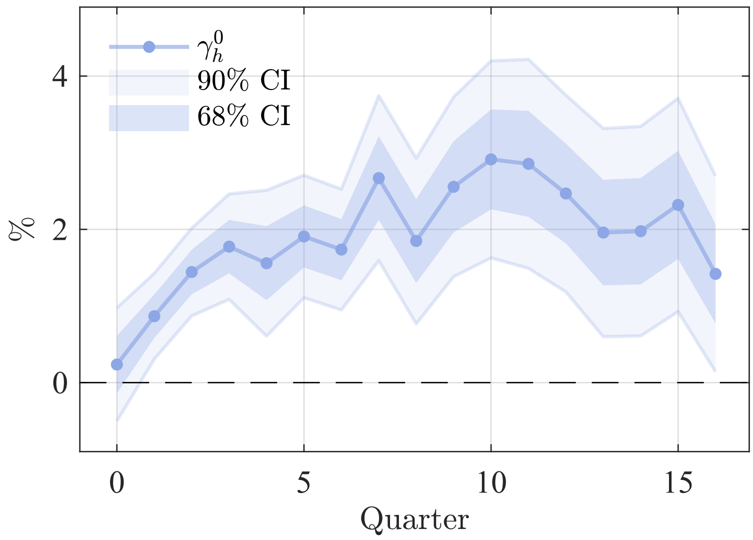


Note. Figure B.4 shows the relative response of old firms' markups to a monetary policy shock (as compared to young firms). In particular, it reports the evolution of the coefficient $\gamma^0_{age,h}$ from the estimation of Equation (71) for $h = 1, ..., 16$. The figure is normalized to a 25 basis points contractionary monetary policy shock. The solid dark blue line with circles reports the point estimates of $\gamma^0_{age,h}$. The dark blue and light blue areas report the 68% and the 90% confidence intervals of our estimates.

B.3.4 Monetary Policy Shock from Gürkaynak, Sack, and Swanson (2005)

Here, we estimate Equation (26) using an alternative monetary policy shock measure from Gürkaynak, Sack, and Swanson (2005). Their series is similar compared to the Jarociński and Karadi (2020) but it does not take into account the information channel of monetary policy.

Figure B.5: Relative Response of Old Firms' Markups to a MP Shock—GSS Shock



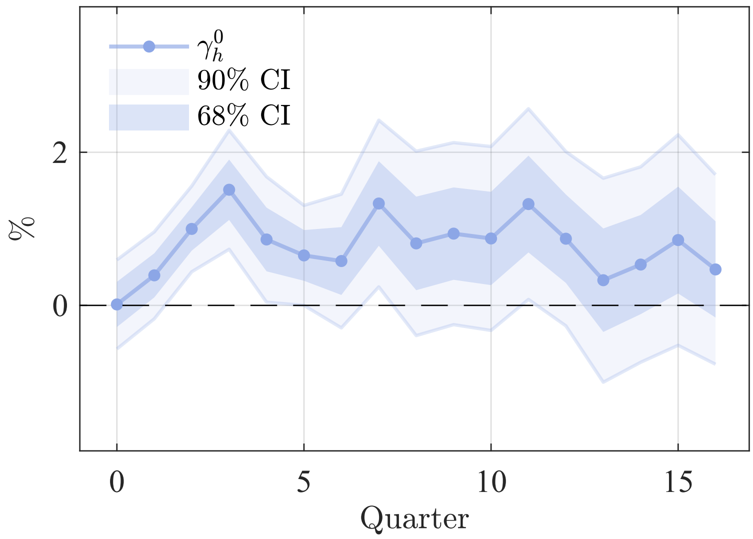
Note. Figure B.5 shows the relative response of old firms' markups to a monetary policy shock from Gürkaynak, Sack, and Swanson (2005) (as compared to young firms). In particular, it reports the evolution of the coefficient $\gamma^0_{age,h}$ from the estimation of Equation (26) for $h = 1, ..., 16$. The figure is normalized to a 25 basis points contractionary monetary policy shock. The solid dark blue line with circles reports the point estimates of $\gamma^0_{age,h}$. The dark blue and light blue areas report the 68% and the 90% confidence intervals of our estimates.

Figure B.5 shows the result. Overall, we see that using the alternative monetary policy shock proposed by Gürkaynak, Sack, and Swanson (2005) produces an IRF that is both qualitatively and quantitatively very close to the one in our benchmark result, showing a 2% stronger markups response for old firms (as compared to young ones) to monetary policy shocks.

B.3.5 Grouping Firms by Sector and Quarter

We also estimate Equation (26) using a different definition for $\mathbb{1}_{i \in \mathcal{I}^x}$. In the main analysis, we define firms' categories by being above or below the median of $x \in \mathcal{X} = \{$*sales share*, *leverage*, *liquidity*, *assets*$\}$, considering the entire sample and according to the previous year. Here, we instead define firms' categories by being above or below the median of those same variables but within a given sector and a given quarter of the previous year. Figure B.6 shows the result. Overall, using a different definition of $\mathbb{1}_{i \in \mathcal{I}^x}$ does not significantly change our conclusions, suggesting that our results do not depend on the particular definition of the dummies capturing firm-level heterogeneity.

Figure B.6: Relative Response of Old Firms' Markups to a MP Shock—Grouping by Sector/Quarter

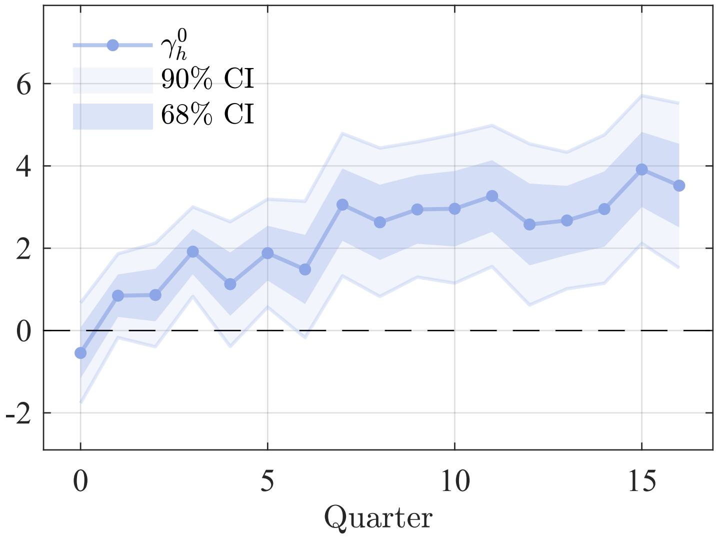


Note. Figure B.6 shows the relative response of old firms' markups to a monetary policy shock (as compared to young firms). In particular, it reports the evolution of the coefficient $\gamma^0_{age,h}$ from the estimation of Equation (26) for $h = 1, ..., 16$. The figure is normalized to a 25 basis points contractionary monetary policy shock. The solid dark blue line with circles reports the point estimates of $\gamma^0_{age,h}$. The dark blue and light blue areas report the 68% and the 90% confidence intervals of our estimates.

B.3.6 Using Founding Age by Jay Ritter

Here, we estimate Equation (26) using the "true" founding age of firms compiled by Jay Ritter instead of corporate age, which is available for a subset of firms in the Compustat sample (about 20% of the observations approximately, hence a significant empirical limitation if we were to use founding age only). Results are shown in Figure B.7. Overall, we see that using this alternative measure of firms' age, which capture firms' founding age, does not alter our main conclusions.

Figure B.7: Relative Response of Old Firms' Markups to a Monetary Policy Shock—Founding Age



Note. Figure B.7 shows the relative response of old firms' markups to a monetary policy shock (as compared to young firms). In particular, it reports the evolution of the coefficient $\gamma^0_{age,h}$ from the estimation of Equation (26) for $h = 1, ..., 16$. The figure is normalized to a 25 basis points contractionary monetary policy shock. The solid dark blue line with circles reports the point estimates of $\gamma^0_{age,h}$. The dark blue and light blue areas report the 68% and the 90% confidence intervals of our estimates.

B.3.7 Using Alternative Elasticities

Here, we estimate Equation (26) using markups calculated with different production function elasticities. In particular, instead of using the elasticities provided by De Loecker, Eeckhout, and Unger (2020), which are common for all the quarters within a year, we calculate our elasticities allowing them to vary at a quarterly level. In particular, we estimate (i) a Cobb-Douglas production function which varies at quarterly and sector levels, and (ii) a Translog production function which also varies at quarterly and sector levels. This second specification also allows us to have production function elasticities that are heterogenous across firms within sectors and quarters.[15]

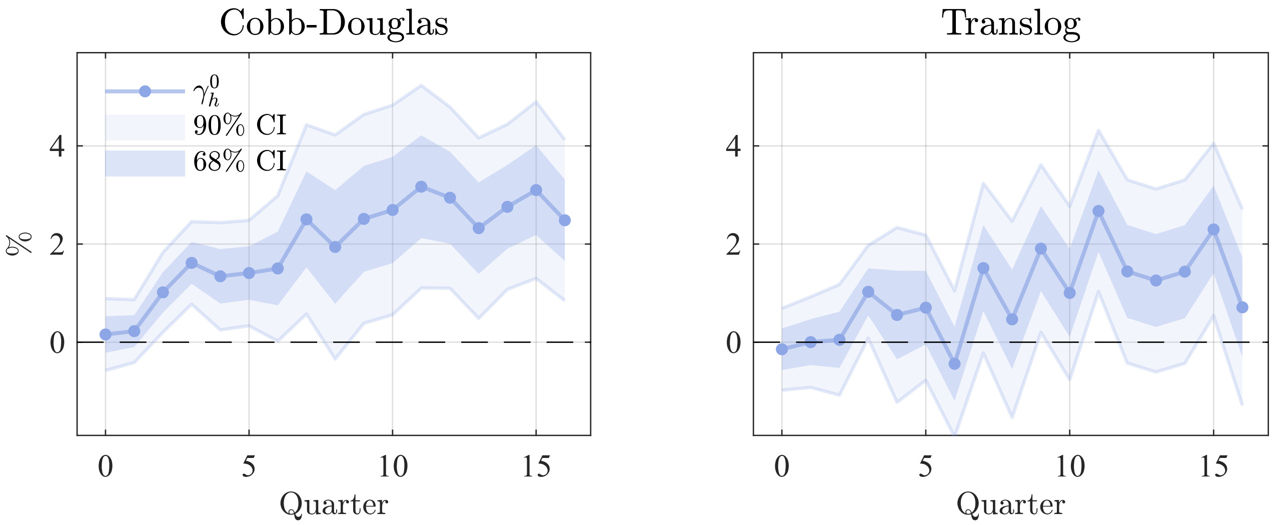
Results are shown in Figure B.8. Both specifications show similar patterns among themselves and compared to our benchmark. This suggests that our results are not driven mainly by heterogeneity in firm-level elasticities but by heterogeneity in the firm-level ratio of sales to variable costs. Different firm-level patterns in production function elasticities do not affect our results.

B.3.8 Using Alternative Markups Measure

Here, we estimate Equation (26) using markups calculated with a different methodology compared to those presented in Equation (25). In particular, we follow the methodology used by Gutiérrez

[15] The elasticity coming from the Translog production function is a function of firm-level capital stock (`PPEGT` in Compustat). For data quality, we interpolate between adjacent points our estimates whenever this item is missing. We check that this does not affect results, although it improves precision.

Figure B.8: Relative Response of Old Firms' Markups to a MP Shock—Alternative Elasticities



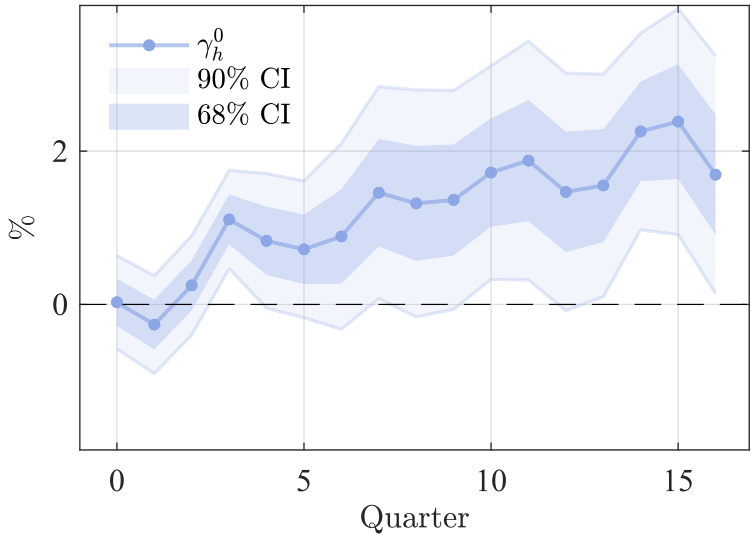
Note. Figure B.8 shows the relative response of old firms' markups to a monetary policy shock (as compared to young firms) when firm-level markups are calculated using alternative production function specifications as explained in the text. In particular, it reports the evolution of the coefficient $\gamma^0_{age,h}$ from the estimation of Equation (26) for $h = 1, ..., 16$. The figure is normalized to a 25 basis points contractionary monetary policy shock. The solid dark blue line with circles reports the point estimates of $\gamma^0_{age,h}$. The dark blue and light blue areas report the 68% and the 90% confidence intervals of our estimates.

and Philippon (2016) and Baqaee and Farhi (2020) given by:

$$\mu = \frac{1}{1-\ell} \quad \text{and} \quad \ell = \frac{\texttt{OIBDPQ} - \texttt{DPQ}}{\texttt{SALEQ}}, \tag{72}$$

where $\texttt{OIBDPQ} - \texttt{DPQ}$ is operating income before depreciation net of depreciation.

Figure B.9: Relative Response of Old Firms' Markups to a MP Shock—Alternative Measure



Note. Figure B.9 shows the relative response of old firms' markups to a monetary policy shock (as compared to young firms). In particular, it reports the evolution of the coefficient $\gamma^0_{age,h}$ from the estimation of Equation (26) for $h = 1, ..., 16$. The figure is normalized to a 25 basis points contractionary monetary policy shock. The solid dark blue line with circles reports the point estimates of $\gamma^0_{age,h}$. The dark blue and light blue areas report the 68% and the 90% confidence intervals of our estimates.

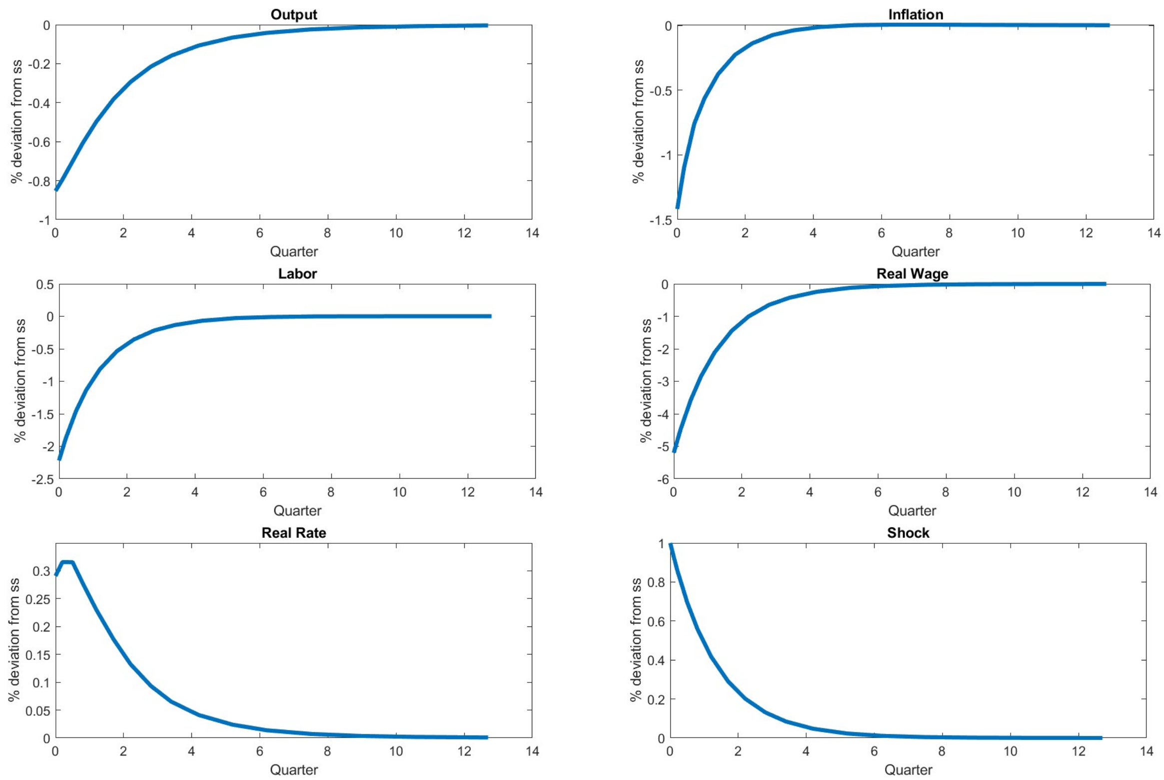
Results are shown in Figure B.9. Overall, we see that using this alternative measure of market power, which should be considered second-best compared to our main measure, does not affect the qualitative behavior of our results, although it dampens slightly the magnitude compared to our benchmark estimates from the main analysis.

C Quantitative Appendix

Figure C.10 shows the aggregate impulse response functions of output, inflation, nominal and real interest rates and the real wage to a transitory and contractionary monetary policy shock in the model economy. The shock captures a negative change of 25 b.p. to ε_t^m, which then transmits to the nominal interest rate i_t through the Taylor Rule discussed in Section 2 and given by:

$$i_t = \rho + \phi_\pi(\pi_t - \pi^*) + \varepsilon_t^m,$$

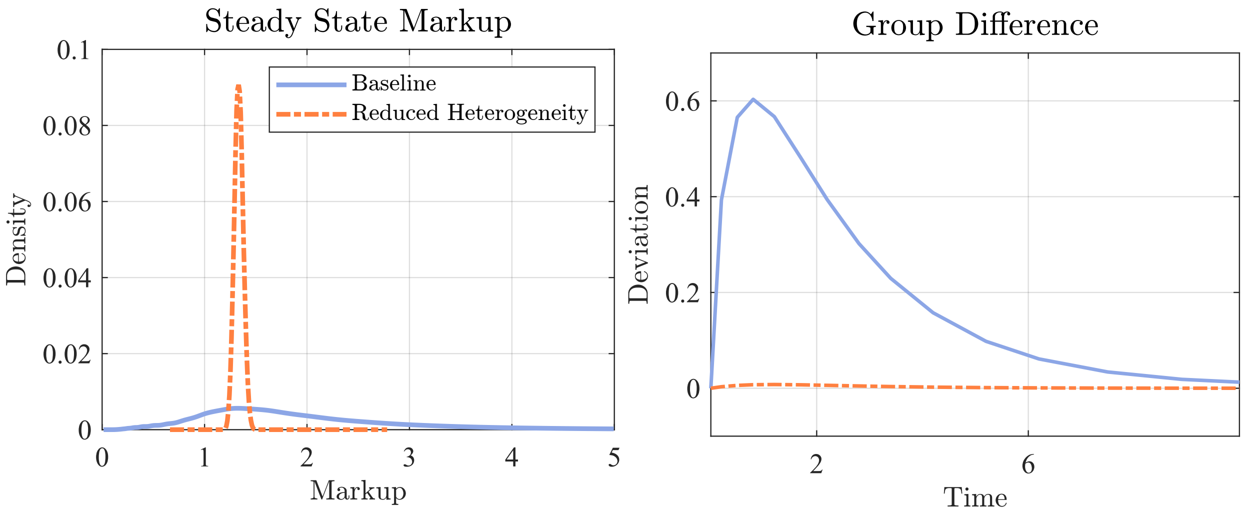
Figure C.10: Aggregate Impulse Response Functions



Note. Figure C.10 shows the aggregate impulse response functions of output, inflation, labor, the real interest rate and the real wage to a transitory and contractionary monetary policy shock in the model economy.

Figure C.11 shows the difference in markups' behaviour in the model under the baseline calibration and an arbitrary calibration with reduced heterogeneity. Specifically, we modify the demand process faced by firms so that there is no implied growth over their life-cycle, meaning $\bar{a} = 0$, and the standard deviation of shocks to the demand process is significantly reduced to $\psi^a = 0.01$.

Figure C.11: Markups under Reduced Heterogeneity



Note. Figure C.11 shows on the left the distribution of markups in the steady state for the baseline calibration and calibration with reduced heterogeneity. And on the right the differential impulse response of markups for old and young firms to a transitory and contractionary monetary policy shock in the model economy with baseline calibration versus the calibration with reduced heterogeneity.