Endogenous Firm Competition and the Cyclicality of Markups*

Hassan Afrouzi†
Columbia University

Luigi Caloi‡
Columbia University

April 4, 2021

Abstract

We show that the cyclicality of *output growth* is a sufficient predictor for the cyclicality of markups in a large class of models that micro-found variable markups through dynamic trade-offs. First, we show that this class of models imply a unified law of motion for markups in which current markup depends on firms' expectations over future *changes* in demand. Second, we use data on markups from the U.S. as well as survey data on firms' expectations from New Zealand to test the predictions of these models and find evidence in favor of their mechanisms. Finally, we study the implications of this law of motion for cyclicality of markups in a calibrated general equilibrium model. In particular, we find that the degree of hump-shaped response in output is crucial for the direction of aggregate markup cyclicality. To highlight the importance of this mechanism, we show that the model with a hump-shaped response of output matches the evidence on cyclicality of markups conditional on TFP shocks, while the model without the hump-shaped response of output completely misses the direction of this cyclicality.

JEL Codes: E3
Key Words: markup cyclicality, implicit collusion models, customer-base models

*We are grateful to the editor and three anonymous referees for their thoughtful comments and suggestions. We also thank Olivier Coibion, Saroj Bhattarai, Andrew Glover, Matthias Kehrig, as well as seminar participants at UT Austin, Midwest Macro Meeting, and Federal Reserve Banks of Dallas and Kansas City.

†Columbia University, Economics Department, 420 West 118 St. New York, NY 10027. ha2475@columbia.edu.

‡Columbia University, Economics Department, 420 West 118 St. New York, NY 10027. luigi.caloi@columbia.edu.

1 Introduction

The cyclicality of markups is a key transmission mechanism in the business cycle literature. Countercyclical markups, for instance, are broadly viewed as a propagation mechanism within New Keynesian models (Christiano, Eichenbaum, and Rebelo, 2011), as well as a potential reason for positive comovement between hours and wages (Rotemberg and Woodford, 1992). Nonetheless, theories that micro-found variable markups have conflicting predictions regarding their cyclical behavior.[1]

In this paper, we provide a unified law of motion for markups in a broad class of models that micro-found variable markups through *dynamic trade-offs*—namely implicit collusion and customer-base models—and show that cyclicality of *output growth* is a sufficient predictor for the cyclicality of markups in these models.

We formally show that, up to a first-order approximation, both implicit collusion and customer-base models yield the same reduced form expression for the dynamics of markups. Specifically, at the firm level, current markups depend on the net present value of all expected sales growths in the future. Once aggregated, the net present value of future sales collapses to the net present value of output growth in the economy. Therefore, the two models relate current markups to expected output growths in the future. They differ, however, in terms of the sign restrictions that they imply for this reduced-form representation of markup dynamics. Hence, one can test and differentiate between these two models by estimating the common law of motion for markups and comparing the empirical estimates with the sign restrictions implied by each model.

Our second contribution is to use data from Compustat to test these predictions for the U.S. To do so, we estimate firm-level markups following the methodology of De Loecker, Eeckhout, and Unger (2020) and subsequently estimate the law of motion for markups. We find this evidence to be consistent with implicit collusion models. First, we find markups comove positively with firms' future sales (using lagged values of sales and markups as instruments for firms' expectations of future sales growth and markups). Second, to further test the implications of these two models, we also study the heterogeneity of these results in the cross-section of firms.

Both models imply that the magnitude of the relationship between markups and future sales growths should be increasing in firms' relative size. We find this prediction consistent with the data: the coefficient on future sales growth is statistically significant and larger than the benchmark estimate for firms above the median of relative sales. In contrast, the point estimate in the below-median group is smaller and statistically insignificant. While this result supports

[1] In particular, two common micro-foundations—implicit collusion and customer-base models—yield different cyclicalities: implicit collusion models are interpreted as implying countercyclical markups, while customer-base models have been used to generate both procyclical and countercyclical markups.

the predictions from theory, it raises the question of which group is more representative of an aggregate economy? To answer this, we compute the sales share of each group in the data. We find that while the two samples are comparable in the number of observations, the sales share of the above-median group is 98% of total sales due to the highly skewed distribution of sales in Compustat.

Another prediction of the two models is that the relationship between markups and future sales growths should be larger for firms that value future profits more. Using insights from previous work on financial frictions (Gilchrist, Schoenle, Sim, and Zakrajšek, 2017), we hypothesize that firms with a higher debt-to-asset ratio should value future profits less relative to current profits due to a higher preference for liquidity in short-run. Therefore, according to the models, we should see a weaker relationship between markups and expected future sales growths among more leveraged firms. We find this prediction to be consistent with the data: when we re-estimate the law of motion among firms in the upper quartile of debt-to-asset ratio, we find no significant relationship between markups and future sales growths.

Finally, to examine the external validity of these findings outside of the Compustat, we estimate the law of motion for markups using survey data on firms' expectations from New Zealand, introduced by Coibion, Gorodnichenko, and Kumar (2018). This approach also allows us to directly rely on data on expectations rather than the actual realization of sales. Consistent with the findings in the Compustat sample, the results favor the implicit collusion model. In particular, we find a positive and significant relationship between markups and expectations of future sales among firms with fewer competitors.

Therefore, both sets of empirical evidence are consistent with the predictions of implicit collusion models but at odds with customer-base models.

Our final contribution is to calibrate a general equilibrium model with implicit collusion and study the implications of these findings for the cyclicality of markups. We find that depending on how hump-shaped the response of output is, markups can either be procyclical or countercyclical in these models. The reason is that the implicit collusion model relates markup cyclicality to output growth rather than output itself. Hence, if output growth is expected to be positive in an expansion (due to a hump-shaped response), then markups increase on impact and covary positively with output. However, if there is no hump-shape in output response, then after an expansionary shock where output jumps to a high level and is expected to fall back to its steady-state after that, output growth is expected to be negative, and markups decrease in expansions.

Thus, accounting for the hump-shaped response of output to shocks, as observed in the empirical literature,[2] is instrumental for the cyclicality of markups. We illustrate the importance

[2]See, e.g., Ramey (2011); Ramey and Shapiro (1998); Monacelli and Perotti (2008) for government spending shocks, Sims (2011); Smets and Wouters (2007) for productivity shocks, and Christiano, Eichenbaum, and Evans

of this mechanism by examining the cyclicality of markups conditional on TFP shocks in the model and relating it to the empirical estimates in Nekarda and Ramey (2020), who find that markups are procyclical conditional on TFP shocks. First, we show that a model without a hump-shaped response of output delivers the wrong cyclicality: when productivity goes up, and output rises on impact, output growth is expected to be negative, and markups fall, leading to countercyclical markups conditional on TFP shocks. However, once we allow for a hump-shaped response for output (using investment adjustment costs), we find that markups become procyclical conditional on TFP shocks and match the conditional correlation of markups and output relatively well as non-targeted moments.

Literature Review. Both implicit collusion and customer-base models are used within macroeconomic models to study the markup setting behavior of firms. In our analysis, we start by building the firm side of the implicit collusion model, and show that markups are determined by the joint distribution of expected growth of output and stochastic discount rates. This allows us to reconcile the seemingly contradictory predictions of these models in a unified framework. For instance, Rotemberg and Saloner (1986) assume that demand shocks are i.i.d., implicitly implying that the expected demand growth is countercyclical, and conclude that markups are countercyclical. On the other hand, Kandori (1991); Haltiwanger and Harrington Jr (1991); Bagwell and Staiger (1997), each by assuming alternative processes for demand shocks find that these models can produce procyclical markups.

Rotemberg and Woodford (1991, 1992) are the first to study the implicit collusion model within a DSGE model. Contrary to the partial equilibrium models, their general equilibrium setting endogenously pins down the joint distribution of output growth and stochastic discount rates, which gives rise to countercyclical markups; however, their result is not robust to the structure of the shocks and is reversed by introducing a hump-shaped response for output.

Phelps and Winter (1970) is the first paper that formalizes the idea for customer-base models. Various papers have used this idea to study the cyclical behavior of markups. Different versions of customer-base models have been shown to create either procyclical or countercyclical markups. For instance, by micro-founding the game between firms and customers, Paciello, Pozzi, and Trachter (2018) find that markups are procyclical, but Ravn, Schmitt-Grohe, and Uribe (2006) argue that they are countercyclical. In this paper, we do not take a stance on the micro-foundations of this friction.[3] Instead, using a simple customer-base model with an exogenous habit formation process on the side of customers, we show that markups can be either pro- or countercyclical depending on whether the response of output to shocks is hump-shaped or not.

(2005) for monetary policy shocks.

[3]There has been a tremendous amount of progress in recent years in micro-founding this friction using search and matching frameworks. See, e.g., Gourio and Rudanko (2014); Kaplan and Menzio (2016); Bornstein (2018).

Another class of models that generate variable markups use demand systems where the elasticity of demand varies with firms' size, either by directly assuming Marshall's second law of demand using the aggregator function as in Kimball (1995), or by assuming nested CES aggregators, as in Atkeson and Burstein (2008). A recent application of Kimball preferences is Edmond, Midrigan, and Xu (2018) who use this demand system to assess the cost of markups, whereas a recent application of the latter is Burstein, Carvalho, and Grassi (2020) who find that the cyclicality of markups depends on the level of aggregation. We discuss these models and their relationship to our framework further in Section 5.

Finally, New Keynesian models also have predictions for the cyclicality of markups due to their assumptions on price stickiness. We discuss the relationship between our paper and these models further in Section 5.

Outline. Section 2 introduces the firm side of implicit collusion and customer-base models and derives the unified law of motion for markups. Section 3 presents the evidence for this law of motion using Compustat data in the U.S. and survey data on firms' expectations from New Zealand. Section 4 discusses the implications of the law of motion for markups in a calibrated general equilibrium model. Section 5 discusses the relationship between our framework and other models of variable markups. Section 6 concludes.

2 A Unified Law of Motion for Markups

In this section, we revisit customer-base and implicit collusion model, both of which micro-found markups through dynamic trade-offs. Our results in this section are twofold. First, we show that both classes of models imply fundamentally similar laws of motion for markups, where markups depend on the net present value of future *sales growth* for firms. Second, we find that despite this similarity, each model associates a different sign to this relationship: in implicit collusion models markups move positively with future sales, but in customer-base models this relationship is negative. In the remainder of this section, we introduce the firm side of implicit collusion and customer-base models respectively, and derive a unified law of motion for markups in both.

2.1 Implicit Collusion Models

We follow Rotemberg and Woodford (1991, 1992) in setting up the implicit collusion model but we depart from their representation by deriving the law of motion for markups, and showing that markups depend on the net present value of future *sales growth*.

There is a final good of consumption in the economy which is produced using a large number of intermediate differentiated goods. There is a unit measure of intermediate good sectors

indexed by $i \in [0,1]$. In each sector, there are N identical firms producing differentiated goods, indexed by i, j where $j \in \{1,\ldots,N\}$.

2.1.1 The Final Good Producer

The final good producer takes the price of consumption good, P_t, as given and produces with

$$Y_t = \left[\int_0^1 Y_{i,t}^{\frac{\sigma-1}{\sigma}} di\right]^{\frac{\sigma}{\sigma-1}}, \quad Y_{i,t} = \Phi(Y_{,1},\ldots,Y_{i,N}) \equiv \left[N^{-\frac{1}{\eta}} \sum_{j=1}^{N} Y_{i,j,t}^{\frac{\eta-1}{\eta}}\right]^{\frac{\eta}{\eta-1}} \tag{1}$$

Therefore, σ is the elasticity of substitution across sectors, and η is the elasticity of substitution within sectors. The profit maximization problem of this firm leads to the following standard demand functions implied by nested CES preferences:

$$\frac{Y_{i,j,t}}{Y_t} = D(\frac{P_{i,j,t}}{P_t}; \frac{P_{i,-j,t}}{P_t}) \equiv \frac{1}{N}\left(\frac{P_{i,j,t}}{P_t}\right)^{-\eta}\left[\frac{1}{N}\sum_{k=1}^{N}\left(\frac{P_{i,k,t}}{P_t}\right)^{1-\eta}\right]^{\frac{\eta-\sigma}{1-\eta}} \tag{2}$$

where $P_t \equiv \left[\int_0^1 [N^{-1}\sum_{k=1}^{N} P_{i,k,t}^{1-\eta}]^{\frac{1-\sigma}{1-\eta}} di\right]^{\frac{1}{1-\sigma}}$ is the aggregate price level. This demand system leads to the following expression for the firm's elasticity of demand, where this elasticity is an average of σ and η weighted by the firm's market share (Atkeson and Burstein, 2008):

$$-\frac{\partial \log(Y_{i,j,t})}{\partial \log(P_{i,j,t})} = (1 - s_{i,j,t})\eta + s_{i,j,t}\sigma, \quad s_{i,j,t} \equiv \frac{P_{i,j,t}Y_{i,j,t}}{\sum_{k=1}^{N} P_{i,k,t}Y_{i,k,t}} \tag{3}$$

A general assumption in two layer CES models is that $\eta > \sigma > 1$. This assumption corresponds to the case that goods within sectors are closer substitutes than goods across sectors, and ensures that Marshall's second law of demand holds in relative terms (that demand is less elastic at higher relative quantity, or firms with higher market share have less elastic demand).

2.1.2 Intermediate Goods

We assume all N firms within each sector i are identical in their production technology and use capital and labor to produce with a Cobb-Douglas production function, $Y_{i,j,t} = Z_t^a K_{i,j,t}^{\alpha} L_{i,j,t}^{1-\alpha}$, where Z_t^a is an economy wide technology shock with an AR1 process:

$$\log(Z_t^a) = \rho_a \log(Z_{t-1}^a) + \sigma_a \varepsilon_{a,t}, \quad \varepsilon_{a,t} \sim \mathcal{N}(0,1)$$

Moreover, we assume factor markets are competitive and firms take the wages, W_t, as well as the rental rate of capital R_t as given.

The Repeated Game of Sector i. Let $Y_{i,t} \equiv \Phi(Y_{i,1,t},\ldots,Y_{i,N,t})$ denote the output of sector i at time t. Firms in any sector i take their demand functions in Equation (2) as given and discount future profits with a stochastic discount factor process, $\{\beta^t Q_{0,t} : t \geq 0\}$,[4] and play the following

[4] In general equilibrium, this process is determined by the relative marginal utilities of households in different states.

infinitely repeated game.

As in every super-game, this repeated game has many potential equilibria. Although there is no rigorous way to rank the multiple equilibria of this game, the literature on the implicit collusion models focuses on the equilibrium in which firms earn the highest possible profit stream subject to an incentive compatibility constraint that eliminates every firm's incentive to deviate from the equilibrium strategy of collusion.[5] Following this literature, we also assume that there exists a perfect monitoring system that detects any deviations with probability one. Therefore, the best cheating strategy for a firm is to best respond to collusion outputs of their rivals, knowing that it will trigger the punishment sub-game.[6]

Characterization of the Repeated Game Equilibrium. The equilibrium strategy is constructed as follows: at time 0 firms form the following contingent plan for all possible states in the future. For every single state at every point of time, every firm chooses a markup that yields the highest profit for the sector and is incentive compatible with collusion relative to the following punishment strategy: in case a firm deviates from the agreement, the game will enter a punishment stage where all firms will charge the static best response markup forever after. However, at each period there is a possibility that the industry will renegotiate this with probability $1-\gamma$ and will move back to the collusion stage. This probability γ is in fact pinning down the expected punishment length such that after a firm cheats, the industry expects to remain in punishment stage for an average of $1/(1-\gamma)$ periods.

Therefore, firms within every sector maximize the discounted value of the industry's life time profits such that no firm in no state has an incentive to cheat.[7] Note that incentive compatibility is the only restricting concern in this setting. Without it, firms would choose the monopoly markup for the industry at every state. However, a firm's incentive to cheat is at its highest level when the rest of the firms are committed to producing the monopoly output of the industry. This incentive declines as the markups of the other firms decrease towards the one in the static best-response equilibrium. Moreover, since firms choose their markups to be incentive compatible with collusion in every possible state, the game stays in the collusion sub-game forever from which no one has an incentive to deviate. Therefore, the proposed strategy is a sub-game perfect

[5] Nonetheless, such an equilibrium is not necessarily the equilibrium with the highest net present value of profits; it may be the case that occasional deviations yield higher profits compared to staying in collusion forever. Therefore, there might be an equilibrium with occasional collusion that dominates the best equilibrium in which firms always collude. We abstract from this case, following Rotemberg and Woodford (1991, 1992, 1999).

[6] In the absence of such a system, however, static best responding may not be the best cheating strategy for a firm. If small deviations were unnoticeable with some probability, characterizing the best strategy is nontrivial. For instance, in an environment with imperfect monitoring, Green and Porter (1984) characterize equilibria in which firms switch to punishment when their price falls below a trigger price, even if it is caused by a negative demand shock rather than a cheating competitor.

[7] Note that this does not require any explicit collusion among firms as each firm simply chooses the highest markup from which no one in the industry has an incentive to deviate.

Nash equilibrium.

Moreover, with a CRS production function, firms' capital to labor ratios are independent of their output level. This can be interpreted as firms having a constant marginal cost of production in a given period that is pinned down by factor prices and is independent of firms' level of production, or their total demand:

$$MC_t = \frac{1}{Z_t^a}\left(\frac{R_t}{\alpha}\right)^{\alpha}\left(\frac{W_t}{1-\alpha}\right)^{1-\alpha} \tag{4}$$

Therefore, given firms' price taking behavior in factor markets, what pins down firm's size and level of production is their market power: given their demand structure, a choice for prices determines firms' demand, which then implies a certain level of production to meet that demand.

Since we only focus on symmetric equilibria, in characterizing the strategy of a firm, we only consider strategies in which a firm's competitors all are charging the same markup which is going to be the collusion markup in the equilibrium. Therefore, firm i, j's profit from the action profile $\mu_i^t \equiv (\mu_{i,j,t}; \mu_{i,t})$, where $\mu_{i,t}$ is the collusion markup chosen by the industry and $\mu_t \equiv P_t / MC_t$ is the average markup in the economy, is given by

$$\Pi_{i,j,t}(\mu_{i,j,t}; \mu_{i,t}) = P_t Y_t \left(\frac{\mu_{i,j,t}-1}{\mu_{i,t}}\right)\left(\frac{\mu_{i,t}}{\mu_t}\right)^{1-\sigma} D\left(\frac{\mu_{i,j,t}}{\mu_{i,t}}; 1\right) \tag{5}$$

Since the model is real and all sectors are symmetric, henceforth we normalize aggregate price to one, $P_t \equiv 1$. The following Proposition formalizes the equilibrium.

Proposition 1. Each firm in sector i, maximizes its net present value of future profits subject to no other firm having an incentive to undercut them:

$$\begin{gathered} \max_{\{\mu_{i,t}\}_{t=0}^{\infty}} \tfrac{1}{N}\mathbb{E}_0 \textstyle\sum_{t=0}^{\infty} (\beta\gamma)^t Q_{0,t} Y_t (1-\tfrac{1}{\mu_{i,t}}) \mu_{i,t}^{1-\sigma} \mu_t^{\sigma-1} \\ s.t. \quad \max_{\rho_{i,t}} \left\{ (\rho_{i,t} - \tfrac{1}{\mu_{i,t}}) D(\rho_{i,t}; 1) \right\} - \tfrac{1}{N}(1-\tfrac{1}{\mu_{i,t}}) \le \beta\gamma \mathbb{E}_t Q_{t,t+1} \tfrac{Y_{t+1}}{Y_t} \left(\tfrac{\mu_{t+1}/\mu_{i,t+1}}{\mu_t/\mu_{i,t}} \right)^{\sigma-1} \Gamma_{i,t+1} \quad , \\ \Gamma_{i,t} \equiv \tfrac{1}{N}\left[(1-\tfrac{1}{\mu_{i,t}}) - \mu_s^{-\sigma}(\mu_s - 1)\mu_t^{\sigma-1} \right] + \beta\gamma \mathbb{E}_t Q_{t,t+1} \tfrac{Y_{t+1}}{Y_t} \left(\tfrac{\mu_{t+1}/\mu_{i,t+1}}{\mu_t/\mu_{i,t}} \right)^{\sigma-1} \Gamma_{i,t+1} \end{gathered} \tag{6}$$

where $\beta^\tau Q_{t,t+\tau}$ is the time t price of a claim that pays a unit of consumption at $t+\tau$, and $\mu_s \equiv \frac{(N-1)\eta+\sigma}{(N-1)\eta+\sigma-N}$ is the equilibrium markup of static best responding for firms at any state. $\eta > \sigma$ guarantees that $\frac{\eta}{\eta-1} \le \mu_s \le \frac{\sigma}{\sigma-1}$. The solution to this problem $\{\mu_{i,t}\}_{t=0}^{\infty}$ *exists, and it is a Sub-game Perfect Nash Equilibrium for the repeated game in sector* i*, in which firms always collude.*

Equation (6) is the incentive compatibility constraint which requires that all firms in a sector prefer collusion to cheating in every possible state. Accordingly, such a sequence of assigned collusion markups are incentive compatible by construction and therefore form an equilibrium.

Now, suppose that the model is calibrated such that the constraint binds in the steady state (otherwise, the oligopoly acts as a monopoly and the model essentially becomes a monopolistic competition model across sectors). Then for small perturbations around that steady state, a first

order approximation yields

$$\hat{\mu}_t = \psi_1 \mathbb{E}_t \left[\Delta \hat{y}_{t+1} + \hat{q}_{t,t+1} \right] + \psi_2 \mathbb{E}_t \left[\hat{\mu}_{t+1} \right] \tag{7}$$

where $\Delta \hat{y}_{t+1} \equiv \frac{\Delta Y_{t+1}}{\bar{Y}}$ is percentage growth of sales with respect to the steady state output, and $\hat{q}_{t,t+1} \equiv \frac{Q_{t,t+1} - \bar{Q}}{\bar{Q}}$ is percentage deviation of the stochastic discount rate from its steady state level. Moreover,

$$\psi_1 \equiv \gamma\beta \frac{\bar{\mu}\bar{\Gamma}}{D(\bar{\rho};1) - 1/N} \geq 0 \tag{8}$$

$$\psi_2 \equiv \gamma\beta \frac{D(\bar{\rho};1) - (\sigma-1)(\mu_C - 1)(\frac{\bar{\mu}}{\mu_C})^{\sigma}/N}{D(\bar{\rho};1) - 1/N} \lessgtr 0 \tag{9}$$

Equation (7) gives the law of motion for average markups in the partial equilibrium of the firm side in this economy. This is the key equation in this paper that will underlie all the results in later sections. Therefore, the following subsection is devoted to discussing this result.

Interpretation. Implicit collusion implies that markups are forward looking variables that depend on the expected change in demand in the next period, the changes in the price of future profits, and the expected change of markup in the next period. ψ_1, which is the coefficient on the first two, is a positive number that is increasing in steady state gains from collusion ($\gamma\beta\bar{\mu}\bar{\Gamma}$) and decreasing in the marginal revenue that a firm makes by cheating in the steady state ($D(\bar{\rho};1) - 1/N$). The intuition behind this equation is the key to understanding the main results of this paper. Two things between current period and the period ahead affect the current period's markup: first, the current price of next period's profit, which is the discount factor of the firms. The more patient the firms are in an industry, the higher their collusion markup will be today as they value future profits more. Second, the expected growth in demand from current period to next period. If firms expect that demand tomorrow will be higher than today, then they do not want to lose the chance of cheating tomorrow by cheating today. Basically, firms want to wait until demand is at its highest to take advantage of cheating, as in that case they will collect the highest cheating gains. This incentive to wait diminishes firms' cheating incentives in the current period, allowing the industry to sustain a higher collusion markup. Therefore, when firms expect output to grow, they will charge markups that are closer to the monopoly one.

ψ_2, however, can theoretically be positive or negative based on the calibration of the model. The reason is that there are two opposite forces that affect the firms' cheating incentives based on their expectation of the future markup. Before explaining these two forces, it is useful to recall that what ultimately determines the sign and the magnitude of the change in the markup is how hard it is to sustain the collusion markup, or in other words, how motivated firms are to cheat given a level of markup for the industry.

Suppose that firms expect that the markup next period will be higher than its steady state level.

On one hand, they know that their industry is going to collude on a higher markup tomorrow, so they do not want to miss that chance by cheating today and pushing the industry to the punishment stage. On the other hand, since firms know that all other industries will also charge high markups, they expect to have a very high demand shift towards their industry from the final good producer, if industry as a whole charge the static best response markup. This second force gives an incentive to every single firm to push the industry to punishment stage by cheating today. Obviously the magnitude of this effect depends on how elastic the final good producer's demand for the industry is; as seen in the expression of ψ_2, when σ is close to 1, this force is negligible because the firms do not expect to get a large demand shift if the industry moves to the static Nash equilibrium.

The previous results in this literature can be seen as special cases of equation 7. For example, Rotemberg and Saloner (1986) setup can be seen as the special case where $\hat{q}_{t,t+1} = 0$ due to a constant discount rate, and $\mathbb{E}_t\left[\Delta\hat{y}_{t+1}\right] = -\hat{y}_t$ as shocks are assumed to be i.i.d. over time. Therefore, in their model

$$\hat{\mu}_t = -\psi_1 \hat{y}_t$$

which is a demonstration of their result that markups should be counter-cyclical. But as (7) implies, assuming other processes for these variables can give rise to different results. With two different random processes, $\Delta\hat{y}_{t+1}$ and $\hat{q}_{t,t+1}$, that are potentially correlated, the spectrum of possibilities for their underlying distribution is large enough to allow for *any* type of result in terms of the cyclicality of markups. Therefore, we need to pin down this joint distribution, which in the case of this paper will be done by introducing a household side for the model.

Finally, ψ_1 and ψ_2 are completely pinned down by the firm side parameters σ, η, N, γ plus β which is going to be the subjective discount factor of the households in the general equilibrium.

2.2 Customer-base Models

Another class of models that micro-found variable markups is based on the notion that there is inertia in how fast customers shift their demand across firms (Phelps and Winter, 1970).[8] Accordingly, firms' pricing decisions in the current period affect their market share in future periods. The dynamics of markups in these models depend on how customers are reacting to pricing of the firms over time.

In this section, we build a simple reduced form customer-base model where the inertia in demand comes from habit formation on the customer side. We then show that these models imply a similar law of motion for markups as the implicit collusion models but have different predictions for the sign of coefficients on output growth and stochastic discount rates.

[8] See Bornstein (2018) for a micro-foundation with an application to recent trends in business dynamism and markups.

2.2.1 Model Specification

Consider the final good producer of Section 2.1. To incorporate the customer-base model, we assume that this final good producer forms external habits over the goods within industries, meaning that

$$Y_t = \left[\int_0^1 Y_{i,t}^{\frac{\sigma-1}{\sigma}} di\right]^{\frac{\sigma}{\sigma-1}} \tag{10}$$
$$Y_{i,t} \equiv \left[N^{-\frac{1}{\eta}} \sum_{j=1}^{N} S_{i,j,t}^{\frac{1}{\eta}} Y_{i,j,t}^{\frac{\eta-1}{\eta}}\right]^{\frac{\eta}{\eta-1}}$$

Where $S_{i,j,t}$ is the external habit of producer in using the input $Y_{i,j,t}$, which is taken as given by the final good producer at time t. We assume that $S_{i,j,t}$ has the following general law of motion

$$S_{i,j,t} = \gamma h(\frac{\mu_{i,j,t}}{\mu_{i,t}}) S_{i,j,t-1} + 1 - \gamma$$

where $h(.)$ is differentiable, $h(1) = 1$, $h'(.) < 0$, and $\gamma \in [0,1)$. Notice that $\gamma h(\frac{\mu_{i,j,t}}{\mu_{i,t}})$ is how fast the habit of the final good producer depreciates over time, $h'(.) < 0$ implying that this depreciation is faster if the firm charges a higher markup relative to its competitors (Phelps and Winter (1970) interpret $S_{i,j,t}$ as the measure of customers matched to the firm at time t, in which case $\gamma h(\frac{\mu_{i,j,t}}{\mu_{i,t}})$ is the separation rate of the firm's customers, which increases with the firm's markup).[9]

Solving the final good producer's problem now implies the following demand structure:

$$Y_{i,j,t} = Y_t S_{i,j,t} D(P_{i,j,t}; P_{i,-j,t})$$

where $D(.;.)$ is defined exactly as in Section 2.1. Firm i,j takes demand as given and maximizes the net present value of all its future profits by choosing a relative markup $\frac{\mu_{i,j,t}}{\mu_{i,t}}$, and $S_{i,j,t}$, where $S_{i,t}$ is the the final good producer's habit for the others in the sector, in the symmetric equilibrium. Therefore, firm i,j's dynamic problem is

$$\max_{\{\mu_{i,j,t}, S_{i,j,t}\}_{t=0}^{\infty}} \mathbb{E}_0 \sum_{t=0}^{\infty} \beta^t Q_{0,t} Y_t S_{i,j,t} \left(\frac{\mu_{i,t}}{\mu_t}\right)^{1-\sigma} (\frac{\mu_{i,j,t}-1}{\mu_{i,t}}) D(\frac{\mu_{i,j,t}}{\mu_{i,t}}; 1)$$
$$s.t. \qquad S_{i,j,t} = \gamma h(\frac{\mu_{i,j,t}}{\mu_{i,t}}) S_{i,j,t-1} + 1 - \gamma$$

Proposition 2. In a symmetric equilibrium where all firms identically solve the problem above, the law of motion for markups, up to a first order approximation, takes the same form as the implicit collusion model, i.e.

$$\hat{\mu}_t = \psi_1 \mathbb{E}_t[\hat{q}_{t,t+1} + \Delta \hat{y}_{t+1}] + \psi_2 \mathbb{E}_t[\hat{\mu}_{t+1}] \tag{11}$$

where

$$\psi_1 \equiv -\frac{\beta\gamma}{1+\zeta} \frac{\zeta(1-\mu_s^{-1})}{\mu_s^{-1}(1-\beta\gamma)+\zeta} \leq 0 \tag{12}$$

[9] For a more recent model of customers as demand shifters see Afrouzi, Drenik, and Kim (2020).

$$\psi_2 \equiv \frac{\beta\gamma}{1+\zeta} \geq 0 \tag{13}$$

with $\zeta \equiv -\frac{\gamma h'(1)}{(1-N^{-1})\eta+N^{-1}\sigma} > 0$ and $\mu_s = \frac{\eta(1-N^{-1})+\sigma N^{-1}}{(\eta-1)(1-N^{-1})+(\sigma-1)N^{-1}}$ being the markup of a firm with no inertia in their demand. Moreover, the magnitudes of ψ_1 and ψ_2 decrease with the number of firms within sectors.

Proof. See Appendix. ■

Equation (11) formalizes the idea that the law of motion implied by this model takes the same form as that of the implicit collusion model. In these models, firms' markups are variable because their incentives to invest in their customer-base varies with their expectations of how much demand there will be in the future. Again, the trade-off boils down to a relative demand argument: if the firm expects demand to be larger in the future relative to today, they have a higher incentive to invest in their customer-base and vice-versa. Therefore, what determines today's markup is the firms' expectation of its sales growth, which implies a similar law of motion for markups as in the implicit collusion model, but the signs of ψ's are different as shown in Equations (12) and (13).

In contrast to the implicit collusion model where markups move positively with expected growth of sales, the customer-base model assigns a negative relationship between the two. The reason is that expected higher demand in the future induces firms to invest more in their customer-bases: they reduce their markups to lure in demand, expecting that they will "harvest" a higher demand when aggregate demand is expected to be higher.[10]

3 Testing the Law of Motion for Markups

The goal of this section is to empirically test this law of motion and provide evidence on the signs and magnitudes of coefficients ψ_1 and ψ_2. To do so, we use data from Compustat as well as data on firms' expectations from New Zealand (introduced by Coibion, Gorodnichenko, and Kumar 2018) to study the relationship between firms' markups and their future (expected) sales.

3.1 Model Predictions

Before presenting our empirical evidence, we draw a series of predictions that are implied by the two models in the previous section to guide our empirical strategy. The following statements summarize these predictions:

[10] See Bornstein (2018) for a detailed and recent discussion of investment and harvesting motives of firms when there is inertia in demand.

1. Both implicit collusion and customer base models relate markups to the net present value of firms' future sales and imply a law of motion for markups of the form

$$\hat{\mu}_t = \psi_1 \mathbb{E}_t[\Delta \hat{y}_{t+1} + \hat{q}_{t,t+1}] + \psi_2 \mathbb{E}_t[\hat{\mu}_{t+1}]$$

where the implication of each model for the signs of the coefficients are different. The following table summarizes these implications:

	Implicit Collusion	Customer Base
ψ_1	> 0	< 0
ψ_2	$\lessgtr 0$	> 0

Table 1: Sign of Coefficients in the Law of Motion for Markups in Different Models

2. In both models, the magnitude of the coefficient on future sales ($|\psi_1|$) increases with a firm's relative size ($1/N$ in the model).
3. In both models, the magnitude of coefficients ψ_1 and ψ_2 decrease with the firms' discount factor.

3.2 Evidence from the United States

In order to analyze firms' markups in the U.S., we have exploited the richness of the firm-level balance sheet information from Compustat dataset. With the caveat that this dataset covers only publicly listed firms, Compustat includes panel data on publicly traded firms since 1960 and it is has been used in the recent literature on price-cost markups. Following De Loecker, Eeckhout, and Unger (2020), we estimate markups as:

$$\mu_{it} = \theta^v_{i,t} \frac{P_{i,t} Q_{i,t}}{P^v_{i,t} V_{i,t}},$$

where $V_{i,t}$ is a variable input of production, $Q_{i,t}$ is output and $\theta^v_{i,t}$ is the output elasticity of the variable input $V_{i,t}$. To estimate $\frac{P_{i,t}Q_{i,t}}{P^v_{i,t}V_{i,t}}$, we use the variables *Sales* and *Cost of Goods Sold (COGS)* from Compustat, which are further explained in Appendix B. In order to estimate the output elasticity of the variable input, we used the production function estimation method also used in De Loecker, Eeckhout, and Unger (2020) with some small changes.[11] Table 2 provides summary statistics for firms' sales and markups in the data.

3.2.1 Benchmark Specification and Results for Prediction 1

Having measures of markups and sales growth in the data (denoting them $M_{i,t}$ and $\Delta \log(Sales_{i,t})$ for firm i at time t), we estimate the following specification:

$$\log(M_{i,t}) = \phi_1 \Delta \log(Sales_{i,t+1}) + \phi_2 \log(M_{i,t+1}) + \varepsilon_{i,t} \tag{14}$$

[11] Specifically, following Traina (2018) we estimated time-invariant but industry-specific output elasticities. We used SIC 2-digit codes for the definition of industries.

where we instrument for expectations of $\Delta\log(Sales_{i,t+1})$ and $\log(M_{i,t+1})$ using the following GMM condition:

$$\mathbb{E}_t[(\log(M_{i,t}) - \phi_1\Delta\log(Sales_{i,t+1}) + \phi_2\log(M_{i,t+1}))\mathbf{z}_{t-1}] = 0 \tag{15}$$

Here, $\mathbf{z}_{t-1}$ includes four lags of log of sales and log of markups for the firm dated $t-1$ and before. This instrumental variable approach is necessary because the model predicts that markups are forward-looking and are determined by firms' *expectations* of their future sales growth. Since we do not observe firms' expectations, assuming rational expectations, we use the realized values of these variables as proxies for firms' forecasts of these variables at time t. Nonetheless, this creates an endogeneity problem because any shocks to markups or sales between time t and $t+1$ are orthogonal to expectations at time t but correlate with realized values of these variables, hence biasing the estimates. Using lags of sales and markups allows us to eliminate this concern by only utilizing the variation in $\Delta\log(Sales_{i,t+1})$ and $\log(\mu_{i,t+1})$ that is predictable at time t, but orthogonal to any shock to these variables after time t.[12]

The results of this estimation exercise are reported in Table 3. Column (3), which reports IV-GMM estimates with both time and SIC 2-digit industry fixed effects, constitutes our benchmark results for the U.S. and shows that markups are positively correlated with firms' future expected sales growths and future markups. For reference, Column (1) shows the OLS estimates, and Column (2) reports IV-GMM estimates with time but without SIC 2-digit industry fixed effects. It is important to note that the OLS estimates give the opposite sign, indicating the necessity of our instrumental variable approach.

The positive signs on the coefficients for future expected sales and markups are consistent with the implicit collusion model but goes against the predictions of the customer-base models (Prediction 1). Having this in mind, we examine the heterogeneity of the effects based on size and discount factor next (Predictions 2 and 3).

Heterogeneity Based on Relative Size (Prediction 2). One major prediction of both models is that the magnitude of the coefficient on relative sales should be increasing in a firm's relative size (Prediction 2). To test this feature, we divide the Compustat sample based on a lagged measure of relative sales—defined as the firms' sales relative to their SIC 1-digit industry sales within a year—and repeat the IV-GMM estimation in Equation (14) for the resulting two subsamples.[13]

The results of this exercise is reported in Table 4 and is consistent with Prediction 2. Among the firms above the median of relative size, the coefficient on $\Delta\log(Sales)_{i,t+1}$ is statistically

[12] This is a common approach to deal with such endogeneity issues in estimating forward-looking equations in macroeconomics. See, for instance, Galí and Gertler (1999).

[13] We construct this measure yearly to capture the variation in firms' sales over time. Moreover, for every year, we group firms above and below median based on their lagged relative sales to account for any mechanical correlation between contemporaneous relative size and markup.

significant and almost twice the size of the coefficient reported in our benchmark regression in Column (3) of Table 3. In contrast, among the firms below the median of relative size, the coefficient on $\Delta \log(Sales)_{i,t+1}$ is statistically insignificant and its magnitude is smaller than the one for the whole sample. Both these observations go in the direction predicted by the models and we cannot reject the null hypothesis posed by Prediction 2.

It is important to note that even though we set the cutoff based on the median of relative sales, this prediction is robust to different cutoffs, and the markups of firms with larger relative sales comove more with their expected future sales growths. Moreover, while we divide the firms based on the median relative size, the total sales of the above median group in Table 4 is, on average, 98% of total sales in Compustat, which is due to the highly skewed distribution of sales in the U.S. Therefore, insofar we weight the two groups based on sales, the significance of the coefficient among the above median group towers over the insignificance of this coefficient among the below median group.[14]

Heterogeneity Based on Debt-to-Asset Ratio (Prediction 3). Since both models micro-found markups based on dynamic incentives, another prediction of the two models is that markups should be more sensitive to future sales of firms if they are more patient, meaning that they assign higher values to future profits relative to contemporaneous profits.

While we do not observe discount rates directly, it is reasonable to assume that financially constrained firms assign lower relative value to future profits due to a preference for liquidity in the short-run (as in Gilchrist, Schoenle, Sim, and Zakrajšek, 2017). Taking these results as given, Prediction 3 would imply that we should see a smaller sensitivity of markups to future sales among more leveraged firms. To test this in the data, we divide the Compustat sample based on a lagged measure of firms' debt-to-asset ratios and repeat the IV-GMM estimation in Equation 14 for the two subsamples.

Table 5 reports the estimates for this exercise, grouping firms above and below the top quartile based on their debt-to-asset ratios. The estimates are in line with the prediction above. In particular, Column (1) shows a significant relationship between markups and future sales for firms with debt-to-asset ratios below the top quartile, while Column (2) shows that this relationship is not statistically significant (with a smaller point estimate) for firms in the top quartile of debt-to-asset ratio.

[14] Figure A.9 also plots the sales share of different percentiles of firms in the Compustat data over time. By 2010s, the top 1% of firms account for almost 45% of sales.

3.3 Evidence from New Zealand

Our estimation strategy in Compustat relies on estimating markups as well as assuming rational expectations with full information,[15] so that we can instrument for these expectations using the realized values of future sales and markups. In this section, we depart from these assumptions and provide direct evidence for the law of motion for markups using data on firm's expectations from an expectations survey conducted by Coibion, Gorodnichenko, and Kumar (2018) in New Zealand. In particular, in this survey, firms were asked to provide information about their number of competitors, average markup, current markup, their expected growth in sales, and their next expected price change, which provides an alternative strategy to estimate the law of motion using firms' answers to these questions. Table 6 provides summary statistics of these variables in the data.

3.3.1 Specification and Results

We estimate the following specification in the survey data:

$$\hat{\mu}_{i,j} = \phi_1 Ex\Delta Sales_{i,j} + \phi_2 Ex\Delta Price_{i,j} + \lambda_i + \varepsilon_{i,j} \tag{16}$$

where i, j denotes firm j in industry i, $\hat{\mu}_{i,j}$ is the deviation of i, j's markup from its average markup, $ExSales_{i,j}$ is the firm's expected sales growth, $Ex\Delta Price$ is its expected price change and λ_i is an industry fixed effect. Appendix D shows how this cross-sectional data allows us to test the predictions of the model. In particular, under fairly regular assumptions signs of ϕ_1 and ϕ_2 correspond directly to signs of ψ_1 and ψ_2 in the model.

While the predictions of the customer base model holds for any number of competitors (including monopolistic competition), a distinct characteristic of implicit collusion model is that the coefficients on the law of motion should be larger for firms with a smaller number of competitors. With that in mind, to test the validity of the the law of motion, we divide the sample into two sub-samples, firms with more than 20 competitors, i.e. competitive firms, and firms with fewer than 20 competitors.[16] Table 7 shows the results of running the regression in Equation (16) for these two subsamples. The results are consistent with the implicit collusion model in several dimensions.

[15] In Appendix E we show that while full information rational expectations is a sufficient condition for our law of motion for markups, it is not necessary. In particular, we show that as long as firms' expectations of their *own* future sales growths coincide with full information rational expectations of those sales growths, aggregation works in the sense that firms' [lack of] knowledge about aggregate variables is irrelevant to the derivation of the law of motion for markups.

[16] We also drop firms with less than 2 competitors considering the possibility of a non-binding incentive compatibility constraint for these firms. Figure A.4 shows how these estimates change as a function of this cutoff. The results are locally robust but the data is very granular, with a substantial number of firms reporting they have a round number of competitors, as shown in Figure A.5.

First, the coefficients on expected sales and on expected price changes are positive and negative respectively for the $N < 20$ sub-sample. Both of these signs are consistent with the implicit collusion model but at odds with the customer-base model, as they suggest that oligopolistic firms (1) increase their markup with higher expected sales, and (2) decreases their markups with positive expected price changes.

The positive sign on expected sales growth is consistent with the prediction of the implicit collusion model that oligopolistic firms can sustain higher markups when sales are expected to be higher in the future. In contrast, this positive sign goes against the prediction of the customer base models that firms with higher expected sales in the future should reduce their markups to attract more customers.

Moreover, the fact that these coefficients are only significant for firms with fewer competitors, but not for the more competitive firms, provides a placebo test for the implicit collusion model—as it should hold more significantly for less competitive firms.

4 Implications for Cyclicality of Markups

Given that our empirical results support the predictions of the implicit collusion model, in this section we investigate the implications of our law of motion for markups in a calibrated implicit collusion model with supply (TFP) and demand (government spending) shocks. To do so, we first extend the partial equilibrium model from Section 2.1 to a general equilibrium model with a representative household that saves and accumulates capital in the economy. We then study the implications of our empirical findings for cyclicality of markups.

Our argument here revolves around the fact that the law of motion for markups under the implicit collusion model relates markups to the net present value of future sales growths of the firm, which in the general equilibrium is closely linked to *output growth* in the economy (as firms' total sales in the economy is equal to total output in the equilibrium). Therefore, what determines the cyclicality of markups is the cyclicality of output growth in the general equilibrium, which in turn depends on how hump-shaped the response of output is to shocks. In particular, we show that incorporating the hump-shaped response of output observed in the data *reverses* the cyclicality of markups in the model.

4.1 Households, the Government and Market Clearing

There is a representative household that solves the following standard problem with investment adjustment costs.

$$\max_{\{(C_t,L_t,I_t,K_{t+1},B_t)\}_{t=0}^{\infty}} \mathbb{E}_0 \sum_{t=0}^{\infty} \beta^t \left[\frac{C_t^{1-\theta}}{1-\theta} - \phi \frac{L_t^{1+1/\epsilon}}{1+1/\epsilon} \right]$$

$$s.t. \quad P_t C_t + P_t I_t + B_t \leq W_t L_t + R_t K_t + (1+i_{t-1}) B_{t-1} + \int_0^1 \sum_{j=1}^{N} \Pi_{i,j,t} di - T_t$$

$$K_{t+1} = (1-\delta)K_t + (1 - S(\tfrac{I_t}{I_{t-1}}))I_t$$
$$S(\tfrac{I_t}{I_{t-1}}) \equiv \tfrac{a}{2}(1 - \tfrac{I_t}{I_{t-1}})^2$$

where C_t is consumption, L_t is labor supply, I_t is investment, K_t is capital, and B_t is a nominal riskless bond with nominal return i_t at $t+1$. The investment adjustment cost is included to allow for a hump-shape in the response of output to shocks. As we discuss later, the extent of this hump-shape response is crucial in determining the cyclicality of markups.

There is also a government that uses lump-sum taxes from households to conduct fiscal policy, G_t. We assume that G_t follows an AR(2) stochastic process

$$G_t = \bar{G} Z_t^g$$
$$\log(Z_t^g) = \rho_1^g \log(Z_{t-1}^g) + \rho_2^g \log(Z_{t-2}^g) + \sigma_g \varepsilon_{g,t}$$
$$\varepsilon_{g,t} \sim \mathcal{N}(0,1)$$

The AR(2) assumption on government spending process is to allow for a hump-shaped response of output to a fiscal policy shock. Moreover, we assume that the monetary policy is set based on the Taylor rule $i_t = \phi_\pi \log(P_t/P_{t-1})$.

Finally, the market clearing conditions for the final good, capital and labor markets are

$$C_t + I_t + G_t = Y_t$$
$$K_t = \int_0^1 \sum_{j=1}^N K_{i,j,t} di$$
$$L_t = \int_0^1 \sum_{j=1}^N L_{i,j,t} di$$

4.2 Calibration and Simulation

In this section, by simulating a log-linearized version of the model around a steady state in which the incentive compatibility constraint binds, we show (1) why implicit collusions models are typically interpreted as implying countercyclical markups and (2) that markups are actually pro-cyclical once the model is calibrated to generate a hump-shape response for output.

4.2.1 Parameters

We set $\beta = 0.993$ to match the a steady state annual real interest rate of 3 percent, $\alpha = 0.35$ to match a steady state share of capital income of 35 percent, $\delta = 0.025$ to match a 10 percent annual rate of depreciation on capital, $\phi = 8$ to match a steady state labor supply of 0.3, $\bar{G} = 0.2$ to match a steady state G/Y of 20 percent, and $a = 2.48$ following Christiano, Eichenbaum, and Evans (2005). We also set the response of the central bank to inflation in Taylor rule, ϕ_π to the common value of 1.5. Furthermore, we set the Frisch labor supply elasticity, ϵ, to 2.5. Moreover, we set the elasticity of substitution across sectoral goods, σ, equal to 4, and the elasticity of substitution within sectoral goods, η, equal to 20. We set $\gamma = 0.8$ and $N = 15$ to match a steady

state markup level of 20 percent.[17]

Although these values are calibrated in an arbitrary fashion, we show in a series of robustness checks in Section 4.2.4, for the given levels of σ and η the model is not very sensitive to these parameters, and reasonable variations in them do not affect the main results of our analysis. The qualitative results in terms of direction of cyclicality of markups are robust to any calibration as long as $\eta > \sigma$.

Finally, we set the persistence of the technology shock to 0.95. For the persistence parameters of the government spending shock, we run the following regression on the quarterly data for real government consumption expenditures and gross investment from 1947Q1 to 2014Q1:

$$\log(G_t) = Constant + \rho_1^g \log(G_{t-1}) + \rho_2^g \log(G_{t-2}) + \varepsilon_t$$

which gives the estimates $\rho_1^g = 1.51$ and $\rho_2^g = -0.52$. We also consider alternative persistence parameters for robustness checks in Section 4.2.4.

4.2.2 Impulse Response Functions

First, consider the case of no investment adjustment cost ($a = 0$). The dashed curves in Figure 1a show the impulse responses of this model to a 1 percent technology shock.[18] The key observation is that in this setting, output jumps up on impact and converges back to zero as the effect of the transitory shock fades away. Moreover, the response of stochastic discount rate, which is given by $Q_{t,t+1} = \beta \frac{u'(C_{t+1})}{u'(C_t)}$, is countercyclical given that households are able to smooth their consumption without being restricted by costly investment. The fact that consumption has an inertial response to the technology shock is a crucial element to the countercyclicality of stochastic discount rates. On impact, households expect that their consumption will peak later in the expansion; therefore, they are not really concerned about future states as they know they will have a higher consumption.

By Equation 7, the combination of countercyclical output growth and discount rates gives rise to countercyclical markups. The interpretation from the firm side is that on impact, firms know that demand is at its highest. This expectation along with the low price of future profits increases increases firms' incentives to deviate from implicit collusion and forces the oligopoly to settle on a lower markup in order to eliminate these incentives.

A similar exercise can be done with the government spending shock. Suppose that Z_t^g is an AR(1) process with persistence 0.95. The impulse response functions of the model to such a shock is illustrated by the dashed curves in Figure 1b. On impact, government spending is at its highest, which means that private consumption is at its lowest. First, since private consumption

[17] Given σ and η, this is the highest level for γ for which the incentive compatibility constraint binds, and the Blanchard Kahn condition for the law of motion for markups holds.

[18] We use Dynare (Adjemian, Bastani, Juillard, Karamé, Maih, Mihoubi, Perendia, Pfeifer, Ratto, and Villemot (2011)) to solve the model.

will increase to its steady state level, such a shock would give rise to countercyclical stochastic discount rates. Moreover, the income effect of G is at its highest on impact, so that Y will peak immediately due to a jump in labor supply and converge back to its steady state as the shock fades away. Again, the combination of countercyclical discount rates and output growth will translate into countercyclical markups.

However, empirical evidence on TFP shocks and government spending shocks suggests that the response of output to these shocks is hump-shaped such that the peak effect happens not on impact but in later periods.[19] To allow for such a response, we introduce investment adjustment costs and an AR(2) process for government spending. Solid curves in Figure 1a show the IRFs of the model to a 1% technology shock when $a = 2.48$. With positive adjustment costs, two things happen. First, investment does not jump on impact and has an inertial response, which translates to a hump-shaped response in output. Second, households now face a stronger trade-off in smoothing their consumption because they face costly investment, which gives rise to procyclical stochastic discount rates. Therefore, on impact firms expect their demand to increase in future periods, which gives them the incentive sustain higher markups as they expect their demand to increase. Hence, on impact one would expect a higher markup than the one in the steady state, making markups procyclical.

A similar exercise can be done with the government spending shock by assuming an AR(2) process for Z_t^g. Figure 1b depicts the IRFs of the model to such a shock. The hump-shaped implementation of the fiscal policy translates to hump-shaped output and consumption responses, as shown by solid curves in Figure 1b, which in turn produce procyclical markups for similar reasons to the case of the technology shock with $a > 0$.[20]

4.2.3 Matching the Evidence for Conditional Cyclicality of Markups

So far we have shown that a direct implication of our law of motion for markups is that it *reverses* the cyclicality of markups once the model matches the hump-shaped response of output to shocks. In particular, our model predicts that under hump-shaped output response to TFP and government spending shocks, markups are *procyclical*.

These predictions directly relate to, and are qualitatively consistent with, the evidence on conditional cyclicality of markups in recent work by Nekarda and Ramey (2020), who find that (1) output response is hump-shaped with respect to TFP and government spending shocks, and

[19] For empirical evidence on hump-shaped response of output to productivity and government spending shocks, see, for example, Sims (2011); Smets and Wouters (2007); Ramey (2011); Ramey and Shapiro (1998); Christiano, Eichenbaum, and Evans (2005); Monacelli and Perotti (2008); Nekarda and Ramey (2020).

[20] For completeness, we have also included IRFs of a customer-base model in Figure A.6, where the calibration is such that the frictionless markup, μ_s, is 11.5%, and the average markup, μ, is 10%. In terms of parameters, the only change compared to the previous section is $\eta = 10$. Notice that in that model markups are countercyclical once the model matches the hump-shaped response of output to shocks.

(2) markups are procyclical conditional on these two types of shocks. It is important to note that without the hump-shaped response of output, the model would completely miss these two facts and would deliver the wrong prediction that markups are countercyclical with respect to both TFP and government spending shocks.

To formally show this, Figure 2 plots the cross-correlation of markup and output conditional on TFP shocks under the model with and without hump-shaped output response against empirical estimates from Nekarda and Ramey (2020) for these correlations.[21] While the model *without* inertia (hump-shaped response of output) fully misses the sign of these correlations, implying that markup and output are negatively correlated, the model *with* inertia gets the sign of and the magnitude of the contemporaneous and lagged correlations of markup and output conditional on TFP shocks right.[22]

It is important to note that while the model matches the contemporaneous and lagged conditional correlations correctly, it predicts that leads of markups should be *countercyclical*, which is not the case in Nekarda and Ramey (2020)'s estimates. This is because in the model markups fall below their steady state level once output peaks, whereas in Nekarda and Ramey (2020)'s estimates, markups remain procyclical for much longer and their point estimates only fall below their steady state level after 7 quarters.

For completeness, Figure A.8 shows the results of a similar exercise in a customer-base model, and illustrates how this model completely misses the sign and structure of these cross-correlations.

4.2.4 Robustness

In this section, we check the robustness of the predictions of the model with respect to different parameters.

Probability of Renegotiation (Discount Factor). Figures A.1a and A.1c show the simulated correlations of leads and lags of markups with output conditional on a technology shock and a government spending shock respectively, for values of γ between 0.4 and 0.8, such that darker curves correspond to higher levels of γ. Aside from the fact that lower γ's create lower steady state markups because of the higher impatience of firms, they also produce lower correlations between output and markups. The reason for the latter is that variations in current markup are a weighted sum of all expected output changes and stochastic discount rate changes in the future,

[21] The empirical estimates for these correlations are constructed based on Nekarda and Ramey (2020)'s TFP SVAR that includes log level of Fernald (2014)'s utilization adjusted measure of TFP, log real GDP per capita, log of the output price deflator, the three month Treasury bill rate and log of Nekarda and Ramey (2020)'s measure of markup based on labor share of output, allowing for overhead labor. Using the results of this estimation, we calculate the correlations of output and markup conditional on shocks to TFP.

[22] The same exercise can be done with government spending shocks, and while the results qualitatively remain the same – meaning that correlations in the inertial model are larger than in the model with no inertia in the output response – the inertia created by the AR(2) process is not enough to make the conditional correlation positive.

and as γ gets smaller, they put lower weights on future values. Nevertheless, all values of γ yield the same structure of correlations of lags and leads of the markup with the output.

Number of Competitors. Figures A.1b and A.1d, respectively, show the correlation of leads and lags of markups with output conditional on a technology shock and government spending shock for values of N between 5 and 25. Again darker curves represent higher values of N. Variation in number of competitors does not change the structure of correlations and has very small level effects. The reason is that what ultimately determines the cheating incentives of firms, and hence markups, is the elasticity of demand for a single firm which is equal to $\eta - \frac{\eta-\sigma}{(N-1)\rho^{\eta-1}+1} \in [\sigma, \eta]$. Note that for small amounts of η ($\eta \leq 20$), which corresponds to a relatively high differentiation among within industry goods, the effect of N on the structure and level of correlations is negligible.

Elasticities of Substitution and Frisch Elasticity. Figures A.2a, A.2b and A.2c show the cross-correlation of markup and output conditional on TFP shocks for different values of $\sigma \in [2, 10]$, $\eta \in [10, 30]$ and $\epsilon \in [0.5, 5]$, respectively. While these values seem to slightly change the size of these correlations, the sign and overall pattern of these correlations are robust to these different values. This confirms the intuition derived from the law of motion for markups that the structure of these correlations should remain unchanged insofar the hump-shaped response of output to the shock at hand is not changed significantly. Since none of these parameters are directly responsible for the hump-shaped response of output to TFP, their effect on the cross-correlations are small. In that sense, the only parameters that does affect these correlations is the degree of investment adjustment costs (a) for TFP shocks, and the shape of the AR(2) process for government spending shocks. We now turn into examining the robustness of our predictions with respect to these parameters.

Degree of Inertia. In the model, investment adjustment cost is the mechanism that generates the hump-shaped response of output to technology shocks. While we have calibrated this parameter to the estimated value of Christiano, Eichenbaum, and Evans (2005), this section examines the question of how large this parameter needs to be for markups to be procyclical. In particular, we investigate the cyclicality of markups conditional of TFP shocks change as we change how hump-shaped the response of output is using different values for the parameter governing the degree of investment adjustment costs.

Figure A.3a depicts the number of periods that markups are procyclical after a technology shock given different values of $a \in [0, 5]$. As soon as a is larger than 0, markups are procyclical on impact. Also, the duration of procyclicality increases as a gets larger. For our calibration of this parameter, markups are pro-cyclical for 5 quarters after the shock hits the economy.

Moreover, Figure A.3b shows the contemporaneous correlation of the markup with output conditional on a TFP shock, which is increasing in a and positive for $a > 1.2$. Hence, any empirically reasonable value of investment adjustment costs will generate procyclical markup in

this model.

Regarding government spending shocks, in the baseline calibration with inertia, we use estimated parameters for the AR(2) process of government spending to generate the hump-shaped response of output to a G shock. Now, we consider a wider range of persistence parameters to check for robustness of results in previous section. Consider the set $\{(\rho_1^g, \rho_2^g) | \rho_2^g \in [-0.7, 0], \rho_1^g + \rho_2^g = \rho_G\}$, where ρ_G is the persistence of government spending shocks, fixed to an estimated value of 0.98. Therefore, this set defines a locus for persistence parameters of G such that when $\rho_2^g = 0$ the process is AR(1) and when $\rho_2^g < 0$ the process is AR(2) with highest inertia achieved when $\rho_2^g = -0.7$. In fact, the magnitude of this parameter, $|\rho_2^g|$, determines the degree of inertia in the response of output. Figure A.3c shows the number of periods that markups are procyclical after a 1% government spending shock given different values of $|\rho_2^g|$. Again, for the most part ($|\rho_2^g| > 0.1$), the inertia causes the markups to be procyclical on impact. For our estimate of persistence parameters, markups are procyclical for 2 periods after the impact. However, as Figure A.3d shows that while this inertia is not enough to make the conditional correlation of markup and output positive, it is still increasing with inertia.

5 Discussion of Other Models

In this section, we discuss the relationship between the law of motion that in implicit collusion and customer-base models and how markups are determined in other related models.

5.1 New Keynesian Models

By assuming that prices are sticky but marginal costs are not, New Keynesian models create variable markups both across time and across firms. In these models, two forces interact in shaping the cyclicality of aggregate markups: among the fraction of firms that are not resetting their prices, markups are countercyclical because in an expansion their marginal costs rise but their prices stay the same. However, among firms that do reset their prices, they might increase their prices by more than the contemporaneous increase in their marginal costs due to the forward-looking nature of price-setting in these environments. Therefore, the aggregate cyclicality would depend on which force dominates in total. To see this formally, consider the linearized version of the firms side of the textbook New Keynesian model (see, e.g., Galí, 2015):

$$p_t^* = (1 - \beta\theta) mc_t + \beta\theta \mathbb{E}_t[p_{t+1}^*] \tag{17}$$

$$p_t = (1 - \theta) p_t^* + \theta p_{t-1} \tag{18}$$

where p_t^* is log-deviation of reset price for firms that are changing their prices at time t from its steady state level, mc_t is the log-deviation of nominal marginal cost of firms (which in our setting is given as a function of aggregate wage and the rental rate of capital in Equation 4) from

its steady state level, p_t is the log-deviation of the aggregate price level from its steady state level, β is the discount rate and finally $1-\theta$ is the probability reseting prices at every period. Notice that we can define two notions of markups in this economy: $\mu_t^* \equiv p_t^* - mc_t$ as the markup of the firms that reset their prices at time t and $\mu_t \equiv p_t - mc_t$ as the aggregate markup based on the aggregate price level. Rewriting the equations above in terms of these markups we get:

$$\mu_t^* = \beta\theta\mathbb{E}_t[\Delta mc_{t+1}] + \beta\theta\mathbb{E}_t[\mu_{t+1}^*] \tag{19}$$

$$\mu_t = (1-\theta)\mu_t^* + \theta\mu_{t-1} + \theta\Delta mc_t \tag{20}$$

Notice that Equation (19), which is the *law of motion* for markups among price setters at time t resembles our law of motion for implicit collusion and customer-base models with one major difference: instead of relating markups to changes in the net present value of future sales growths, it relates firms' markups to a discounted average of future changes in marginal cost growths. Therefore, insofar growth in marginal costs, output growth and stochastic discount rates comove positively in response to a shock, the mechanics of how μ_t^* evolves over the business cycle is akin to the other models considered in this paper. Nonetheless, the dynamics of aggregate markups is more complicated and also depends on the history of prices and marginal costs due to the stickiness of prices among the fraction of firms that are not changing their prices at time t.

Furthermore, this model does not capture the heterogeneity that we observe in the Compustat based on relative size (Predictions 2 in Section 3). In the New Keynesian model markups of all firms have the same comovement with their future expected marginal cost growths which is uniquely determined by the discount factor and price stickiness ($\beta\theta$).

On a final note, the standard New Keynesian models take a center stage in the analysis of Nekarda and Ramey (2020), where they find the predictions of these models for cyclicality of markups conditional on demand shocks to be inconsistent with the empirical evidence.

5.2 Models with Variable Elasticities of Demand

Another set of models that generate variable markups over the business cycle rely on preferences that, in contrast to CES preferences, generate variable demand elasticities along firms' demand curves and populate firms along those demand curves based on heterogeneity in size. These models broadly break down to the following two classes.

Kimball Preferences. One approach for variable demand elasticities is to use a generalized aggregator (as in Kimball, 1995) that varies demand elasticity according to Marshall's second law of demand—which requires demand elasticity to decrease with relative price along the demand curve. For instance, this approach has recently been used by Edmond, Midrigan, and Xu (2018) who study the cost of markups in a firm dynamics model. In these models, markups are static but are determined as a function of a firm's demand elasticity given their marginal cost of production.

Formally, let $e(p)$ denote the elasticity of demand at relative price p along the demand curve of firm. Then, the optimal pricing strategy of a firm with real marginal cost mc in these models is to choose p such that

$$p = \frac{e(p)}{e(p)-1} mc \tag{21}$$

which is an implicit equation in the relative price of the firm and determines this relative price as a function of mc. Marshall's second law of demand then implies that firms with higher marginal costs charge higher relative prices *and* lower markups (see, e.g., Lemma 1 in Afrouzi, Drenik, and Kim, 2020).

Since these models relate markups to *relative* prices, heterogeneity plays a crucial role for their conclusions on how markups change over the business cycle. To see this, notice that in a model with no heterogeneity, by symmetry, all firms have to charge the same price which implies all relative prices are always 1, independent of whether the economy is booming or in a recession. Therefore, in such an economy markups are *always* constant and given by $e(1)/(e(1)-1)$. It is when there is heterogeneity in size that these models start to shine, and create prediction for cyclicality of markups as the distribution of prices across the economy starts to change over the business cycle. The implications of these effect for cyclicality of markups, however, are unexplored to the best of our knowledge and remains a promising area for future work.

Nested CES with Heterogeneity in Size. Variable demand elasticities also arise in nested CES preferences (similar to what we assumed in Equation 1), combined with heterogeneity in productivity (Atkeson and Burstein, 2008; Burstein, Carvalho, and Grassi, 2020). Without any dynamic incentives (i.e. implicit collusion or customer-acquisition), pricing decisions in these models are static, where firms set a markup over their marginal cost and their markups are determined by their demand elasticity, which in turn depends on the firm's market share within its sector.

Formally, in such a model, demand elasticity of a firm j in a sector i with market share $s_{i,j}$ can be written as an average of two elasticities of substitution implied by the nested CES system, weighted by the firm's market share:

$$e(s_{i,j}) = s_{i,j}\eta + (1 - s_{i,j})\sigma \tag{22}$$

which then implies the following markup for the firm (Equation 5 in Burstein, Carvalho, and Grassi 2020):

$$\mu(s_{i,j}) = \frac{\eta}{\eta - 1}\left[\frac{1 - (\frac{\eta-\sigma}{\eta})s_{i,j}}{1 - (\frac{\eta-\sigma}{\eta-1})s_{i,j}}\right] \tag{23}$$

Moreover, defining the aggregate markup of a sector as the inverse labor share of that sector implies that the sectoral markup is the harmonic mean of the individual firms' markups, weighted

by their market share (Equation 7 in Burstein, Carvalho, and Grassi 2020):

$$\mu_i^{-1} = \sum_j s_{i,j} \mu(s_{i,j})^{-1} \tag{24}$$

Notice now that sectoral cyclicality of markups depends on two objects: (1) cyclicality of individual firms' markups, and (2) redistribution of sales across the sector during the business cycle. Therefore, the cyclicality of sectoral markups could go against the cyclicality of individual firms' markups within that sector *if there is enough heterogeneity in size*.

To relate our model in Section 2.1 to this framework, we depart from this setting in two ways. First, we consider environments with dynamic incentives which introduces a different law of motion for markups than the static case presented here. These dynamic incentives create intertemporal considerations for firms in choosing their markups. However, for tractability, we do not allow for heterogeneity in size within sectors, as our symmetric equilibria imply that all firms have the same and *constant* market share ($1/N$) within their sectors. This assumption is of course restrictive but allow for tractability as well as simplified testable predictions that we take to the data. Extending these predictions and models to environments with heterogeneity in market shares would be a natural next step for future work.

6 Conclusion

In this paper, we revisit the implicit collusion and customer-base models and show they both imply a forward looking law of motion that relates markups to firms' expectations of the net present value of their future sales growths. We then use data on markups and sales from Compustat and survey data on firms' expectations from New Zealand to test this implied law of motion and find the evidence to be in favor of the implicit collusion models.

In a general equilibrium model, we also show that this law of motion reduces the net present value of firms future sales growths to their expectations of output growth and stochastic discount rates. Because markups are related to the expected output growth, and not to its level, the conditional expectations of firms for the dynamics of output are key in the model for cyclicality of markups. In particular, if firms expect a hump-shaped response for output during the business cycle, the predictions of these models are reversed.

Previous work using these models has not allowed for sufficiently rich dynamics in output which has lead to the conclusion that implicit collusion models lead to counter-cyclical markups. We show that this prediction is overturned once empirically realistic dynamics of output are incorporated into the model, which also helps the implicit collusion model to match the empirical evidence on the dynamic cross correlation of output and markup conditional on TFP shocks, as documented in Nekarda and Ramey (2020).

References

ADJEMIAN, S., H. BASTANI, M. JUILLARD, F. KARAMÉ, J. MAIH, F. MIHOUBI, G. PERENDIA, J. PFEIFER, M. RATTO, AND S. VILLEMOT (2011): "Dynare: Reference Manual Version 4," Dynare Working Papers 1, CEPREMAP.

AFROUZI, H., A. DRENIK, AND R. KIM (2020): "Growing by the Masses: Revisiting the Link between Firm Size and Market Power," *Available at SSRN 3703244*.

ATKESON, A., AND A. BURSTEIN (2008): "Pricing-to-market, trade costs, and international relative prices," *American Economic Review*, 98(5), 1998–2031.

BAGWELL, K., AND R. W. STAIGER (1997): "Collusion over the Business Cycle," *The RAND Journal of Economics*, 28(1), 82–106.

BORNSTEIN, G. (2018): "Entry and profits in an aging economy: The role of consumer inertia," Discussion paper, mimeo.

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

CHRISTIANO, L., M. EICHENBAUM, AND S. REBELO (2011): "When Is the Government Spending Multiplier Large?," *Journal of Political Economy*, 119(1), pp. 78–121.

CHRISTIANO, L. J., M. EICHENBAUM, AND C. L. EVANS (2005): "Nominal rigidities and the dynamic effects of a shock to monetary policy," *Journal of political Economy*, 113(1), 1–45.

COIBION, O., Y. GORODNICHENKO, AND S. KUMAR (2018): "How do firms form their expectations? new survey evidence," *American Economic Review*, 108(9), 2671–2713.

DE LOECKER, J., J. EECKHOUT, AND G. UNGER (2020): "The rise of market power and the macroeconomic implications," *The Quarterly Journal of Economics*, 135(2), 561–644.

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

FERNALD, J. (2014): "A quarterly, utilization-adjusted series on total factor productivity," .

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

GALÍ, J., AND M. GERTLER (1999): "Inflation Dynamics: A Structural Econometric Approach," *Journal of Monetary Economics*, 2.

GILCHRIST, S., R. SCHOENLE, J. SIM, AND E. ZAKRAJŠEK (2017): "Inflation dynamics during the financial crisis," *American Economic Review*, 107(3), 785–823.

GOURIO, F., AND L. RUDANKO (2014): "Customer capital," *Review of Economic Studies*, 81(3), 1102–1136.

GREEN, E. J., AND R. H. PORTER (1984): "Noncooperative Collusion under Imperfect Price Information," *Econometrica*, 52(1), 87–100.

HALTIWANGER, J., AND J. E. HARRINGTON JR (1991): "The impact of cyclical demand movements on collusive behavior," *The RAND Journal of Economics*, pp. 89–106.

KANDORI, M. (1991): "Correlated demand shocks and price wars during booms," *The Review of Economic Studies*, 58(1), 171–180.

KAPLAN, G., AND G. MENZIO (2016): "Shopping externalities and self-fulfilling unemployment fluctuations," *Journal of Political Economy*, 124(3), 771–825.

KIMBALL, M. (1995): "The Quantitative Analytics of the Basic Neomonetarist Model," *Journal of Money, Credit and Banking*, 27(4), 1241–77.

MONACELLI, T., AND R. PEROTTI (2008): "Fiscal policy, wealth effects, and markups," Discussion paper, National Bureau of Economic Research.

NEKARDA, C. J., AND V. A. RAMEY (2020): "The cyclical behavior of the price-cost markup," *Journal of Money, Credit and Banking*, 52(S2), 319–353.

PACIELLO, L., A. POZZI, AND N. TRACHTER (2018): "Price Dynamics with Customer Markets," *International Economic Review*, pp. 413–446.

PHELPS, E. S., AND S. G. WINTER (1970): "Optimal price policy under atomistic competition," *Microeconomic foundations of employment and inflation theory*, pp. 309–337.

RAMEY, V. A. (2011): "Identifying Government Spending Shocks: It's all in the Timing*," *The Quarterly Journal of Economics*, 126(1), 1–50.

RAMEY, V. A., AND M. D. SHAPIRO (1998): "Costly capital reallocation and the effects of government spending," in *Carnegie-Rochester Conference Series on Public Policy*, vol. 48, pp. 145–194.

RAVN, M., S. SCHMITT-GROHE, AND M. URIBE (2006): "Deep Habits," *The Review of Economic Studies*, 73(1), 195–218.

ROTEMBERG, J. J., AND G. SALONER (1986): "A Supergame-Theoretic Model of Price Wars during Booms," *The American Economic Review*, 76(3), 390–407.

ROTEMBERG, J. J., AND M. WOODFORD (1991): "Markups and the Business Cycle," in *NBER Macroeconomics Annual 1991, Volume 6*, NBER Chapters, pp. 63–140. National Bureau of Economic Research, Inc.

ROTEMBERG, J. J., AND M. WOODFORD (1992): "Oligopolistic Pricing and the Effects of Aggregate Demand on Economic Activity," *Journal of Political Economy*, 100(6), 1153–1207.

——— (1999): "The cyclical behavior of prices and costs," *Handbook of macroeconomics*, 1, 1051–1135.

SIMS, E. R. (2011): "Permanent and Transitory Technology Shocks and the Behavior of Hours: A Challenge for DSGE Models," Manuscript.

SMETS, F., AND R. WOUTERS (2007): "Shocks and Frictions in US Business Cycles: A Bayesian DSGE Approach," *The American Economic Review*, 97(3), 586–606.

TRAINA, J. (2018): "Is Aggregate Market Power Increasing? Production Trends Using Financial Statements," Manuscript.

Tables and Figures

Table 2: Descriptive statistics for Compustat data

Statistic	Real sales$_{it}$	% change in sales$_{i,t+1}$	markup$_{it}$	% change in markup$_{i,t+1}$
Mean	1565.05	25.60	1.41	10.52
P 25	27.56	-5.35	1.06	-3.25
P 50	138.01	5.14	1.23	0.03
P 75	648.23	18.77	1.51	3.33
SD	8458.53	841.10	0.79	322.06

Notes: Each column of the table shows the summary statistics of the corresponding variable in the Compustat data. The dataset contains 242,155 observations for 20,252 firms across 67 years (1960-2016). Real sales$_{it}$ for firm i and year t are reported in millions. To calculate markups, we followed the procedure from De Loecker, Eeckhout, and Unger (2020); Traina (2018)–we first estimated time-invariant but industry-specific (SIC 2-digits) output elasticities using the production function estimation method from De Loecker, Eeckhout, and Unger (2020). We then define markup$_{it}$ as output elasticities$_j$ times sales over cost of goods sold (COGS).

Table 3: Law of motion for the U.S.

	(1) log(M$_{i,t}$)	(2) log(M$_{i,t}$)	(3) log(M$_{i,t}$)	(4) log(M$_{i,t}$)
Δ log(sales$_{i,t+1}$)	-0.199***	0.085	0.168**	-0.163***
	(0.007)	(0.085)	(0.083)	(0.010)
log(M$_{i,t+1}$)	0.836***	1.074***	1.078***	0.832***
	(0.004)	(0.006)	(0.006)	(0.006)
R^2	0.711	0.621	0.550	0.729
Year FE	Yes	Yes	Yes	Yes
Industry FEs	Yes	No	Yes	Yes
Method	OLS	IV-GMM	IV-GMM	OLS
N	217,980	145,700	145,700	145,700

$*p<0.10, **p<0.05, ***p<0.01$; standard errors are clustered at the firm-level. The first column contains the results of an OLS regression of log(markup$_{it}$) on Δ log(sales$_{i,t+1}$) and log(markup$_{i,t+1}$). We also include year and industry (SIC 2-digit codes) fixed effects. The second and third columns report the results of the GMM estimator when using four lags of log(sales$_{i,t}$) and log(markup$_{i,t}$) as instruments. The fourth column report the results of the OLS specification from column 1, but this time restricting the observations to the IV-GMM sample used in columns 2 and 3. Our dataset contains 66 industries, 67 years and 242,155 observations.

Table 4: Law of motion for the U.S. split for above and below the median of lag of relative sales

	(1)	(2)
	$\log(M_{i,t})$	$\log(M_{i,t})$
$\Delta\log(\text{sales}_{i,t+1})$	0.079	0.290^{***}
	(0.094)	(0.081)
$\log(M_{i,t+1})$	1.100^{***}	1.030^{***}
	(0.007)	(0.005)
R^2	0.513	0.664
Above median	No	Yes
market share	2%	98%
Industry FEs	Yes	Yes
N	66019	79681

$*p < 0.10, **p < 0.05, ***p < 0.01$; standard errors are clustered at the firm-level. We report the GMM estimator of the effects of $\Delta\log(\text{sales}_{i,t+1})$ and $\log(\text{markup}_{i,t+1})$ on $\log(\text{markup}_{i,t})$ using four lags of $\log(\text{sales}_{i,t})$ and $\log(\text{markup}_{i,t})$ as instruments. We also include year and industry (SIC 2-digit codes) fixed effects. The first column reports the results for observations below the median of lag relative sales, while the second column reports the results for observations above the median. We define relative sales as the total sales for a firm in a given year, divided by the total sales of the industry (SIC 1-digit) in that same year.

Table 5: Law of motion for the U.S. split for above and below the top quartile of lag of debt-to-asset ratio

	(1)	(2)
	$\log(M_{i,t})$	$\log(M_{i,t})$
$\Delta\log(\text{sales}_{i,t+1})$	0.204^{**}	0.097
	(0.104)	(0.154)
$\log(M_{i,t+1})$	1.063^{***}	1.126^{***}
	(0.006)	(0.020)
R^2	0.565	0.438
Top quartile	No	Yes
market share	82.5%	17.5%
Industry FEs	Yes	Yes
N	107415	30994

$*p < 0.10, **p < 0.05, ***p < 0.01$; standard errors are clustered at the firm-level. We report the GMM estimator of the effects of $\Delta\log(\text{sales}_{i,t+1})$ and $\log(\text{markup}_{i,t+1})$ on $\log(\text{markup}_{i,t})$ using four lags of $\log(\text{sales}_{i,t})$ and $\log(\text{markup}_{i,t})$ as instruments. The second column reports the results for observations in the top quartile of the lag of debt-to-asset ratio, while the first column report the results for observations below the top quartile. We include year and industry (SIC 2-digits) fixed effects.

Table 6: Descriptive statistics for New Zealand data

Statistic	Firm production value	Expected % change in sales	Markup	Expected size of next price change
Mean	992,136.69	4.39	25.02	4.64
P 25	211,600.00	0.00	15.00	2.00
P 50	369,500.00	5.00	25.00	5.00
P 75	1,061,500	8.50	34.50	8.00
SD	1,935,236	5.84	12.18	5.27

The table provides summary statistics for the survey of firms' expectations from New Zealand. The dataset contains 3,153 observations for 3,153 firms in 18 industries and corresponds to the first wave of the survey conducted by Coibion, Gorodnichenko, and Kumar (2018).

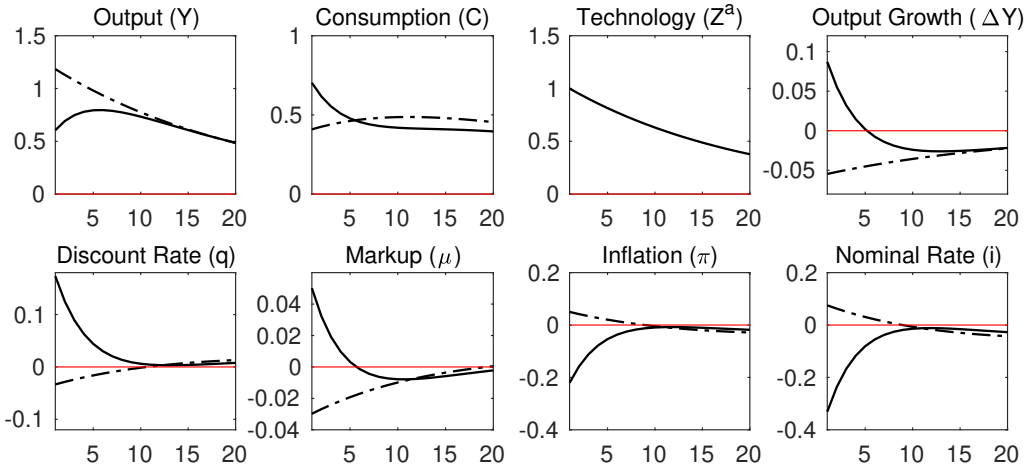
Table 7: Law of Motion for Markups: Survey Data from New Zealand

	(1)	(2)
	Markup	Markup
Expected size of next price change	-0.178***	0.025
	(0.062)	(0.090)
Expected growth in sales	0.163**	-0.037
	(0.082)	(0.107)
R^2	0.118	0.153
Industry FE	Yes	Yes
Number of competitors	$2 \leq$ competitors ≤ 20	competitors > 20
N	495	200

$*p < 0.10, **p < 0.05, ***p < 0.01$; robust standard errors. The table reports the coefficients for the regression specified in Equation (16) allowing for industry fixed effects. The first column reports the coefficients for firms that report less than 20 but more than 2 competitors. Second column reports the coefficients for firms that report more than 20 competitors.

Figure 1: Impulse Response Functions: Implicit Collusion Model

(a) The dashed curves plot the impulse response functions of the implicit collusion model to a 1% technology shock with no adjustment cost in which markups are counter-cyclical as output growth and stochastic discount rates are counter-cyclical. Solid curves illustrate the impulse response functions of the same model to a 1% technology shock with investment adjustment cost. Markups are pro-cyclical as long as firms expect output to grow. See Section 4.2.2 for details.



(b) The dashed curves plot the impulse response functions of the implicit collusion model to a 1% government spending shock without inertia in which markups are counter-cyclical as output growth is negative during the expansion. Solid curves illustrate the impulse response functions of the same model to an inertial government spending shock that peaks at 1%. Markups are pro-cyclical on impact as output growth and stochastic discount rates are pro-cyclical. See Section 4.2.2 for details.

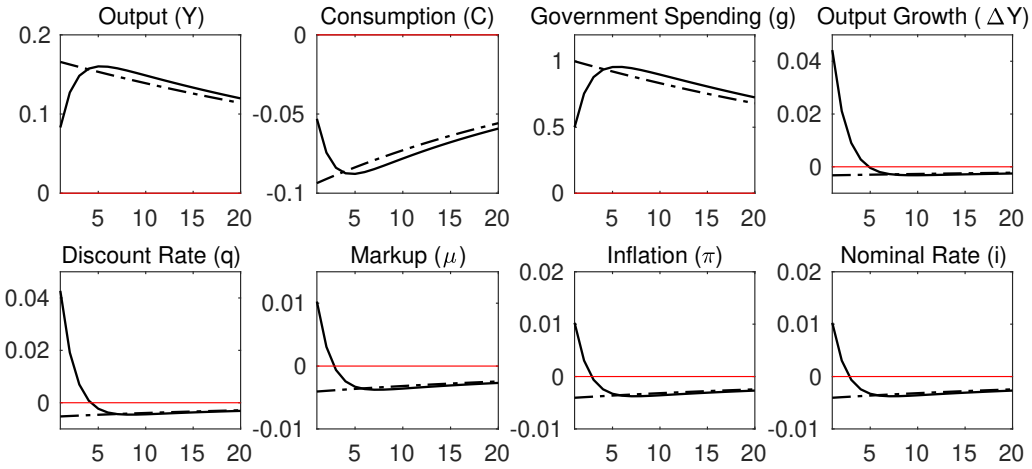
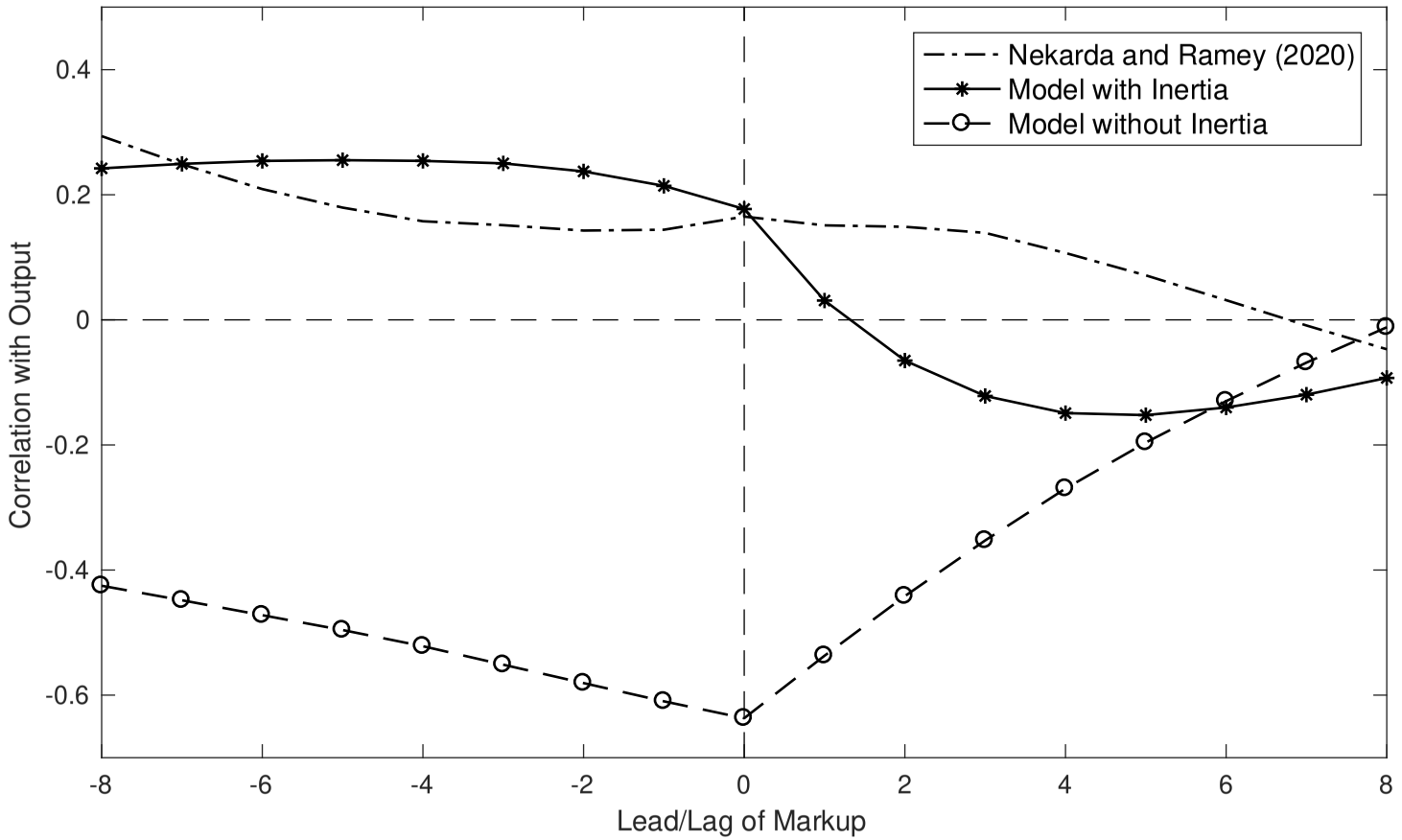


Figure 2: Cross-correlation of Markup and Output in the Implicit Collusion Model



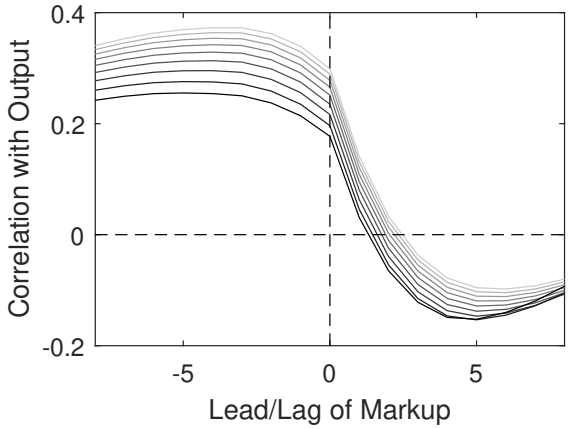
Notes: The black curve with square markers depicts correlation of μ_{t+j} with Y_t from the simulated implicit collusion model without inertial response of output conditional on TFP shocks. The dotted curve shows cross-correlation of the cyclical components of markups with real GDP from Nekarda and Ramey (2020) conditional on TFP shocks. The black curve with circle markers illustrate this cross-correlation from the simulated implicit collusion model with inertial response of output conditional on TFP shocks. Inertia is crucial in matching the positive correlation between output and markups from Nekarda and Ramey (2020). See Section 4.2.3 for details.

APPENDIX
(FOR ONLINE PUBLICATION)

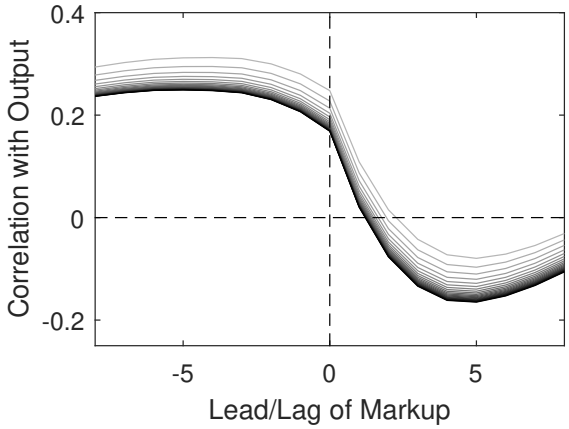
A Additional Figures

Figure A.1: Robustness to number of firms in sectors N, and the renegotiation probability γ

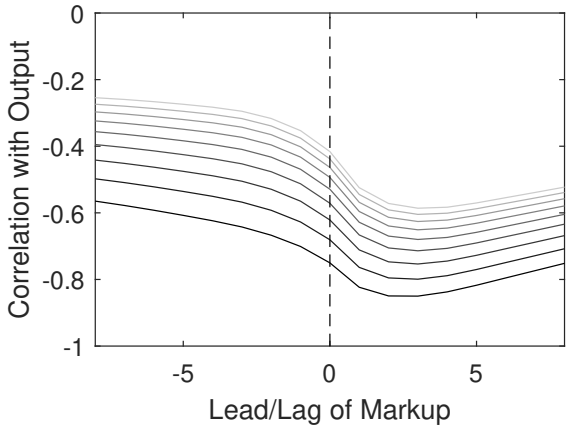
(a) Simulated correlation of μ_{t+j} with Y_t conditional on a TFP shock for $\gamma \in [0.4, 0.8]$. See section 4.2.4 for details.



(b) Simulated correlation of μ_{t+j} with Y_t conditional on a TFP shock for $N \in \{5, \ldots, 25\}$. See section 4.2.4 for details.



(c) Simulated correlation of μ_{t+j} with Y_t conditional on a government spending shock for $\gamma \in [0.4, 0.8]$. See section 4.2.4 for details.



(d) Simulated correlation of μ_{t+j} with Y_t conditional on a government spending shock for $N \in \{5, \ldots, 25\}$. See section 4.2.4 for details.

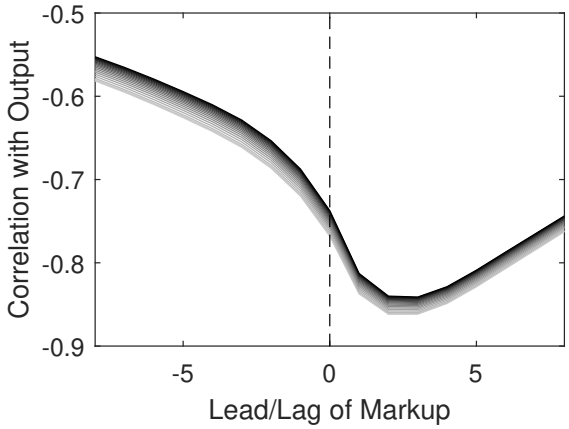
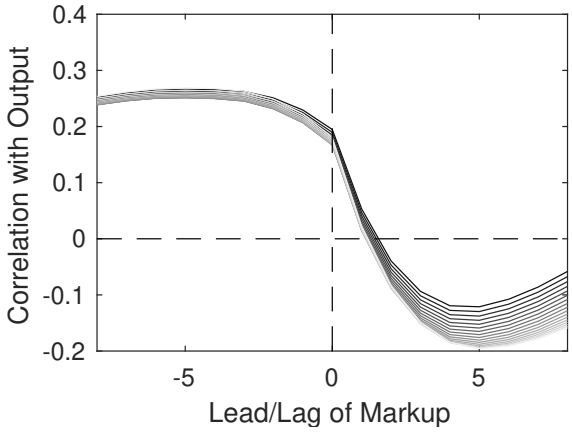
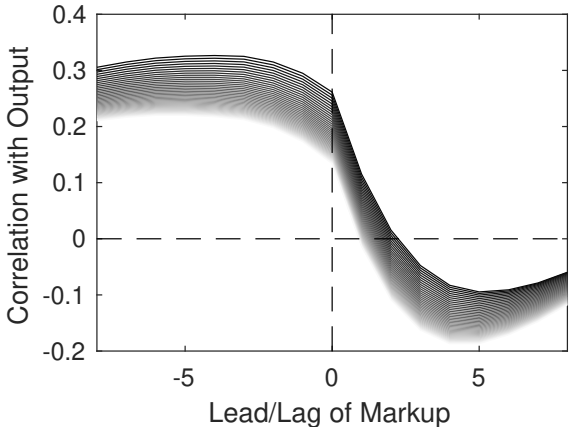


Figure A.2: Robustness to Elasticities of Substitution and Frisch Elasticity of Labor Supply

(a) Simulated correlation of μ_{t+j} with Y_t conditional on a TFP shock for $\sigma \in [2, 10]$. See section 4.2.4 for details.



(b) Simulated correlation of μ_{t+j} with Y_t conditional on a TFP shock for η between 10 and 30. See section 4.2.4 for details.



(c) Simulated correlation of μ_{t+j} with Y_t conditional on a government spending shock for $\epsilon \in [0.5, 5]$. See section 4.2.4 for details.

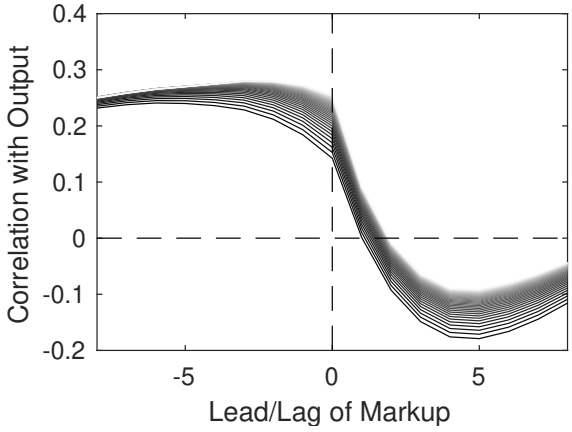
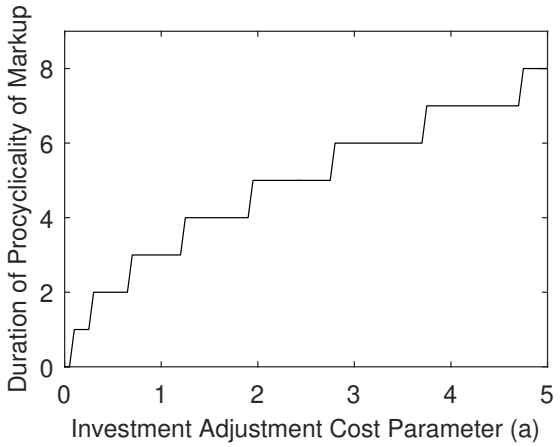
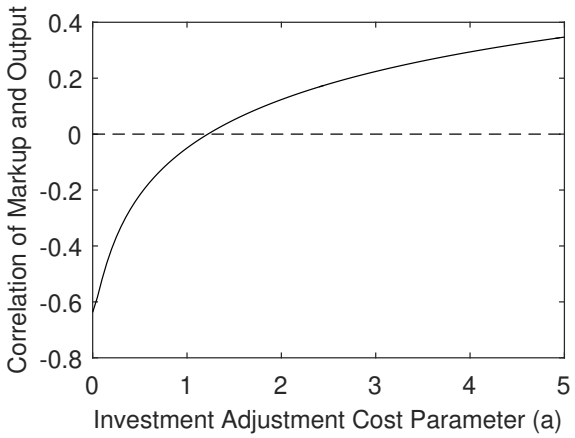


Figure A.3: Robustness to inertia

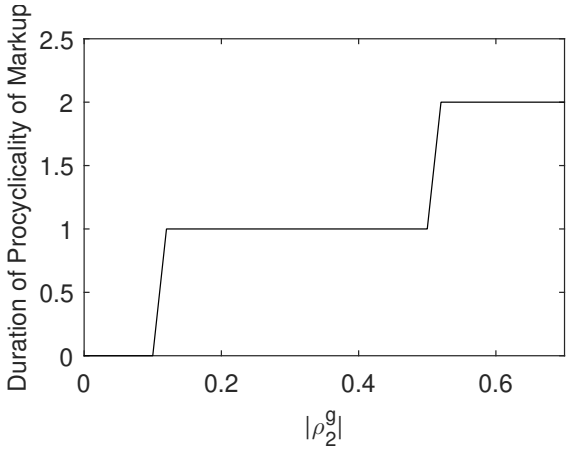
(a) Duration of procyclicality of markup after a 1% TFP shock for different value of investment adjustment cost parameter, $a \in [0,5]$. See Section 4.2.4 for details.



(b) Simulated correlation of μ_t with Y_t conditional on a TFP shock for different value of investment adjustment cost parameter, $a \in [0,5]$. See Section 4.2.4 for details.



(c) Duration of procyclicality of markup after a 1% government spending shock for different values of the inertia parameter in the AR(2) process, $|\rho_2^g| \in [0,0.7]$. See Section 4.2.4 for details.



(d) Simulated correlation of μ_t with Y_t conditional on a government spending shock for different values of the inertia parameter in the AR(2) process, $|\rho_2^g| \in [0,0.7]$. See Section 4.2.4 for details.

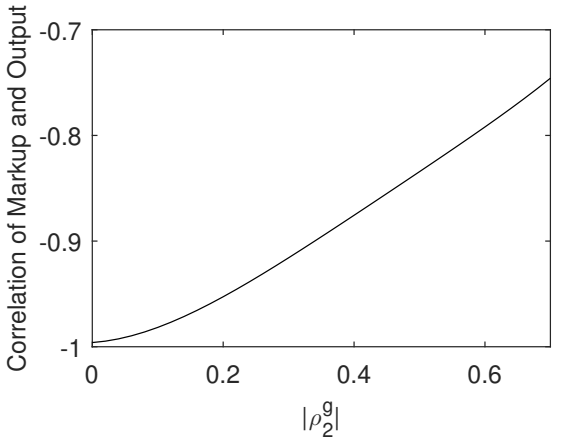
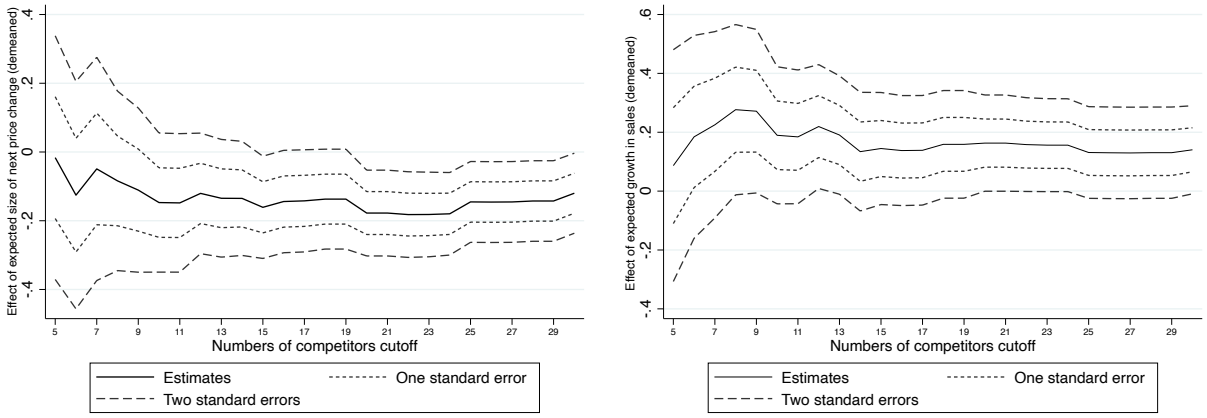


Figure A.4: Different cutoffs for N in regression for New Zealand

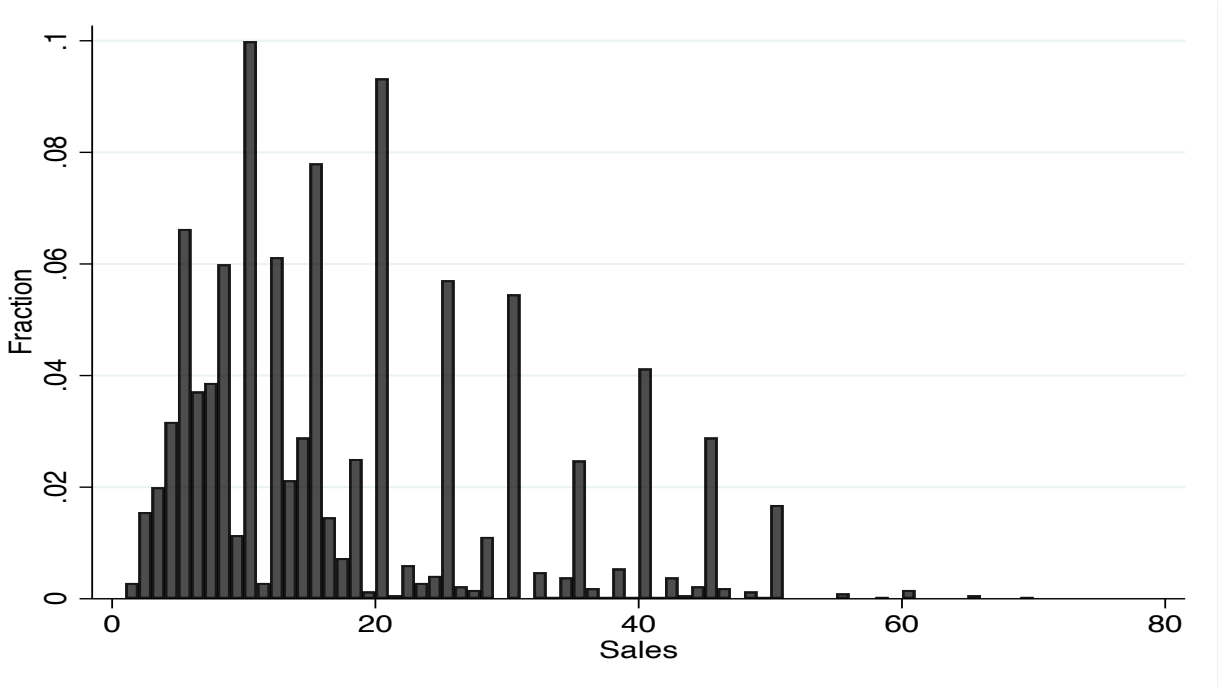


(a) effect of expected size of next price change

(b) effect of expected growth in sales

Notes: This figure plots the coefficients for the regression in Equation (16) for different cutoffs of number of competitors, while allowing for industry fixed effects. We have limited the regression for firms that report less than n but more than 2 competitors. The plot shows how the coefficients change as we pick different values for n.

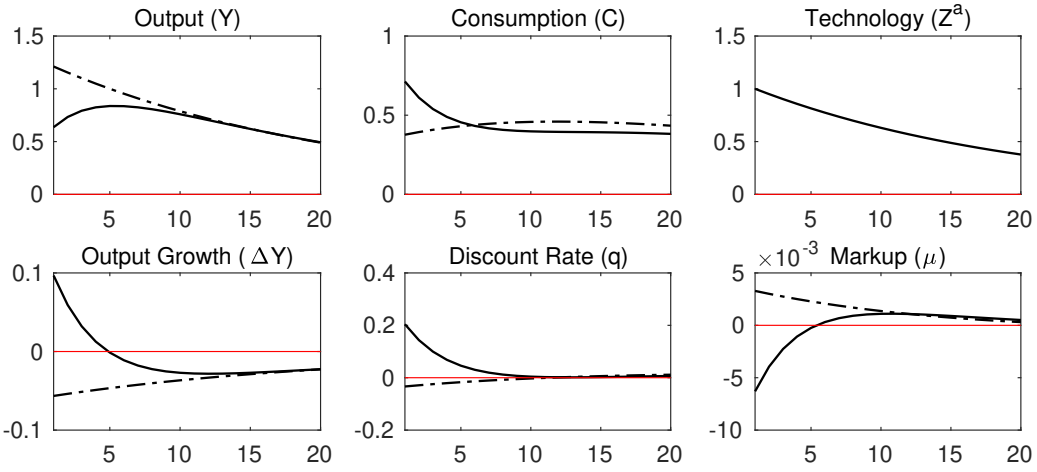
Figure A.5: Number of competitors histogram



Notes: This figure plots the histogram of the number of competitors that firms report they directly face in the New Zealand survey.

Figure A.6: Impulse Response Functions: Customer-Base Model

(a) The dashed curves plot the impulse response functions of the customer-base model to a 1% technology shock with no adjustment cost in which markups are pro-cyclical as output growth and stochastic discount rates are counter-cyclical. Solid curves illustrate the impulse response functions of the same model to a 1% technology shock with investment adjustment cost. Markups are counter-cyclical as long as firms expect output to grow. See Section 2.2.1 for details.



(b) The dashed curves plot the impulse response functions of the customer-base model to a 1% government spending shock without inertia in which markups are pro-cyclical as output growth is negative during the expansion. Solid curves illustrate the impulse response functions of the same model to an inertial government spending shock that peaks at 1%. Markups are counter-cyclical on impact as output growth and stochastic discount rates are pro-cyclical. See Section 2.2.1 for details.

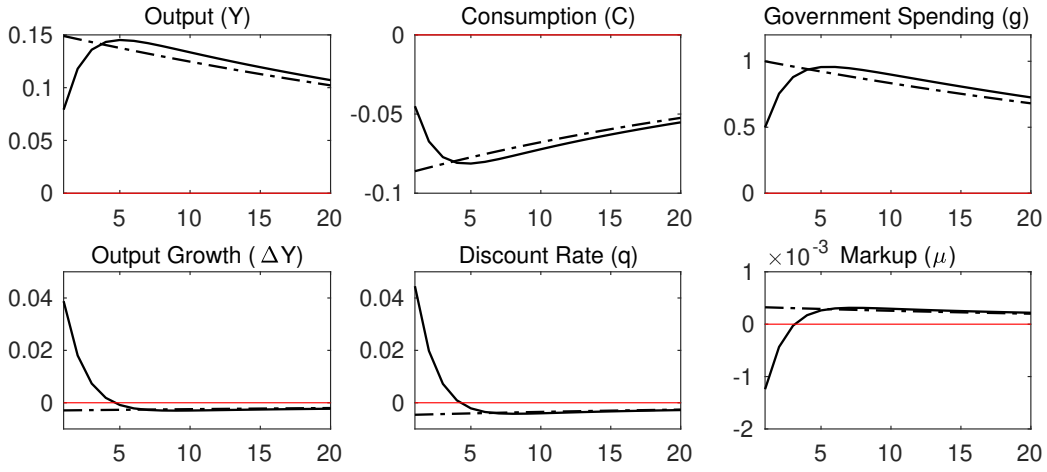
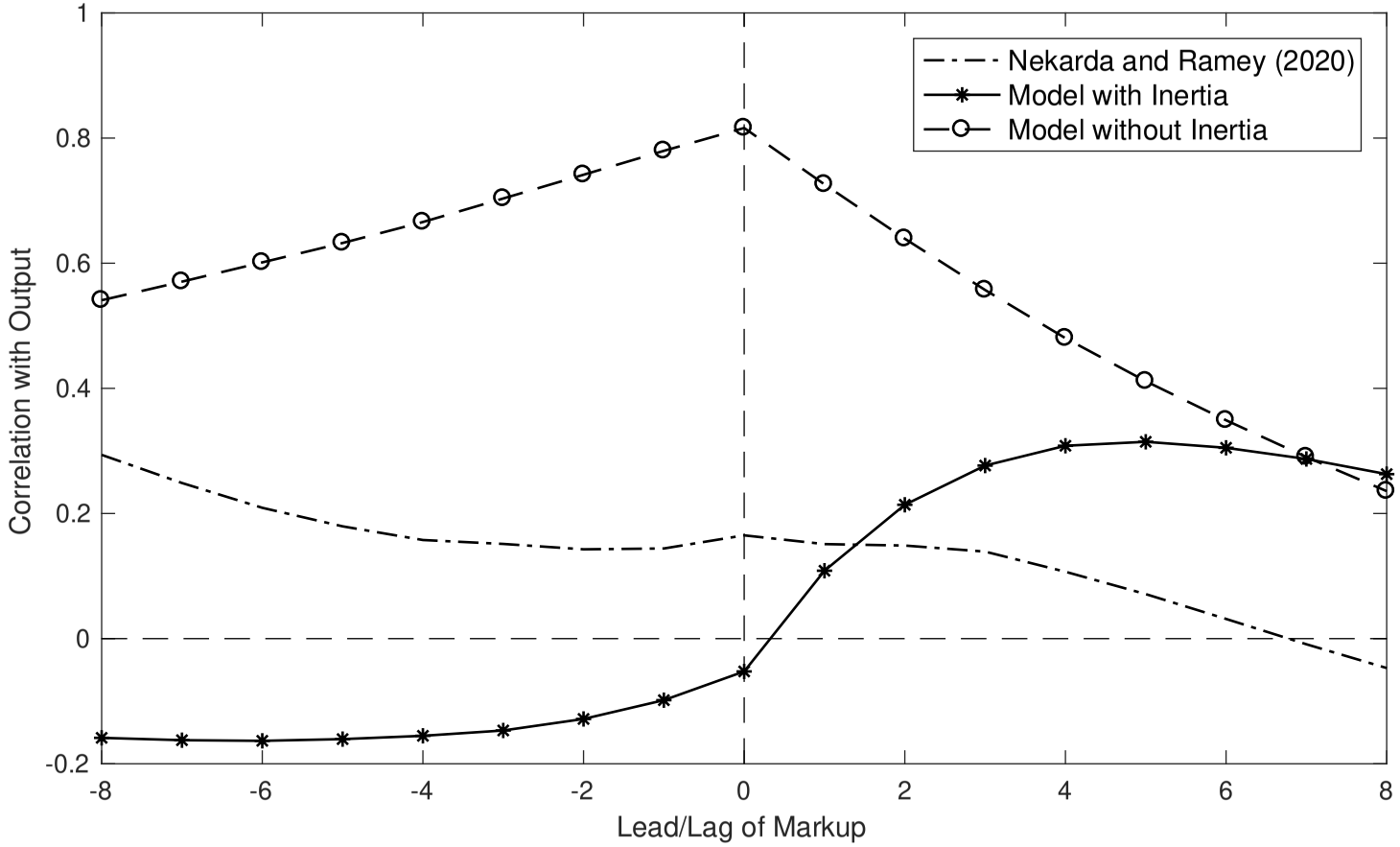
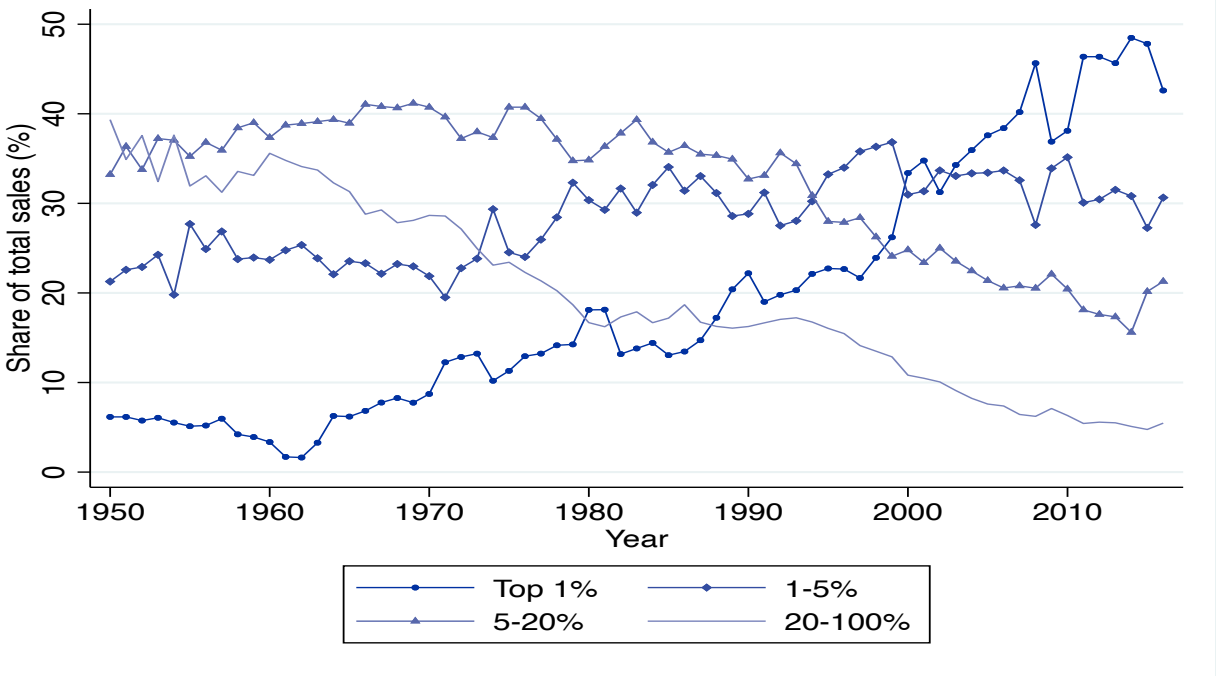


Figure A.8: Cross-correlation of Markup and Output in the Customer-base Model



Notes: The black curve with square markers depicts correlation of μ_{t+j} with Y_t from the simulated implicit collusion model without inertial response of output conditional on TFP shocks. The dotted curve shows cross-correlation of the cyclical components of markups with real GDP from Nekarda and Ramey (2020) conditional on TFP shocks. The black curve with circle markers illustrate this cross-correlation from the simulated implicit collusion model with inertial response of output conditional on TFP shocks. The customer-base model misses these conditional cross-correlations.

Figure A.9: Sales Share of Top Percentiles of Firms in Compustat



Notes: This figure plots the sales share different percentiles of firms over time in the Compustat data. By 2010s, the top 1% of firms account for around 45% of sales.

B Compustat: Variables Selection and Construction

We download and construct the following variables from Compustat:

- *Global company key* (mnemonic `gvkey`): Compustat's firm id.
- *Year* (mnemonic `fyear`): the fiscal year.
- *Costs of goods sold* (mnemonic `COGS`): the COGS sums all "expenses that are directly related to the cost of merchandise purchased or the cost of goods manufactured that are withdrawn from finished goods inventory and sold to customers." They include expenses such as labor and related expenses (including salary, pension, retirement, profit sharing, provision for bonus and stock options, and other employee benefits), operating expense, lease, rent, and loyalty expense, write-downs of oil and gas properties, and distributional and editorial expenses.
- *Operating expenses, total* (mnemonic `XOPR`): OPEX represents the sum of COGS, SG&A and other operating expenses.
- *Sales (net)* (mnemonic `SALE`): this variable represents gross sales, for which "cash discounts, trade discounts, and returned sales and allowances for which credit is given to customer" are discounted from the final value.
- *Assets, total* (mnemonic `AT`).
- *Standard industry classification code* (mnemonic `SIC`): the `SIC` is a four-digit classification of a company's operations.
- *Debt in current liabilities* (mnemonic `DLC`): this variable represents "the total amount of short-term notes" and the portion of the long-term debt that is due in one year.
- *Long-term debt* (mnemonic `DLC`): all debt obligation that is due in more than one year from the company's balance sheet date.
- *Total debt*: we define total debt as the sum of the debt in current liabilities and long-term debt.
- *Leverage*: we follow Ottonello and Winberry (2020) and define it as the debt-to-asset ratio, where debt is the total debt described above and assets is the book value of assets.
- *HHI index*: we calculate the HHI index as the sum of the squared market share of each firm, where we have used a SIC 2-digits industry-specific market share.
- *markup*: following De Loecker et al. (2018), we first estimate time-invariant but industry-specific (SIC 2-digits) output elasticities using the production function estimation method from De Loecker et al. (2018). We then define markup as output elasticities $\times \frac{\text{sales}_{it}}{\text{COGS}_{it}}$

We used NIPA Table 1.1.9. GDP deflator (line 1) to generate the real value for the variables `sale`, `COGS`, `XOPR`.

B.1 Compustat Sample Selection

We downloaded the dataset "Compustat Annual Updates: Fundamentals Annual," from Wharton Research Data Services, from Jan 1950 to Dec 2016. The following options were chosen:

- Consolidated level: C (consolidated)
- Industry format: INDL (industrial)
- Data format: STD (standardized)
- Population source: D (domestic)
- Currency: USD
- Company status: active and inactive

We took the following steps for the cleaning process:

1. To select American companies, we filtered the dataset for companies with Foreign Incorporation Code (FIC) equal to "USA."
2. We replace industry variables (`sic` and `naics`) by their historical values whenever the historical value is not missing.
3. We drop utilities (`sic` value in the range [4900, 4999]) because their prices are very regulated and financials (`sic` value in the range [6000,6999]) because their balance sheets are exceptionally different than the other firms in the analysis.
4. To ensure quality of the data, we drop missing or non-positive observations for sales, COGS, OPEX, sic 2-digit code, gross PPE, net PPE, and assets. For each year, we exclude the top bottom and top 1% of the COGS-to-sales ratio and the SG&A-to-sales ratio. We also exclude observations in which acquisitions are more than 5% of the total assets of a firm.
5. A portion of the data missing for sales, COGS, OPEX, and capital in between years for firms. We input these values using a linear interpolation, but we do not interpolate for gaps longer than two years. This exercise inputs data for 4.6% of our sample.

The final data set contains 242,155 observations for 20,252 firms across 67 years.

C Proofs

Proof of Proposition 1.

First, observe that the set of solutions is not empty as $\mu_{i,t} = \mu_{COU}, \forall\, t$ satisfies the constraint for all periods. Moreover, if the constraint is not binding, the firms will simply act like a monopoly and choose $\mu_{i,t} = \mu_{MON}$, as it maximizes their joint profits. Hence, the choice set of firms can be compactified so that $\mu_{i,t} \in [\mu_{COU}, \mu_{MON}]$, and as the usual assumption of continuity holds, the problem has a solution. Finally, for the solution to be a sub-game Nash equilibrium, two conditions have to hold: first, that firms do not have an incentive to deviate from the chosen markups in the equilibrium path, which is true by construction, and second, that if ever the game were to go to punishment stage, firms would have an incentive to revert back to this strategy, which is also true as collusion is always at least as good as best responding.

Proof of Proposition 2.

Taking the first order optimality conditions for a firm's problem and imposing the symmetry conditions $\mu_t = \mu_{i,t} = \mu_{i,j,t}$, and $S_{i,j,t} = 1$, $\forall\, i, j$, we get

$$\mu_t^{-1} - \mu_s^{-1} = \frac{\zeta}{1+\zeta}(1-\mu_s^{-1}) + \frac{\beta\gamma}{1+\zeta}\mathbb{E}_t\left[Q_{t,t+1}\frac{Y_{t+1}}{Y_t}(\mu_{t+1}^{-1} - \mu_s^{-1})\right]$$

where $\mu_s = \frac{\eta(1-N^{-1})+\sigma N^{-1}}{(\eta-1)(1-N^{-1})+(\sigma-1)N^{-1}}$ is the markup of a firm with no inertia in their demand ($\gamma = 0$). Hence, in the steady state

$$\mu^{-1} - \mu_s^{-1} = \frac{\zeta(1-\mu_s^{-1})}{1+\zeta-\beta\gamma} \tag{C.1}$$

Taking a first order approximation to the first order condition above and replacing μ from Equation (C.1) we get the law of motion in the proposition along with coefficients ψ_1 and ψ_2. Moreover, the comparative statics with respect to N follow directly from the fact that μ_s is decreasing with N.

D Regression Specification in New Zealand Survey

Let industries be indexed by i and firms within them be indexed by j. Consider the following regression

$$\begin{aligned}\hat{\mu}_{ij} - \sum_i\sum_j \hat{\mu}_{ij} &= Industry\_FE_i + \beta_1\{Ex\Delta Sales_{ij} - \sum_i\sum_j Ex\Delta Sales_{ij}\} \\ &+ \beta_2\{Ex\Delta Price_{ij} - \sum_i\sum_j Ex\Delta Price_{ij}\} + \varepsilon_{ij}\end{aligned}$$

where $\hat{\mu}_{ij}$ is the deviation of current markup of firm ij from its average level, $Ex\Delta Sales_{ij}$ is the expected growth in sales for firm ij, and $Ex\Delta Price_{ij}$ is its next expected price change. Now

consider the following decomposition of firms' errors in expecting stochastic discount rates and changes in marginal costs:

$$\mathbb{E}_t^{ij}\{\hat{q}_{t,t+1}\} - \sum_i \sum_j \mathbb{E}_t^{ij}\{\hat{q}_{t,t+1}\} = u_{1,t}^i + u_{2,t}^{ij}$$
$$\mathbb{E}_t^{ij}\{\Delta\hat{mc}_{t+1}\} - \sum_i \sum_j \mathbb{E}_t^{ij}\{\Delta\hat{mc}_{t+1}\} = v_{1,t}^i + v_{2,t}^{ij}$$

where $u_{1,t}^i$ and $v_{1,t}^i$ are industry specific errors that are orthogonal to the firm level errors $v_{2,t}^{ij}$ and $u_{2,t}^{ij}$. Assuming that $v_{2,t}^{ij}$ and $u_{2,t}^{ij}$ are independent across firms and are orthogonal to the other terms in the above regression, ψ_1 and ψ_2 can be identified from β_1 and β_2 up to the elasticity of substitution across sectors, σ.

To see why, notice that

$$\begin{aligned} Ex\Delta Sales_{i,j,t} &= E_t^{ij}\frac{P_{i,j,t+1}Y_{i,j,t+1} - P_{i,j,t}Y_{i,j,t}}{P_{i,j,t}Y_{i,j,t}} \\ &\approx E_t^{ij}\left[(1-\sigma)\Delta\hat{p}_{i,j,t+1} + \Delta\hat{y}_{t+1}\right] \\ &= (1-\sigma)Ex\Delta Price_{i,j,t} + E_t^{ij}\left[\Delta\hat{y}_{t+1}\right] \end{aligned}$$

where the second line is derived using the demand structure $Y_{i,j,t} = Y_{i,t} = Y_t D(P_{i,t}; P_{i,t})$. Now, rewriting the law of motion

$$\begin{aligned} \hat{\mu}_{i,j,t} &= \frac{\psi_1}{1-\psi_2}\mathbb{E}_t^{ij}\left\{\Delta\hat{y}_{t+1} + \hat{q}_{t,t+1}\right\} + \frac{\psi_2}{1-\psi_2}\mathbb{E}_t^{ij}\{\Delta\hat{\mu}_{i,t+1}\} \\ &= \frac{\psi_1}{1-\psi_2}\mathbb{E}_t^{ij}\left\{Ex\Delta Sales_{i,j,t} + (\sigma-1)\Delta\hat{p}_{i,t+1} + \hat{q}_{t,t+1}\right\} + \frac{\psi_2}{1-\psi_2}\mathbb{E}_t^{ij}\{\Delta\hat{p}_{i,j,t+1} - \Delta\hat{mc}_{t+1}\} \\ &= \frac{\psi_1}{1-\psi_2}\mathbb{E}_t^{ij}\left\{\hat{q}_{t,t+1}\right\} + \frac{\psi_1}{1-\psi_2}Ex\Delta Sales_{i,j,t} + \frac{(\sigma-1)\psi_1+\psi_2}{1-\psi_2}Ex\Delta Price_{i,j,t} - \frac{\psi_2}{1-\psi_2}\mathbb{E}_t^{ij}\{\Delta\hat{mc}_{t+1}\} \end{aligned}$$

Now sum over i and j and subtract the two to get

$$\begin{aligned} \hat{\mu}_{ij} - \sum_i\sum_j \hat{\mu}_{ij} &= \frac{\psi_1}{1-\psi_2}\{Ex\Delta Sales_{ij} - \sum_i\sum_j Ex\Delta Sales_{ij}\} \\ &+ \frac{(\sigma-1)\psi_1+\psi_2}{1-\psi_2}\{Ex\Delta Price_{ij} - \sum_i\sum_j Ex\Delta Price_{ij}\} \\ &+ \frac{\psi_1}{1-\psi_2}(\mathbb{E}_t^{ij}\{\hat{q}_{t,t+1}\} \\ &- \sum_i\sum_j \mathbb{E}_t^{ij}\{\hat{q}_{t,t+1}\}) - \frac{\psi_2}{1-\psi_2}(\mathbb{E}_t^{ij}\{\Delta\hat{mc}_{t+1}\} - \sum_i\sum_j \mathbb{E}_t^{ij}\{\Delta\hat{mc}_{t+1}\}) \\ &= \frac{\psi_1}{1-\psi_2}\{Ex\Delta Sales_{ij} - \sum_i\sum_j Ex\Delta Sales_{ij}\} \\ &+ \frac{(\sigma-1)\psi_1+\psi_2}{1-\psi_2}\{Ex\Delta Price_{ij} - \sum_i\sum_j Ex\Delta Price_{ij}\} + Industry\_FE_i + \varepsilon_{i,j,t} \end{aligned}$$

where

$$Industry\_FE_i \equiv \tfrac{\psi_1}{1-\psi_2}u_{1,t}^i + \tfrac{\psi_2}{1-\psi_2}v_{1,t}^i$$

$$, \qquad \varepsilon_{i,j,t} \equiv \frac{\psi_1}{1-\psi_2} u^{ij}_{2,t} + \frac{\psi_2}{1-\psi_2} v^{ij}_{2,t}$$

Since $u^{ij}_{2,t}$ and $v^{ij}_{2,t}$ are independent of $Indsutry\_FE_i$ by construction and the other two terms by assumption, we have,

$$\psi_1 = \frac{\hat{\beta}_1}{1+\hat{\beta}_2-(\sigma-1)\hat{\beta}_1}, \psi_2 = 1 - \frac{1}{1+\hat{\beta}_2-(\sigma-1)\hat{\beta}_1}$$

E Deviations from Full Information Rational Expectations about Aggregates

In this section, we show that as long as firms know their own future sales growths up to full information rational expectations (FIRE), even if their aggregate expectations do not coincide with FIRE, the law of motion holds in aggregate with FIRE as well. In other words, there is no need to assume that firms know everything in the economy up to FIRE.

To see this, suppose firms within a sector face sector specific demand or supply shocks so that the sectoral output is not necessarily the same as the aggregate output. Moreover, suppose that firms within sectors do not necessarily have full information rational expectations but share the same expectation operator with their competitors (so that there are no imperfect common knowledge issues confounding the problem). Then, we can write the incentive compatibility constraint in the implicit collusion model as

$$(\rho_i - \mu_{i,t}^{-1}) D(\rho_i;1) - N^{-1}(1-\mu_{i,t}^{-1}) \leq \beta\gamma \mathbb{E}_{i,t} Q^i_{t,t+1} \frac{Y_{i,t+1}}{Y_{i,t}} \Gamma_{i,t+1}$$

$$\Gamma_{i,t} = N^{-1}(\mu_s^{-1} - \mu_{i,t}^{-1}) + \beta\gamma \mathbb{E}_{i,t} Q^i_{t,t+1} \frac{Y_{i,t+1}}{Y_{i,t}} \Gamma_{i,t+1}$$

where, for simplicity, we have assumed $\sigma = 1$ ($\sigma > 1$ would require firms to make forecasts of how sales are reallocated across the aggregate economy which adds another layer of complexity to this simple example and for now we abstract away from it to focus on our aggregation result). It follows that, up to a first order approximation,

$$\hat{\mu}_{i,t} = \psi_1 \mathbb{E}_{i,t}[\Delta\hat{y}_{i,t+1} + \Delta\hat{q}_{i,t,t+1}] + \psi_2 \mathbb{E}_{i,t}[\hat{\mu}_{i,t+1}]$$

whre $\Delta\hat{y}_{i,t+1}$ and $\Delta\hat{q}_{i,t,t+1}$ are sectoral output growth and discount factors respectively and $\mathbb{E}_{i,t}[.]$ is the firms' expectation operator in sector i. Let us define

$$\xi_{i,t+h} \equiv \Delta\hat{y}_{t+h} + \Delta\hat{q}_{t,t+h} - \Delta\hat{y}_{i,t+h} - \Delta\hat{q}_{i,t,t+h}$$

as the wedge between economywide growth in output and sector i's growth in output, and let us assume that $\xi_{i,t+h}$ is orthogonal to the economywide output growth and stochastic discount rate. Notice that the variance of $\xi_{i,t+h}$ determines how much sectoral level variables deviate from aggregate variables.

New evidence on firms' expectations (see for instanace Meyer, Parker and Sheng (2021)) shows that firms are very well aware of their own environment. So let us assume that

$$\mathbb{E}_{i,t}[\Delta\hat{y}_{i,t+h} + \Delta\hat{q}_{i,t,t+h}] = \mathbb{E}_t^f[\Delta\hat{y}_{i,t+h} + \Delta\hat{q}_{i,t,t+h}]$$

where $\mathbb{E}_t^f[.]$ is the FIRE operatore. Notice that this simply assumes that firms are very well aware of their own environement and is much weaker than assuming that firms know aggregate variables according to FIRE. In fact, it only requires the firm's expectations about their own future sales growths to coincide with FIRE, but does not impose a restriction on how informed the firm should be about aggregates. For instance, if variance of $\xi_{i,t+h}$ is large, then all firms could have very well-informed expectations about their own sales but since aggregates are only known up to $\xi_{i,t+h}$, their expectations of aggregate variables will be very noisy. Nonetheless, we can use the above equation to achieve the following aggregation results:

$$\int_i \mathbb{E}_{i,t}[\Delta\hat{y}_{i,t+h} + \Delta\hat{q}_{i,t,t+h}]di = \int_i \mathbb{E}_t^f[\Delta\hat{y}_{t+h} + \Delta\hat{q}_{t,t+h}]di + \underbrace{\int_i \mathbb{E}_t^f[\xi_{i,t+h}]di}_{=0}$$

Meaning that if firms only know their own sales growths up to FIRE, their average expectations will collapse to expectations of aggregate sales growth based on FIRE, eventhough that no firms knows the aggregates perfectly. Hence, while FIRE does not hold about aggregate variables at the firm level, it holds at the aggregate level and we can derive the law of motion as before

$$\begin{aligned}
\hat{\mu}_t &\equiv \int_i \hat{\mu}_{i,t}di \\
&= \psi_1 \sum_{h=1}^{\infty} \psi_2^h \int_i \mathbb{E}_{i,t}[\Delta\hat{y}_{i,t+h} + \Delta\hat{q}_{i,t,t+h}]di \\
&= \psi_1 \sum_{h=1}^{\infty} \psi_2^h \mathbb{E}_t^f[\Delta\hat{y}_{t+h} + \Delta\hat{q}_{t,t+h}]di \\
&= \psi_1 \mathbb{E}_t^f[\Delta\hat{y}_{t+1} + \Delta\hat{q}_{t,t+1}] + \psi_2 \mathbb{E}_t^f[\hat{\mu}_{t+1}]
\end{aligned}$$