Endogenous Firm Competition and the Cyclicality of Markups*

Hassan Afrouzi† Columbia University

Luigi Caloi‡ Columbia University

November 8, 2022

*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.

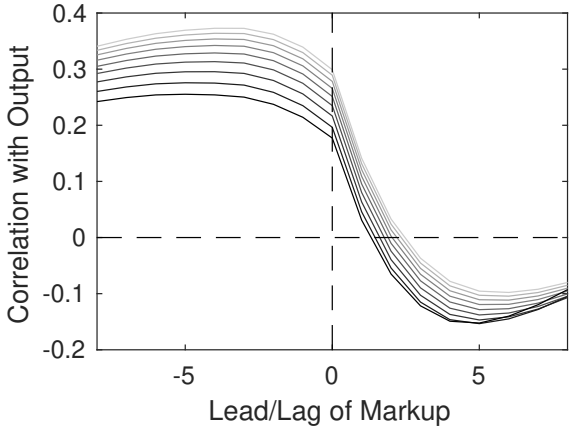
†Columbia University, Department of Economics, 420 West 118 St. New York, NY 10027. Email: hassan.afrouzi@columbia.edu.

‡Columbia University, Department of Economics, 420 West 118 St. New York, NY 10027.

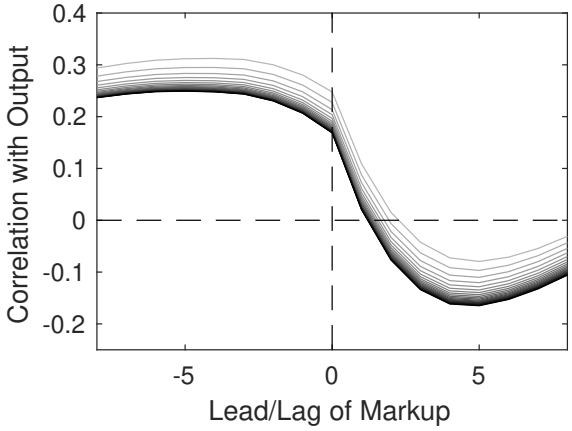
APPENDIX
(FOR ONLINE PUBLICATION)

A Additional Figures

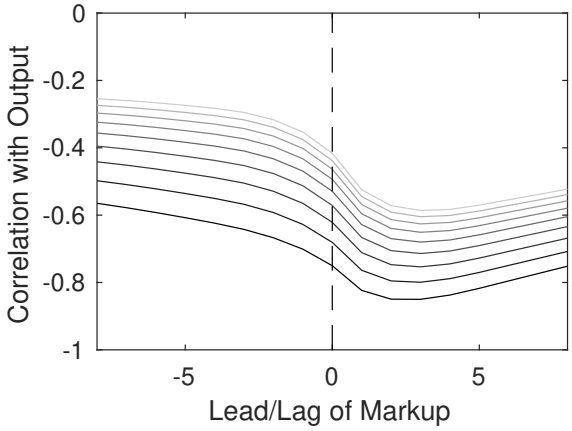
Figure 1: Robustness to number of firms in sectors N, and the renegotiation probability γ



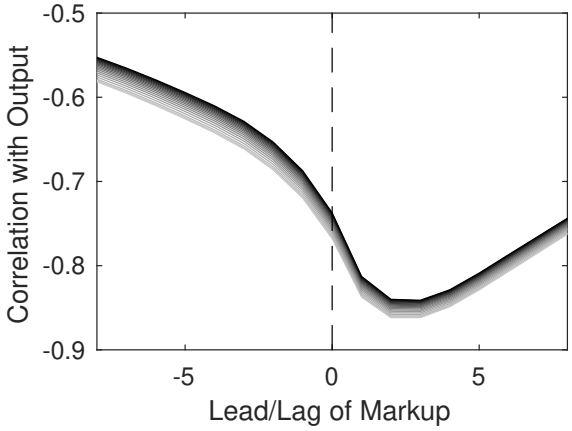
(a) Simulated correlation of μ_{t+j} with Y_t conditional on a TFP shock for $\gamma \in [0.4, 0.8]$. See section C for details.



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

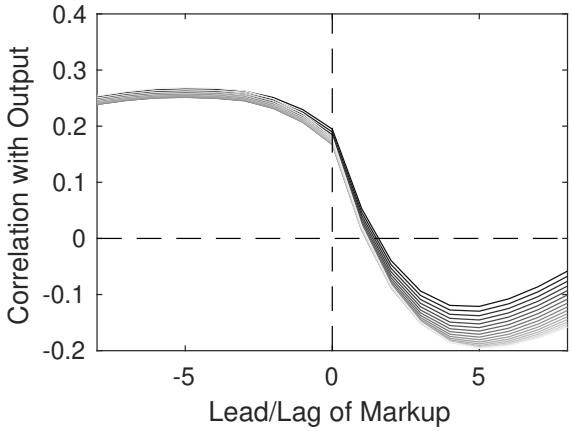


(c) Simulated correlation of μ_{t+j} with Y_t conditional on a government spending shock for $\gamma \in [0.4, 0.8]$. See section C for details.

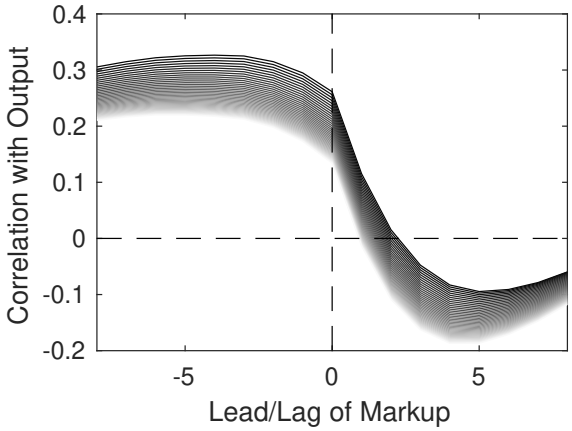


(d) Simulated correlation of μ_{t+j} with Y_t conditional on a government spending shock for $N \in \{5, \ldots, 25\}$. See section C for details.

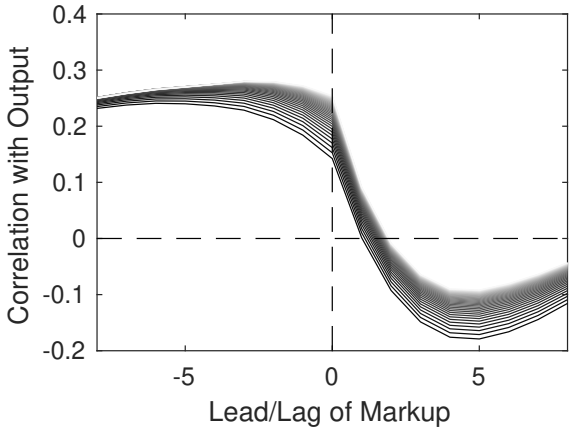
Figure 2: Robustness to Elasticities of Substitution and Frisch Elasticity of Labor Supply



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

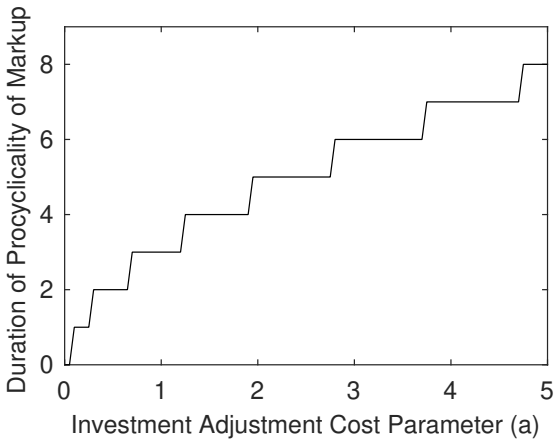


(b) Simulated correlation of μ_{t+j} with Y_t conditional on a TFP shock for η between 10 and 30. See section C for details.

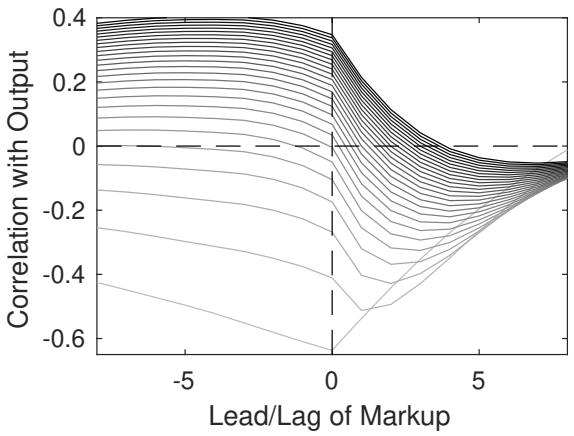


(c) Simulated correlation of μ_{t+j} with Y_t conditional on a government spending shock for $\epsilon \in [0.5, 5]$. See section C for details.

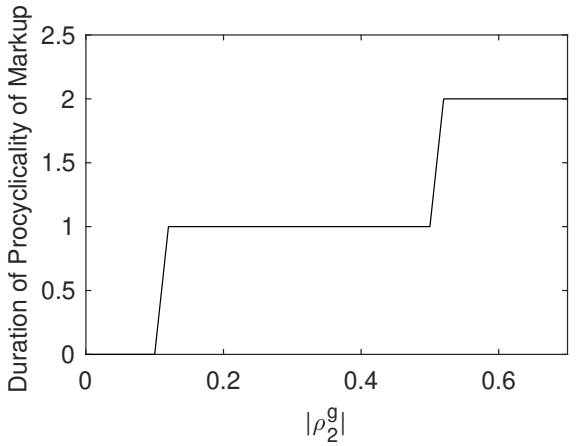
Figure 3: Robustness to inertia



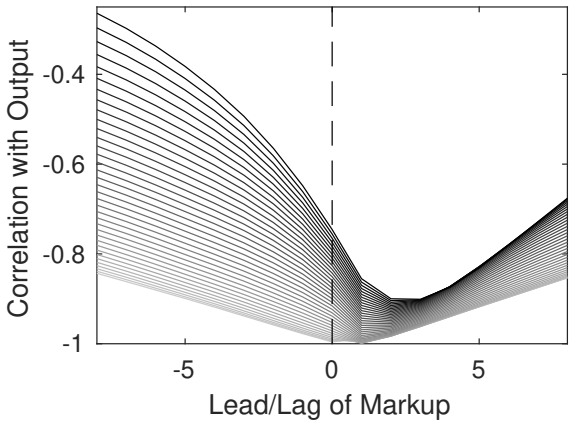
(a) Duration of procyclicality of markup after a 1% TFP shock for different value of investment adjustment cost parameter, $a \in [0, 5]$. See Section C for details.



(b) Simulated correlation of μ_{t+j} with Y_t conditional on a TFP shock for different value of investment adjustment cost parameter, $a \in [0, 5]$. See Section C for details.

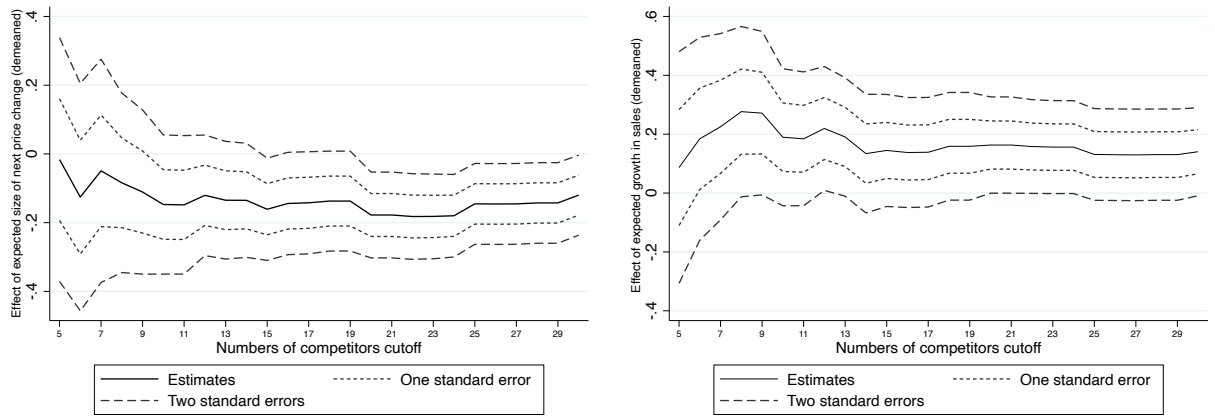


(c) Duration of procyclicality of markup after a 1% government spending shock for different values of the inertia parameter in the AR(2) process, $|\rho_2^g| \in [0, 0.7]$. See Section C for details.



(d) Simulated correlation of μ_{t+j} 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 C for details.

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

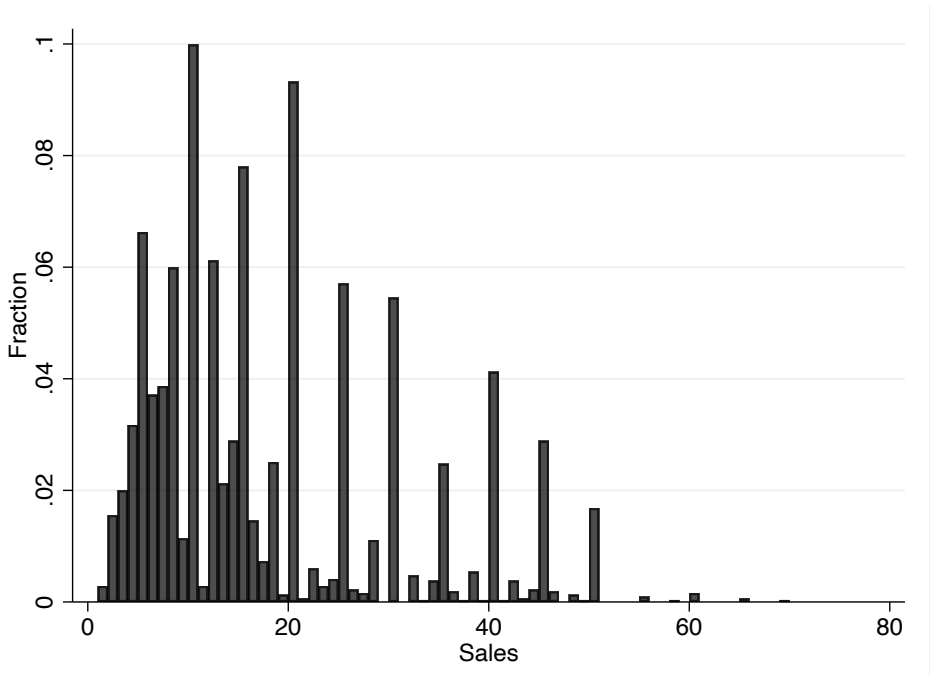


(a) effect of expected size of next price change

(b) effect of expected growth in sales

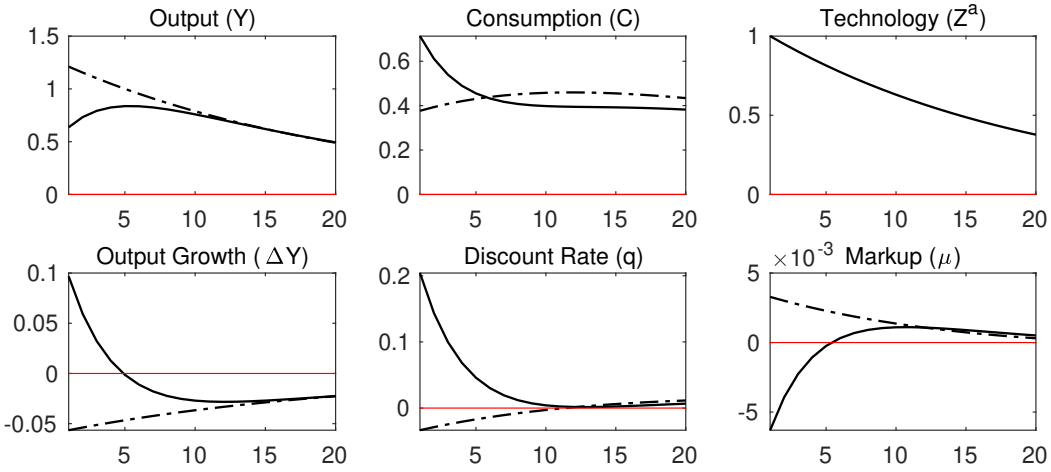
Notes: This figure plots the coefficients for the regression in Equation (15) for different cutoffs of number of competitors, while allowing for industry fixed effects. We have limited the regression for firms that report less than n but more than 2 competitors. The plot shows how the coefficients change as we pick different values for n.

Figure 5: Number of competitors histogram

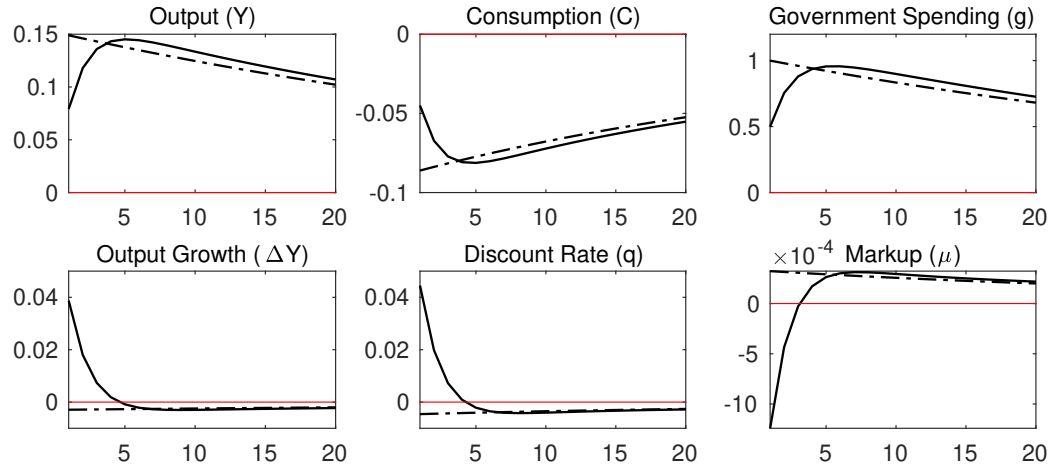


Notes: This figure plots the histogram of the number of competitors that firms report they directly face in the New Zealand survey.

Figure 6: Impulse Response Functions: Customer-Base Model

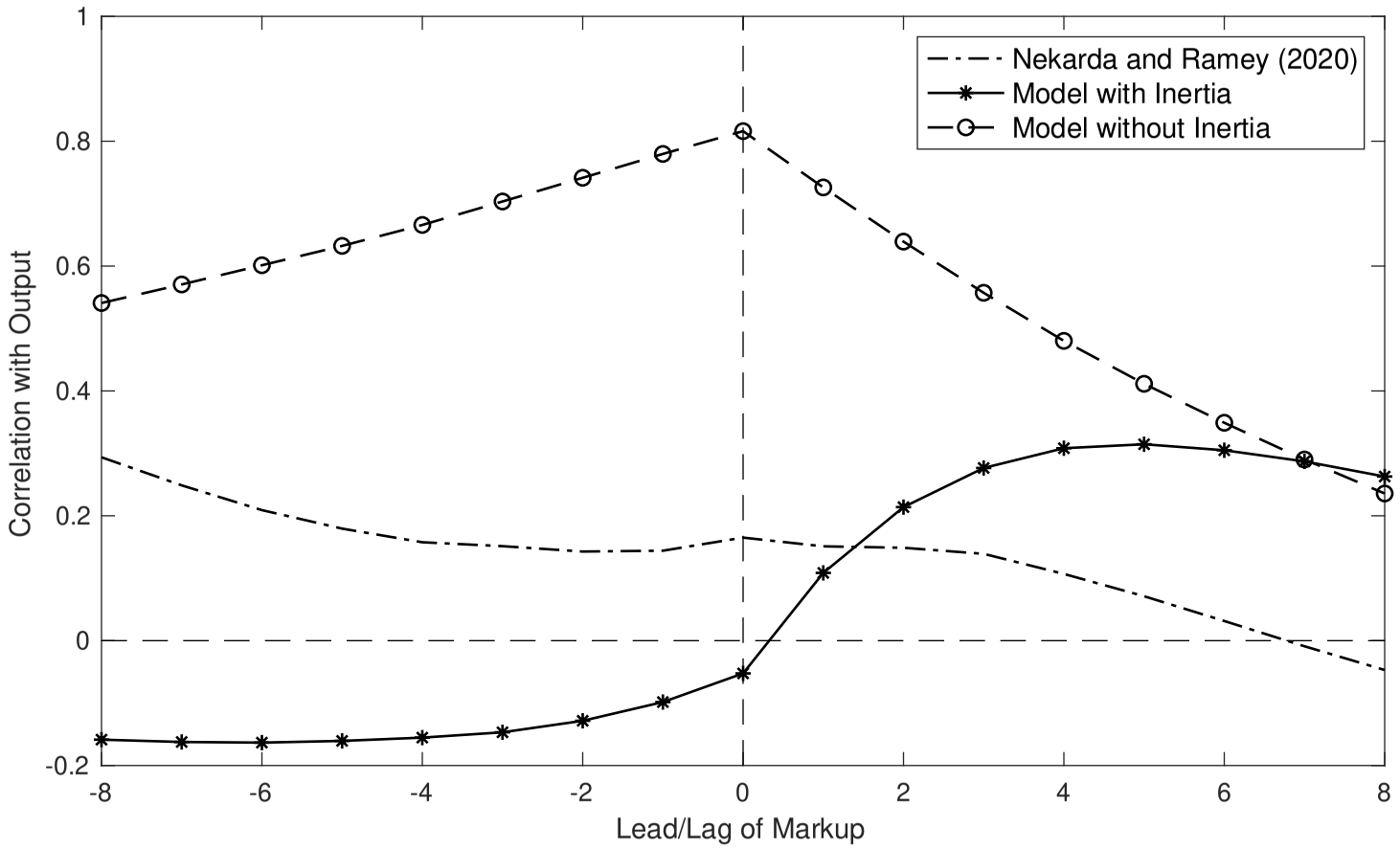


(a) The dashed curves plot the impulse response functions of the customer-base model to a 1% technology shock with no adjustment cost in which markups are pro-cyclical as output growth and stochastic discount rates are counter-cyclical. Solid curves illustrate the impulse response functions of the same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



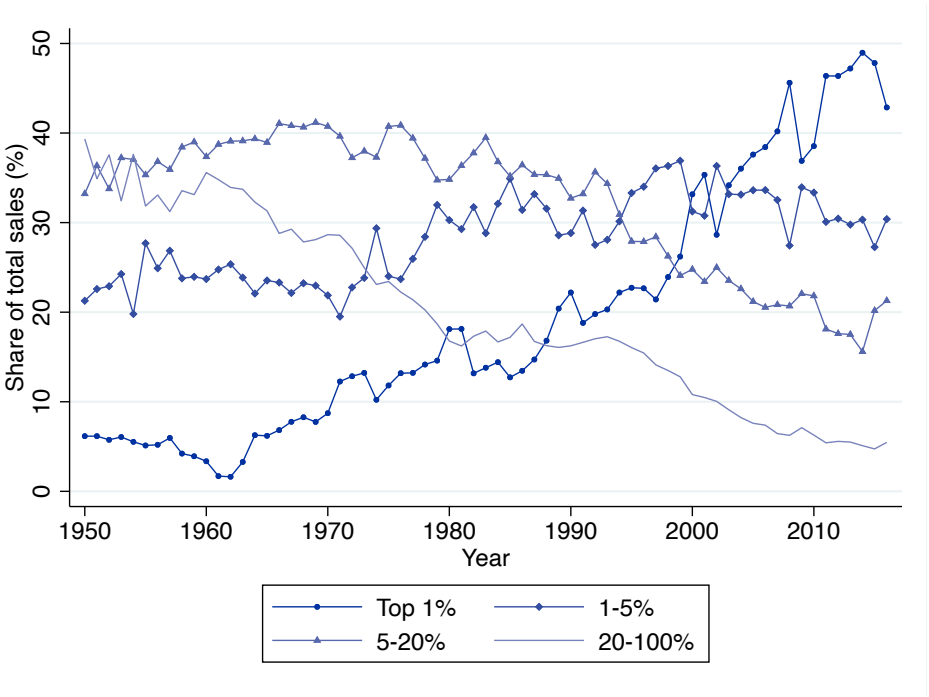
(b) The dashed curves plot the impulse response functions of the customer-base model to a 1% government spending shock without inertia in which markups are pro-cyclical as output growth is negative during the expansion. Solid curves illustrate the impulse response functions of the same model to an inertial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

Figure 7: 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 8: 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 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

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

Taking a first order approximation to the first order condition above and replacing μ from Equation (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.

C Model Robustness Checks

Here, we check the robustness of the model predictions with respect to different parameters.

Probability of Renegotiation (Discount Factor). Figures 1a and 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, 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 1b and 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 2a, 2b and 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 as 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 et al. (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 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 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 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 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.

D 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`.

D.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 one year. 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.

E Challenges with Measuring Markups using Compustat Data

In this section, we start with a brief overview of the different methods to estimate markups, and we then argue why we think Compustat and the production function approach are still the most appropriate way to measure markups for answering the questions in this paper.

E.1 Methods to Measure Markups

There are multiple methods that have been employed to estimate markups. First, the accounting approach relies on gross (or net) margins of profits and has the benefit of being easy to implement.[1] However, this method doesn't estimate marginal cost of production and rely instead in the average cost. Another approach, commonly used in the modern industrial organization literature, relies on a specification of the demand system to generate price elasticities of demand.[2] While these methods might be relevant for other studies, we (and De Loecker et al. (2020)) do not want to impose how

[1] See Karabarbounis and Neiman (2014) for a recent implementation and De Loecker et al. (2020) for a further discussion of the benefits and issues with this method.

[2] See, for instance, Berry et al. (1995) for an implementation. Again, see De Loecker et al. (2020) for a discussion of the merits and pitfalls of this method.

firms compete when analyzing data across several different industries and time periods. Instead, we rely on a production function approach which relies on cost-minimization to generate an equation for markup.

Within the approaches that use cost-minimization, there are three leading methods. To investigate them, Basu (2019) provide a unifying framework. Consider a firm that produces output Y using capital K, labor L and technology Z via the production function $F(K, L, Z)$. Given that the profits are $F(K, L, Z) \times P - K \times R - L \times W$, and assuming that firms take the price of capital R and of labor W as given, then a profit maximizing firm will set a cost-minimizing use of labor. That is, it will set the marginal product of labor equal to wage times the markup:

$$PF_L = \mu W. \tag{2}$$

Similarly, we can get a condition equating the marginal product of capital and the rental rate, R. Note that these conditions do not arise from assumptions of what form of competition generates the markup, e.g. if the firm is a monopoly or an oligopoly, but simply from an optimality condition with minimum assumptions. From these observations, we can analyze the three leading estimation procedures:

1. The first approach estimates a production function for various firms or sectors, based on a variety of inputs. It allows for increasing returns to scale and recovers the markup by applying conditions for cost minimization. To arrive at the first estimation procedure, note that we can rewrite equation (1) as follows:

$$\frac{F_L L}{Y} = \mu \frac{WL}{PY} \tag{3}$$

The left-hand side is the elasticity of output with respect to labor input and the right-hand side is labor's share in revenue, multiplied by the markup. After deriving a very similar equation for efficient use of capital and adding them, we obtain:

$$\frac{F_L L}{Y} + \frac{F_K K}{Y} = \mu \left(1 - s_\pi\right)$$

where s_π is the ratio of profit to revenue and the left-hand side is the sum of the output

elasticities, or the degree of returns to scale. Using the equation above, one can infer the markup for each firm[3]. However, the equation also clarifies what one needs to assume (or estimate) using this method. Note that to estimate markups, one must obtain the degree of returns to scale. Another major challenge is that one must impute a required return to capital to estimate the profit rate.

The main challenge with this framework is estimating economic profits. A typical assumption made is that profits are paid only to owners of capital. Another challenge with this approach is that disaggregated stocks of capital are not usually measured at the firm level. Specifically with Compustat, we can only observe the book, and not the market, value of a firm's capital stock.

2. The second approach, used by De Loecker et al. (2020) and us, again estimates a production function, typically using firm-level data. However, in contrast to the first approach, we recover the markup from the optimization condition for a single input bundle. If we replace labor in the framework from Basu (2019) with a vector $\boldsymbol{V} = \left(V^1, \ldots, V^J\right)$ of variable inputs, then equation one can be rewriten in the same form as in De Loecker et al. (2020):

$$\mu_{it} = \frac{\partial Q(\cdot)}{\partial V_{it}} \frac{V_{it}}{Q_{it}} \times \frac{P_{it} Q_{it}}{P_{it}^V V_{it}} \tag{4}$$

where P_{it} is the price of the final good, Q_{it} is the quantity, P_{it}^V is the price of the variable input and V_{it} is the quantity of the variable input. In words, one can estimate the markup as the elasticity of output with respect to the variable input divided by the factor payment to the selected input as a share of the firm's revenue.

Note that this approach avoids the need to estimate the rate of economic profit. On the other hand, to infer markups within this framework, one must estimate the output elasticity, as we do following De Loecker et al. (2020).

3. Third, one can first use a first-order approximation in logs to the production function and then

[3]See Gutiérrez and Philippon (2017) for an application of this approach with the Compustat data

obtain an equation for markups using again cost minimization:

$$\Delta y \simeq \mu \left[\frac{WL}{PY} \Delta l + \frac{RK}{PY} \Delta k \right] + \Delta z$$

To estimate μ, one can take the unobserved Δz as the error term and estimate the above equation. Since changes in technology are probably correlated with changes in input choices, Hall (2018) uses instrumental variables to estimate markups using a small modification of this method. There are two main disadvantages with this method. First, it requires extra assumptions that come with the estimation with instrumental variables. Second, and most importantly to this paper, is that this method does not generate firm-level markups. In the equation above, for instance, we impose one aggregate markup. Hall (2018) estimates markups on the industry-level and allow the markups to be the sum of a constant and a time trend, which remediates this issue, but still would not allow us to study the heterogeneity of markups on the firm level.

What are the advantages of using the production function approach from De Loecker et al. (2020)? The answer to this question depends on the relevant question of interest and the dataset being used. Since we want to analyze a large panel of firms with a broad economic coverage, we need a method that can be applied across a range of different industries and firms. Thus, we find the Compustat data and the cost-minimizing approaches better suited than others used in the industrial organization literature, since it doesn't make any assumptions about the structure of competition in a given market.

Among the options of cost-minimization approaches, we think that the three methods mentioned above have their merits and are useful for analyzing markups with financial statements data. Nonetheless, we find that the approach from De Loecker et al. (2020) makes assumptions which are easier to accept. Note that in all the frameworks above the markup does not vary by inputs. Thus, we can select any variable input which plausibly does not receive pure profits (e.g. intermediate inputs) to deal with the issue of measuring profit rate. Following De Loecker et al. (2020), we use cost of goods sold (COGS), which contains most intermediate inputs and some labor, as the main

bundle of input of the analysis.

Despite the improvements that we perceive with this methodology, there are also several limitations. First, as highlighted by Traina (2018) and Basu (2019), financial statements are not constructed to measure variable and fixed costs. Moreover, the definition of COGS and SG&A is not as clear as one hoped for all industries over time. Thus, we would benefit from datasets that measure each input comprehensively over time and across all firms, and not bundles of inputs such as COGS and SG&A, for which the definitions might vary. Second, as stressed by Syverson (2019) and Basu (2019), studies such as Hall (2018) using aggregate data have found different patterns in the aggregate markups. Moreover, Syverson (2019) also argues that the rise in markups estimated in De Loecker et al. (2020) should have several macro economic implications which seem inconsistent with the data. However, as argued by De Loecker et al. (2020), given the great heterogeneity in markups across firms and the increasing correlation between firms' sales and markup, it is not a surprise that we see a difference between the economy-wide averages and the sales-weighted markup aggregated from micro data. Third, as highlighted by Bond et al. (2021), in Compustat we observe sales instead of output quantities, which means we effectively estimate the revenue elasticity instead of the output elasticity. They show that this can be problematic whenever the two differ. We share this concern with Bond et al. (2021), but we are not aware of economy-wide datasets with prices or quantities.

F 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

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

$$, \qquad \varepsilon_{i,j,t} \equiv \tfrac{\psi_1}{1-\psi_2} u_{2,t}^{ij} + \tfrac{\psi_2}{1-\psi_2} v_{2,t}^{ij}$$

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

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

G 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

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

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

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,

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

where $\Delta \hat{y}_{i,t+1}$ 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 instance 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 operator. Notice that this simply assumes that firms are very well aware of their own environnement 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, even though 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

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

References

Basu, Susanto, "Are price-cost markups rising in the United States? A discussion of the evidence," *Journal of Economic Perspectives*, 2019, *33* (3), 3–22.

Berry, Steven, James Levinsohn, and Ariel Pakes, "Automobile Prices in Market Equilibrium," *Econometrica*, 1995, *63* (4), 841–890.

Bond, Steve, Arshia Hashemi, Greg Kaplan, and Piotr Zoch, "Some unpleasant markup arithmetic: Production function elasticities and their estimation from production data," *Journal of Monetary Economics*, 2021, *121*, 1–14.

Christiano, Lawrence J, Martin Eichenbaum, and Charles L Evans, "Nominal rigidities and the dynamic effects of a shock to monetary policy," *Journal of political Economy*, 2005, *113* (1), 1–45.

Gutiérrez, Germán and Thomas Philippon, "Declining Competition and Investment in the U.S.," *NBER Working Paper Series*, 2017, (23583).

Hall, Robert E., "New Evidence on the Markup of Prices over Marginal Costs and the Role of Mega-Firms in the US Economy," *NBER Working Paper Series*, 2018, p. 21.

Karabarbounis, Loukas and Brent Neiman, "The Global Decline of the Labor Share," *The Quarterly Journal of Economics*, 2014, *129* (1), 61–103.

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

Nekarda, Christopher J and Valerie A Ramey, "The cyclical behavior of the price-cost markup," *Journal of Money, Credit and Banking*, 2020, *52* (S2), 319–353.

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

Traina, James, "Is Aggregate Market Power Increasing? Production Trends Using Financial Statements," 2018. Manuscript.