# Pricing Inequality

Simon Mongey
Federal Reserve Bank of Minneapolis and NBER

Michael E. Waugh
Federal Reserve Bank of Minneapolis and NBER

January 2025

ABSTRACT

This paper studies household inequality and product market power in dynamic, general equilibrium. In our model, households' price elasticities of demand endogenously vary with wealth. Heterogeneous firms set their price as oligopolistic competitors given the endogenous distribution of demand. A firm's market power varies with the distribution of demand as households with different elasticities sort into high- and low-price varieties. Under standard preferences, larger firms' products are more appealing, sell at higher prices, to more households, and a relatively richer customer base, face less elastic demand, and set higher markups. Quantitatively (a) our model rationalizes a wide set of recent empirical studies in the cross-section of households and firms, (b) we find household heterogeneity to be a dominant source of markup variation across firms, and (c) a one-time fiscal transfer of one percent of GDP to households leads to a 0.3 percentage point increase in the aggregate markup.

Email: michael.e.waugh@gmail.com, simonmongey@gmail.com. The views expressed herein are those of the authors and not necessarily those of the Federal Reserve Bank of Minneapolis or the Federal Reserve System.

## 1. Introduction

Customers and their composition matter for understanding a firm's size and pricing strategies. Firm and customer-level evidence has emphasized that a firm's number of customers, or the extensive margin of demand, is the dominant determinant of sales (Argente, Fitzgerald, Moreira, and Priolo, 2021; Einav, Klenow, Levin, and Murciano-Goroff, 2021; Afrouzi, Drenik, and Kim, 2023). Further evidence suggests that the primary reason some firms sell more than others is product quality or appeal (Hottman, Redding, and Weinstein, 2016; Eslava, Haltiwanger, and Urdaneta, 2024).

Behind a firm's customer base are households which are heterogeneous in systematic ways. Bils and Klenow (2001), Handbury (2021), Jaimovich, Rebelo, Wong, and Zhang (2019) find that high-income households tend to purchase higher-priced varieties. This sorting is interpreted as households making choices based on quality. Auer et al. (2024) emphasize that households' price sensitivity differs, with high-income households being less price-sensitive than low-income households. Faber and Fally (2022) show that richer households tend to purchase from larger firms. Taken together, these studies suggest that larger firms offer higher-quality products and sell to larger, relatively richer, and less price-sensitive customer bases. An implication of this is that larger firms may have higher markups because they face households with relatively inelastic demand curves, rather than due to size-based market power considerations

When asked, general equilibrium models of heterogeneous firms and imperfect competition provide narrow, limited answers to questions about why firms are big or why they have larger markups (e.g. Atkeson and Burstein, 2008; Edmond et al., 2023). Large firms are large because they generate more sales per customer, rather than having more customers. These models are agnostic as to whether size is driven by quality or productivity variation. Markups arise solely from market power considerations and depend on a firm's position within the market.

This paper develops a dynamic, general equilibrium model of household inequality and product market power that is consistent with the facts discussed above: larger firms sell to more households, their products are more appealing, at higher prices, to a relatively richer customer base, endogenously face less elastic demand, and set higher markups. We show how this model rationalizes a wide range of un-targeted empirical studies in the cross-section of households and firms. We then quantify how a fiscal transfer to households leads to an increase in the aggregate markup, with underlying differential price and output responses of firms.

Two canonical models form the basis of the demand side. Heterogeneity is introduced via the standard incomplete markets model (Bewley, 1979; Imrohoroğlu, 1989; Huggett, 1993; Aiyagari, 1994), where households face incomplete insurance against idiosyncratic shocks. Incomplete markets models are powerful tools for rationalizing empirical evidence on heterogeneity in household behavior observed in microdata — large, positive marginal propensities to consume

and the lack of consumption smoothing (Kaplan and Violante, 2010). These models are also useful for policy analysis, as they allow for understanding both the aggregate and distributional consequences of fiscal and monetary policy.

We combine the incomplete markets model with a nested-logit demand system, where households make a discrete choice each period over different goods and different varieties of each good (McFadden, 1974). Discrete choice parsimoniously delivers an extensive margin of demand. We show that when combined with the incomplete markets model, discrete choice also endogenizes heterogeneous elasticities of demand, and differential sorting of households across varieties. Income or asset poor households strongly value the resources left over after a purchase, making them more price-sensitive when selecting a product, which leads to sorting into relatively cheaper varieties. Unlike static, partial equilibrium applications of the nested-logit demand system, changes in income or wealth, policy interventions, and other economic fundamentals differentially affect the marginal value of wealth, altering households' price sensitivity, and firms' customer bases.

The supply side of our model features heterogeneous firms, differing in both productivity and quality, which compete strategically based on the distribution of demand they face and the prices set by competitors. In our model, markups arise for two reasons. First, as in Atkeson and Burstein (2008), a large firm will want to markup its product given that is large and has market power. Second, and unlike canonical macro approaches, the composition of demand matters as well when setting its markup, specifically whether the firm faces elastic or inelastic households. Thus, the demand side of our model, with heterogeneous price sensitivity and sorting, interacts non-trivially with the supply side, giving rise to two plausible forces that explain why markups vary across firms.

We use our framework to revisit the facts discussed above and draw out implications for the determination of markups and how markups are influenced by fiscal transfers. Specifically, we have three sets of results. **First**, when disciplined by data on the cross-section of firms and households, our framework quantitatively accounts for a broad range of empirical facts regarding firms and households. **Second**, we show that household heterogeneity is important for pricing in the cross-section: across firms, heterogeneity in customer bases explains more markup variation than size-based market power mechanisms. **Third**, household heterogeneity is important for how prices respond to fiscal transfers: a one-time fiscal transfer of one percent of GDP to households results in a 0.3 percentage point increase in the aggregate markup.

**First**, we calibrate our model to match salient facts about household heterogeneity in price sensitivity, sorting on quality, and markup variation. A key feature of our model is its parsimony, as it relies on a small and relatively standard set of parameters. The parameters are chosen so that the model replicates three key empirical relationships on both the household and firm side.

Specifically, we (i) replicate the empirical exercise in Auer et al. (2024) which finds a negative relationship between household income and demand elasticities; (ii) replicate the exercise in Jaimovich et al. (2019), which finds that households in the top quintile of expenditure buy more expensive varieties; and (iii) match the cross-sectional relationship between markups and sales shares as found in Edmond et al. (2023). Thus, in a parsimonious way, we match key facts about how elasticities vary across households and markups vary across firms.

Our calibration reveals (a) the importance of product quality in determining firm size, and (b) a positive correlation between quality and marginal cost, which is captured by decreasing returns to scale. These outcomes are consistent with empirical findings emphasizing that the dominant source of variation in firm sales is product quality and appeal, as highlighted in Hottman et al. (2016) and Eslava et al. (2024). Alternative approaches, in which productivity determines firm size, imply counterfactual patterns of household sorting across firms and how markups vary with firm size.

We then show how our model delivers a unified interpretation of the facts discussed above about how firms and households interact in product markets. In addition to highlighting the importance of quality, our model is quantitatively consistent with the fact that the extensive margin of demand is the dominant source of firm sales as shown in Afrouzi et al. (2023). Our model is also quantitatively consistent with the sorting patterns of households to firms, where richer households tend to purchase from larger firms as in Faber and Fally (2022). We also show how our model is quantitatively consistent with the evidence from Stroebel and Vavra (2019) on markup responses to wealth-induced demand shocks. This evidence is particularly relevant because it directly supports a core mechanism in our model, where the marginal value of wealth shapes both demand elasticities and markups.[1]

**Second**, we show that household heterogeneity is quantitatively important in accounting for the cross-section of firm markups. An outcome of our calibration—consistent with the facts discussed above—is that larger firms tend to sell to higher-income customers who are less price-sensitive. As a result, large firms have larger markups for two reasons (i) size-based market power forces and (ii) the composition of their customer base. We use the calibrated model to decompose markup variation in the cross-section of firms into a component due to size-based market power forces and a component due to household heterogeneity. We find that the household heterogeneity component accounts for 58 percent of the difference across firms. This result is striking in that canonical macro approaches assign *all* markup variation to size-based market power. Prominent examples include Atkeson and Burstein (2008), Edmond et al. (2023), Loecker et al. (2021), Baqaee et al. (2024a), Baqaee et al. (2024b), Boar and Midrigan (2019).

[1]Similar evidence is found by Gupta (2024) in India. Following a shock that increases income of low income households, markups of firms that sell to lower income households increase, while those that sell to high income households do not.

**Third**, we show that household heterogeneity is important for understanding the dynamics of markups in response to an aggregate shock. We analyze a one-time fiscal transfer—such as the lump-sum checks issued during the Covid-19 pandemic—equal to one percent of GDP. This produces a modest but meaningful increase in the aggregate markup of 0.3 percentage points, with significant heterogeneity across firms based on their market position.[2] Household heterogeneity is at the essence of this result — a fiscal transfer lowers the marginal value of wealth, households become less price sensitive, and firms increase their markups accordingly. When we turn off heterogeneous price sensitivity, a fiscal transfer has no impact on the shape of firms' demand curves, leaving markups unchanged.

**Literature.** We present a clear deviation from recent approaches that incorporate household heterogeneity, sorting of households across price and quality dimensions of goods, and heterogeneity in households' elasticities of demand for goods. Relative to the approaches discussed below, we emphasize that our theory combines workhorse models and leverages the economic interaction between them, allowing for a richer data generating process with less parameters.

First, recent research across trade, microeconomics, macroeconomics, and urban economics has used functional forms for preference aggregators over products to explain empirical patterns of sorting and heterogeneity in demand elasticities across households, e.g., non-homothetic preferences used in Handbury (2021), Auer et al. (2024), Faber and Fally (2022) and Olivi et al. (2024). In these models, preferences—and by extension, demand elasticities—are typically assumed to be exogenous functions of income. In contrast, our model endogenizes the distribution of demand elasticities, linking them to market incompleteness and households' consumption and savings decisions. Section 6 illustrates this point by showing how fiscal transfers impact households' demand elasticities and firms' pricing. More broadly, our model offers an equilibrium framework to study how shocks to demand, supply, financial constraints, interest rates and so on, affect demand elasticities and pricing.

Second, a recent literature in macroeconomics has studied household inequality and firm price setting in search frameworks following Burdett and Judd (1983). In Kaplan and Menzio (2016) unemployed workers search more intensively than employed workers, leading to lower markups during periods of high unemployment. Pytka (2018), Nord (2023) and Sangani (2023) endogenize search intensity and embed this model of demand in an incomplete markets setting. In each case *demand elasticities* depend on search intensity, while the cost of search is parameterized to vary with income. Similarly, *sorting* depends on preference heterogeneity across consumers, which is also parameterized to vary with income. By contrast, our framework assumes identical preferences and budget sets for all households, and generates heterogeneity in demand elastic-

[2]We view these results as suggestive about mechanisms behind how prices behaved during the Covid-19 pandemic in which fiscal transfers amounted to more than seven percent of GDP. However, a complete account of this episode is beyond the scope of the paper as our model does not include nominal rigidities nor supply shocks.

ities and sorting by leveraging the interaction between (i) heterogeneity in the marginal value of wealth via a standard incomplete markets model, and (ii) a standard discrete choice framework.[3] Unlike Burdett and Judd (1983) models, our framework also allows for the analysis of concentrated markets, thereby incorporating market-share-based explanations for markups.

Third, a long literature in industrial organization includes individual heterogeneity in price sensitivity in estimated empirical models of demand—most famously Berry et al. (1995). Two recent studies in industrial organization estimate such demand models across a wide range of goods and markets over time (Atalay et al., 2023; Döpper et al., 2024). Independently, they find that declining price sensitivity over time drives the path of average markups in the economy. Our contributions are twofold: (a) demonstrating that a standard incomplete markets economy provides a theoretical foundation for this heterogeneity, and (b) making price-sensitivity endogenous to individual, aggregate and policy shocks, rather than treating it as an estimated parameter. When compared to the estimated random-coefficients model of Nakamura and Zerom (2010), we find that our model produces a similar estimate of the relationship between income and price elasticities of demand.

Our paper emphasizes that incomplete markets play a central role in shaping firm pricing. Incomplete markets models are a powerful framework for explaining empirical evidence on heterogeneity in household behavior observed in microdata: large, positive MPC's and lack of consumption smoothing (Kaplan and Violante, 2010). We emphasize that inequality in the marginal value of wealth, inherent in incomplete markets models, is a key driver of heterogeneity in individual demand elasticities, which in turn influence sorting, firms' demand elasticities, and ultimately pricing. Mongey and Waugh (2024) points out that the industrial organization approach to discrete choice demand is inherently one of *incomplete markets*, lacking insurance against idiosyncratic preference shocks. Mongey and Waugh (2024) further demonstrates that in a discrete choice economy with *complete markets* the equalization of marginal utilities of consumption across agents implies little scope for heterogeneity in demand elasticities. Thus, studying discrete demand within an incomplete markets framework is both a natural extension of our theory and consistent with prevailing practice in economics.

**Overview.** Section 2 defines the model environment and equilibrium. Section 3 characterizes heterogeneity in demand elasticities and the sorting of households across products. Section 4 calibrates the model, with a focus on replicating the quasi-experiment from Auer et al. (2024), and demonstrates that the model qualitatively and quantitatively reproduces a wide range of empirical findings, including those in Stroebel and Vavra (2019). Section 5 and Section 6 then apply the model to two questions. Section 5 studies the role of household heterogeneity in ac-

[3] In its simplest form, our theory explains sorting, price elasticity, and markup heterogeneity across households in a manner consistent with recent empirical findings, while introducing no additional parameters or functional forms to either the Bewley model or a discrete choice model.

counting for markup variation in the cross-section of firms. Section 6 investigates the dynamics of markups and prices following a lump-sum fiscal transfer. The Appendix includes important extensions of our theory including (a) continuous time, (b) a "many goods purchased" version, (c) a simplified pedagogical model in which we establish all results "pencil and paper".

## 2. Model

Section 2.1 describes the economic environment, Section 2.2 describes agents' decision problems, and Section 2.3 defines general equilibrium across markets and Nash equilibrium within.

### 2.1. Environment

Time is discrete. The economy consists of three types of agents: a continuum of firms, a continuum of households, and a government. Firms are heterogeneous in the goods they produce and their productivity. Households are heterogeneous in their assets, labor productivity, and preferences for goods. Fiscal policy is passive: the government taxes labor income to fund government spending and interest payments on debt, which is held by households.

**Goods, varieties, and firms.** The economy features two types of goods, each characterized by heterogeneity in production and preferences, leading to different competitive structures. A "homogeneous good" is produced by many homogeneous firms, hence product markets are competitive. "Differentiated goods" are produced by many heterogeneous firms. Finitely many firms produce each good, and each of these firms produces a differentiated variety of that good. Producers of the different varieties within each good engage in oligopolistic competition, while they remain competitive with producers of other goods.

**Competitive good.** The homogeneous good, which we also call the *competitive good*, is produced by firms with a linear technology. We consider a representative firm, with output and profits:

$$Y_c = Z_c N_c \quad , \quad \Pi_c = P_c Y_c - W N_c \,, \tag{1}$$

where the subscript $c$ denotes that this is the competitive good. Output depends on total factor productivity $Z_c$ and efficiency units of labor $N_c$, hired in a competitive labor market at price $W$. Competition implies that price equals marginal cost:

$$P_c = \frac{W}{Z_c}. \tag{2}$$

We choose $P_c$ as the numeraire and normalize it to one; thus, all other prices are relative to the price of the competitive good. Infinitely elastic demand for labor in the competitive sector implies $W = Z_c$ in all periods.

**Differentiated goods.** There are $m \in \{1, \ldots, M\}$ goods, and we refer to the *market* for each good. Within each market there are $j \in \{1, \ldots, J_m\}$ firms, each producing a unique variety. Thus, firm and variety are synonymous, as are good and market. The total number of markets $M$, is so large that firms within a market are effectively infinitesimal with respect to all other markets. However, the number of competitors $J_m$, within a market is finite, and firms understand that their decisions impact market-level variables.

Firms $jm$ are heterogeneous in two dimensions. First, each firm $jm$ differs in their productivity, $z_{jm}$. Second, each firm $jm$'s product has quality $\psi_{jm}$ that is valued commonly by all consumers; discussed below in preferences. The literature often refers to this as product *appeal* (Hottman et al., 2016; Eslava et al., 2024).

Given this environment, a firm $jm$ operates a technology with potentially non-constant returns to scale, with output and profits:

$$y_{jm} = z_{jm} n_{jm}^{\alpha} \quad , \quad \pi_{jm} = p_{jm} y_{jm} - W n_{jm} \, , \tag{3}$$

where $z_{jm}$ is the firm-specific productivity term, and $\alpha$ controls the returns to scale.

**Households.** A unit mass of households is indexed by $i$. Each period, $t$, household $i$ consumes a continuous amount of the competitive good $(c_t^i)$ and chooses a single good-variety $jm$ and a continuous amount of that good to purchase $(x_{jmt}^i)$.[4]

Preferences over sequences of consumption of both goods are given by:

$$\mathbb{E}\left[\sum_{t=0}^{\infty} \beta^t \sum_{m \in M} \sum_{j \in J_m} \widetilde{u}_{jmt}^i\right] \quad , \quad \widetilde{u}_{jmt}^i = \begin{cases} u\left(x_{jmt}^i, c_t^i\right) + \psi_{jm} + \zeta_{jmt}^i & \text{, if purchase } jm \\ 0 & \text{, otherwise.} \end{cases} \tag{4}$$

Households' period utility function is of the additive random utility class. The term $u\left(x_{jmt}^i, c_t^i\right)$ describes the mapping from consumption of continuous quantities of both goods into utils. The term $\psi_{jm} + \zeta_{jmt}^i$ provides utils from choosing differentiated good $jm$. The first component $\zeta_{jmt}^i$, is random *iid* across time and households. The second component $\psi_{jm}$, reflects permanent heterogeneity in quality differences across varieties, as discussed above, and is common to all consumers.

We assume that the vector of idiosyncratic tastes of individual $i$ at date $t$ is $iid$ over time and individuals, and distributed according to a Generalized Extreme Value distribution with pa-

[4] An intensive margin of demand for the differentiated good is not necessary for our results. We note below what happens in the special case of unit demand.

rameters $\eta$ and $\theta$:

$$F\left(\boldsymbol{\zeta}_t^i\right) = \prod_{m \in M} \exp\left\{ -\left( \sum_{jm \in J_m} e^{-\eta \zeta_{jmt}^i} \right)^{\theta/\eta} \right\}. \tag{5}$$

This is commonly referred to as a *nested-logit* with an outer nest of goods $m$ and inner nest of varieties $jm$. This approach allows us to mimic properties of nested CES (Verboven, 1996) and in turn oligopolistic pricing behavior similar to Atkeson and Burstein (2008). Parameters $\theta$ and $\eta$ control dispersion in tastes across and within markets with larger values implying less dispersed tastes. These parameters also shape patterns of substitution similar to nested CES with the parameters $\theta$ and $\eta$ controlling within- and between-market elasticities of substitution. To be well behaved, we assume $\eta \geq \theta$: varieties within a market are more substitutable (less dispersion in tastes) than across markets (more dispersion in tastes across markets). If $\eta = \theta$, varieties within and across markets are equally substitutable, and $F$ collapses to a simple Type 1 Extreme Value distribution. Later, we separately estimate this restricted version of the model. Dynamically, since shocks are $iid$ and there are a large number of markets, an individual will purchase from different markets and producers in each period.

A household's labor productivity is stochastic and evolves according to a Markov process. Let $e_{it}$ be a household's efficiency units and $\mathcal{P}(e_{it}, e')$ describe the transition density to $e'$. We assume that $\mathcal{P}$ is well behaved in the necessary ways and time-invariant.

Households pay labor income taxes to the Government, and their net labor income is $(1 - \tau_t)W_t e_{it}$. Households also receive $T_t$ in transfers. Households can trade a non-state contingent asset which pays $a_{t+1}^i$ tomorrow with return $R_{t+1}$. An exogenous debt limit $\underline{a}$ constrains borrowing such that $a_{t+1}^i \geq \underline{a}$. Finally, households are the owners of firms and receive equal shares of aggregate profits $\Pi_t$.

Household $i$'s budget constraint is as follows. Conditional on choosing variety $jm$:

$$c_t^i + x_{jmt}^i p_{jmt} + a_{t+1}^i \;\leq\; R_{t+1} a_t^i + (1 - \tau_t) W_t e_t^i + T_t + \Pi_t \,. \tag{6}$$

Expenditure is on the competitive good, the differentiated good, and asset purchases. Resources available are asset payments, net labor income, transfers and profits.

**Government.** The Government issues one-period, non-state contingent, interest-bearing debt, which is held by households. Labor income taxes finance expenditures of $G_t$ units of the competitive good, interest payments on outstanding debt, and transfers $T_t$ to households. Government spending is not valued by households. In all exercises we will keep $G_t$ constant at $G$. The

Government budget constraint is:

$$G + R_t B_t + T_t = \tau_t W_t N_t + B_{t+1}. \tag{7}$$

One role of the government is to tax households and supply safe assets to them. In this sense, the government is providing an elastic supply curve of assets in response to the elastic demand for assets from the households.

### 2.2. Decision problems

This section describes the problems faced by firms and households, given the economic environment described above.

**Differentiated goods firms' problem.** A key assumption is what the game entails and what the firm perceives it can influence. Throughout we make an assumption of price competition (i.e. Bertrand), although there is no methodological constraint on solving the model under Cournot competition.

> **Assumption 1 (Bertrand)** *Firms play a static game of price competition. Specifically, each firm $jm$ chooses its price, taking as given the prices of its competitors in the market $\boldsymbol{p}_{-jm}$, as well as the aggregates that we gather in $\boldsymbol{S} = \{W, R, P_c\}$. Each firm in market $m$ recognizes that market $m$ quantities and prices vary when that firm changes its price.*

Given Assumption 1, the firm chooses its price and labor inputs to maximize profits $\pi_{jm}$, given its perceived demand curve $x_{jm}(p_{jm};\ \boldsymbol{p}_{-jm}, \boldsymbol{S}\ )$:

$$\pi_{jm} = \max_{p_{jm}, n_{jm}} \ p_{jm} x_{jm}\big(p_{jm};\ \boldsymbol{p}_{-jm},\ \boldsymbol{S}\ \big) - W n_{jm}, \quad \text{subject to} \quad x_{jm}\big(p_{jm};\ \boldsymbol{p}_{-jm},\ \boldsymbol{S}\ \big) = z_{jm} n_{jm}^{\alpha}.$$

Firm $jm$'s perceived demand curve aggregates over all households' demands and is a function of its price $p_{jm}$, aggregates $\boldsymbol{S}$ and the vector of prices of its competitors in market $m$, denoted $\boldsymbol{p}_{-jm}$. All firms take their competitors' prices $\boldsymbol{p}_{-jm}$ as given, and choose their best response $p_{jm}$.

The firm's profit maximizing price is a markup over marginal cost:

$$p_{jm} = \mu_{jm} \times mc_{jm} \quad , \quad \mu_{jm} = \frac{\varepsilon_{jm}(p_{jm};\ \boldsymbol{p}_{-jm},\ \boldsymbol{S})}{\varepsilon_{jm}(p_{jm};\ \boldsymbol{p}_{-jm},\ \boldsymbol{S}) - 1} \quad , \quad mc_{jm} = \frac{1}{\alpha}\left(\frac{W}{z_{jm}}\right) n_{jm}^{1-\alpha} \quad , \tag{8}$$

where $\varepsilon_{jm}(p_{jm};\ \boldsymbol{p}_{-jm}, \boldsymbol{S})$ is the elasticity of demand for firm $jm$'s product, defined as

$$\varepsilon_{jm}(p_{jm};\ \boldsymbol{p}_{-jm},\ \boldsymbol{S}) := -\frac{\partial x_{jm}(p_{jm};\ \boldsymbol{p}_{-jm},\ \boldsymbol{S}\ )/x_{jm}(p_{jm};\ \boldsymbol{p}_{-jm},\ \boldsymbol{S}\ )}{\partial p_{jm}/p_{jm}}\bigg|_{\boldsymbol{p}_{-jm}}. \tag{9}$$

**Household problem.** The state variables of a household are its asset holdings and efficiency units. For brevity we detail a stationary economy. When we later study transition dynamics following an unforeseen shock, we will describe how we extend the model.

Let $v(a, e, \boldsymbol{\zeta})$ be the expected present discounted value of lifetime utility of a household with assets $a$, productivity $e$ and vector of idiosyncratic tastes $\boldsymbol{\zeta}$. The value is given by:

$$v\Big(a, e, \boldsymbol{\zeta}\Big) = \max_{jm} \Big\{ \; v_{jm}(a, e) + \psi_{jm} + \zeta_{jm} \; \Big\}. \tag{10}$$

This represents the maximum value across the choices associated with the different varieties. The value of choosing variety $jm$, net of quality $\psi_{jm}$ and idiosyncratic utility $\zeta_{jm}$, is:

$$v_{jm}(a, e) = \max_{a'} \Big\{ u\Big(x_{jm}(a, e),\; c_{jm}(a, e)\Big) + \beta\, \mathbb{E}\Big[ v\Big(a', e', \boldsymbol{\zeta}'\Big)\Big] \Big\} \tag{11}$$

$$\text{subject to:} \quad \text{borrowing constraint } a' \geq \underline{a} \quad \text{and budget constraint } (6)$$

where households choose asset holdings. We index the consumption of the competitive good $c_{jm}(a, e)$ with $jm$ as its value depends upon the choice of $jm$. The continuation value in (11) is an expectation with respect to future efficiency units $e'$ and taste shocks $\boldsymbol{\zeta}'$. The solution includes an asset policy function $a'_{jm}(a, e)$, that maps states into asset holdings tomorrow contingent on choosing $jm$.

**Choice probabilities.** The distribution of taste shocks implies the following choice probabilities for each differentiated good. Conditional on purchasing good $m$, the probability that variety $jm$ is chosen is

$$\rho_{jm|m}(a, e) = \left( \frac{\exp\left\{ v_{jm}(a, e) + \psi_{jm} \right\}}{\exp\left\{ \widetilde{v}_m(a, e) \right\}} \right)^{\eta}, \tag{12}$$

$$\text{where} \quad \exp\left\{ \widetilde{v}_m(a, e) \right\} := \left[ \sum_{j'm \in J_m} \exp\left\{ v_{j'm}(a, e) + \psi_{j'm} \right\}^{\eta} \right]^{1/\eta}. \tag{13}$$

The term $\widetilde{v}_m(a, e)$ represents the value that—prior to drawing preference shocks—an individual would expect from consuming the value maximizing choice in market $m$. This expected value term plays a similar role to a CES price-index in summarizing the value of all options within market $m$. The probability that market $m$ is chosen is

$$\rho_m(a, e) = \left( \frac{\exp\left\{ \widetilde{v}_m(a, e) \right\}}{\exp\left\{ \overline{v}(a, e) \right\}} \right)^{\theta}, \quad \text{where} \quad \exp\left\{ \overline{v}(a, e) \right\} := \quad \left[ \sum_{m' \in M} \exp\left\{ \widetilde{v}_{m'}(a, e) \right\}^{\theta} \right]^{1/\theta}. \tag{14}$$

The term $\overline{v}(a, e)$ represents the expected value across all markets. Combining these, the total

probability that good $jm$ is chosen depends on the value of buying from $jm$ relative to the average value of the market, and the average value of the market relative to all other markets:

$$\rho_{jm}(a,e) = \rho_{jm|m}(a,e)\rho_m(a,e) = \left(\frac{\exp\left\{v_{jm}(a,e)+\psi_{jm}\right\}}{\exp\left\{\widetilde{v}_m(a,e)\right\}}\right)^{\eta}\left(\frac{\exp\left\{\widetilde{v}_m(a,e)\right\}}{\exp\left\{\overline{v}(a,e)\right\}}\right)^{\theta}. \tag{15}$$

If $\eta$ is large, dispersion in idiosyncratic tastes within markets is small, so within-market choices respond strongly to small differences in values within market $m$. Similarly, if $\theta$ is large, across-market choices respond strongly to small differences in values across markets.

**Euler equation.** An Euler Equation holds for households away from the borrowing constraint:

$$u_c\Big(x_{jm}(a,e),c_{jm}(a,e)\Big) = \mathbb{E}_{e'}\left[\sum_{m\in M}\sum_{jm'\in J_m}\Big\langle\beta R\,\rho_{jm}(a',e')\Big\rangle u_c\Big(x_{jm}(a',e'),c_{jm}(a',e')\Big)\right]. \tag{16}$$

In our stationary environment $\mathbb{E}_{e'}$ represents the expectation over the household's idiosyncratic labor productivity shock. As is standard, the household equates marginal utility today with the discounted expected marginal utility tomorrow. The only modification is that the discount factor $\langle\cdot\rangle$ reflects choice probabilities over varieties tomorrow, which weight marginal utilities conditional on consuming each potential variety.

### 2.3. Aggregation and equilibrium

Given the solutions to the firm and household problem, we now define aggregate variables and provide a formal definition of equilibrium.

**Aggregation.** Determining firm demand requires aggregation, which requires a distribution $\Lambda(a,e)$ of households across the individual states. Here, the mass of households with $(a,e)$ evolves according to:

$$\Lambda(a',e') = \int_e \int_{a:a'=a'_{jm}(a,e)} \sum_{m\in M}\sum_{jm\in J_m} \rho_{jm}(a,e)\Lambda(a,e)\mathcal{P}(e,e')\,da\,de. \tag{17}$$

A mass $\rho_{jm}(a,e)\Lambda(a,e)$ of households choose variety $jm$. Of this group, fraction $\mathcal{P}(e,\mathcal{E})$ transitions to $e'\in\mathcal{E}$, which is only integrated over situations where $a'_{jm}(a,e)=a'$ capturing transitions to asset holdings. Given the distribution $\Lambda(a,e)$, all other aggregates follow.

Aggregate demand of variety $jm$ is

$$x_{jm} = \int_e\int_a \rho_{jm}(a,e)x_{jm}(a,e)\Lambda(a,e)\,da\,de. \tag{18}$$

This is the demand curve firm $jm$ faces for its variety. Given demand, aggregate profits are

obtained by summing differentiated goods producers:

$$\Pi = \sum_{m \in M} \sum_{jm \in J_m} \Big( p_{jm} x_{jm} - W n_{jm} \Big). \tag{19}$$

Aggregate private assets are calculated by integrating over asset choices, conditional on the differentiated good variety and weighted by choice probabilities $\rho_{jm}(a, e)$:

$$A' = \int_e \int_a \sum_{m \in M} \sum_{jm \in J_m} a'_{jm}(a, e) \rho_{jm}(a, e) \Lambda(a, e) da \, de. \tag{20}$$

Government tax revenues are given by aggregate payments to labor times the tax rate:

$$\tau W N = \int_e \int_a \tau W e \Lambda(a, e) da \, de. \tag{21}$$

**Simplifying assumption.** From here on we assume that households do not value the competitive good in utility: $u(x^i_t, c^i_{jmt}) = u(x^i_{jmt})$. The competitive good remains the numeraire in the economy and is used for savings, taxes, wage payments, and government spending.

**Equilibrium.** In a Stationary Recursive Competitive Equilibrium, aggregate variables, firm variables, and prices remain constant: $\tau$, $\Lambda$, $T$, $\overline{G}$, $B$, $W$, $Q$, $p_{jm}$. Private market participants take prices as given and solve their problems, the distribution of households is stationary, prices are consistent with market clearing, and the Government respects its budget constraint. Importantly, a consistency condition requires that a firm's perceived demand curve is consistent with the demand curve induced by household behavior. And given the finiteness of firms within each market, prices constitute a Nash equilibrium in each market.

**Definition 1 (Stationary Recursive Competitive Equilibrium)** *A Stationary Recursive Competitive Equilibrium is a government policy* $\{G, B, \tau, T\}$*, household value functions, asset policy functions, consumption functions, and variety choice probabilities* $\{v_{jm}(a, e), a'_{jm}(a, e), x_{jm}(a, e), c_{jm}(a, e), \rho_{jm}(a, e)\}$*, demand functions* $x_{jm}(p_{jm}, \boldsymbol{p}_{-jm})$*, probability distribution* $\Lambda(a, e)$*, and prices* $\{W, R\}$ *and* $p_{jm}$ *such that*

*i. Competitive firm prices satisfy the condition (2) given $W$;*

*ii. Differentiated goods firm prices $p_{jm}$ satisfy (8) and (9) given $W$ given their demand curve $x_{jm}(p_{jm}, \boldsymbol{p}_{-jm})$;*

*iii. The value functions, policy functions, and choice probabilities solve the household's optimization problem in (10) and (11);*

*iv. The probability distribution $\Lambda(a, e)$ induced by the policy functions, choice probabilities, and primitives is stationary and satisfies (17);*

*v. The aggregate demand functions that firms take as given are consistent with household choice probabilities and the distribution of types satisfying (18);*

*vi. Goods markets clear for the competitive and differentiated goods;*

*vii. Government budget constraint holds (7);*

*viii. Bond market clears: $A' = B'$.*

In summary, firms have some market power, and the demand curve they face, as described in (18), reflects the heterogeneity among households. This heterogeneity arises from differences in productivity, taste shocks, and the households' limited ability to fully insure against these shocks. In a stationary equilibrium, firms' strategies and household decisions are optimal, consistent with each other, and market clearing. Below, we work through key properties of the demand curve and their connections to the household side of the model.

## 3. Extensive Margin of Demand, Elasticities, and Sorting

This section emphasizes three issues: (i) that demand arises from both the intensive and extensive margins of customers, (ii) how the marginal utility of wealth determines household elasticities of demand, (iii) how the same forces in (ii) result in sorting of consumers across firms by income and assets. Together, these issues determine the firm's elasticity of demand and, consequently, its markup.

**Notation.** To ease the notation, we suppress $p_{jm}$ as an argument, for example writing $\varepsilon_{jm}$ instead of $\varepsilon(p_{jm}, \boldsymbol{p}_{-jm}, \boldsymbol{S})$. We also suppress $(a, e)$ as arguments and use superscript $i$ to capture household heterogeneity, for example writing $\rho^i_{jm}$ instead of $\rho_{jm}(a, e)$.

### 3.1. Demand and the extensive margin

Using our simplified notation, demand of variety $jm$ is

$$x_{jm} = \int \rho^i_{jm} x^i_{jm} \Lambda^i \, di. \tag{22}$$

We highlight three issues. First, a firm's demand curve depends on the number of customers it faces—the extensive margin. This is captured through choice probabilities $\rho^i_{jm}$ and the measure of consumers of type $i$, $\Lambda^i$. Thus, our model can confront evidence on the extensive margin being the main determinant of firms' sales as documented in Argente et al. (2021), Afrouzi et al. (2023), and Einav et al. (2021).

The extensive margin of demand distinguishes our model from many benchmark models of heterogeneous firms and imperfect competition, where only the intensive margin is relevant. For example, Atkeson and Burstein (2008) and Edmond et al. (2023) with representative agents,

heterogeneous agents with homothetic demand as in Boar and Midrigan (2019), or heterogeneous agents with non-homothetic demand as in Comin et al. (2021), Faber and Fally (2022), and Auer et al. (2024).

Second, we allow for an intensive margin to partially determine demand via $x^i_{jm}$. This margin distinguishes our model from the unit demand assumption in Fajgelbaum et al. (2011) and widely used in industrial organization literature, e.g., Berry et al. (1995) or Nevo (2000). Active extensive and intensive margins enable tests of our model against empirical evidence on the relative importance of each.

Third, the distribution, $\Lambda^i$, over which demand is aggregated, is endogenous. Through the law of motion in (17), household behavior determines the distribution of wealth, which in turn determines aggregate demand. Cross-equation restrictions are imposed between aggregate demand and individual demands via the distribution. In other words, the distribution is not a free parameter, and changes in response to shocks to policy and economic fundamentals.

### 3.2. Elasticities of demand

The firm's own price elasticity, $\varepsilon_{jm}$, is central in connecting how a firm sets its price in (8) and (9) with household behavior. We proceed in the following steps. First is an identity decomposing a firm's own price elasticity:

$$\varepsilon_{jm} = \int \varepsilon^i_{jm}\, \omega^i_{jm}\, di \quad , \quad \underbrace{\omega^i_{jm} := \frac{\rho^i_{jm} x^i_{jm} \Lambda^i}{x_{jm}}}_{\text{Share of } jm\text{'s sales to type } (a^i, e^i)} \quad , \tag{23}$$

$$\varepsilon^i_{jm} := \underbrace{-\frac{\partial \rho^i_{jm}/\rho^i_{jm}}{\partial p_{jm}/p_{jm}}}_{\text{Extensive margin elasticity: } \varepsilon^{\rho,i}_{jm}} + \underbrace{-\frac{\partial x^i_{jm}/x^i_{jm}}{\partial p_{jm}/p_{jm}}}_{\text{Intensive margin elasticity: } \varepsilon^{x,i}_{jm}} . \tag{24}$$

The firm's elasticity of demand is the sales-share-weighted ($\omega^i_{jm}$) average of household demand elasticities $\varepsilon^i_{jm}$. The household demand elasticity has an extensive margin component that reflects how customers reallocate spending across firms ($\varepsilon^{\rho,i}_{jm}$), and an intensive margin component that reflects how the quantity demanded changes conditional on purchasing from $jm$ ($\varepsilon^{x,i}_{jm}$). This formulation is quite general. Any model of the household can be inserted into (23) to link household-level elasticities and sales shares with aggregate demand elasticities.[5] The second step imposes our formulation of the household problem from Section 2.1.

[5] As two benchmarks, consider the following. Under unit demand, the intensive-margin elasticity is zero. Under CES, non-homothetic CES, or Kimball preferences with a continuum of goods the extensive-margin elasticity is zero.

**Extensive margin elasticity.** Household $i$'s extensive margin demand elasticity has two pieces:

$$\varepsilon^{\rho,i}_{jm} = \left[\frac{\partial \rho^i_{jm}/\rho^i_{jm}}{\partial v^i_{jm}}\right] \times \left[-\frac{\partial v^i_{jm}}{\partial p_{jm}/p_{jm}}\right] = \underbrace{\left[\eta\Big(1-\rho^i_{jm|m}\Big) + \theta\rho^i_{jm|m}\right]}_{\text{Oligopoly}} \times \underbrace{\left[-\frac{\partial v^i_{jm}}{\partial p_{jm}/p_{jm}}\right]}_{\text{Wealth}}. \tag{25}$$

The *oligopoly component* reflects the firm's share of market $m$ *for consumers of type* $(a, e)$. This term is multiplied by what we call the *wealth component* which captures the household's marginal value of the change in price, conditional on buying from $jm$.

The *oligopoly component* resembles the form found in the Bertrand competition version of Atkeson and Burstein (2008), with a key difference. As in Atkeson and Burstein (2008), if the conditional choice probability $\rho^i_{jm|m}$ is large, then firm $jm$ influences the value of market $m$ to the consumer. Consequently, the across-market dispersion in tastes $\theta$ primarily determines the elasticity of demand. In contrast, if $\rho^i_{jm|m}$ is small, the firm has minimal influence on outcomes across markets. In this case, the within-market dispersion in tastes, $\eta$, determines the elasticity of demand.

The key distinction relative to Atkeson and Burstein (2008) is that these statements are in terms of the *market for individual* $i$. Instead of overall market share, the individual elasticity is shaped by the firm's share in the consumption basket of individual $i$, $\rho^i_{jm|m}$. A firm might hold a small overall market share yet dominate the market share of a particular customer group, thereby facing inelastic demand from that group.[6]

The *wealth component* in (25) reflects how the household's value function changes with price, i.e., how sensitive a household is to price on the extensive margin. If a household's value function is highly sensitive to price, even small price changes can prompt the household to switch to another seller, leading to a high elasticity of demand. This derivative is captured by the household's Lagrange multiplier on its budget constraint. A price increase effectively reduces the household's available resources, with the multiplier measuring the marginal value of these resources. Using the household's optimal quantity choice, we can express its price sensitivity in terms of the marginal utility of consumption:

$$-\frac{\partial v_{jm}(a,e)}{\partial p_{jm}/p_{jm}} = \lambda^i_{jm} x^i_{jm} p_{jm} = u_x\Big(x^i_{jm}, c^i_{jm}\Big) x^i_{jm} = \left(x^i_{jm}\right)^{-(\sigma-1)}, \quad \text{with } u \text{ being separable CRRA.} \tag{26}$$

Equation (26) begins generally, while the last equality specializes to separable CRRA utility

[6]For example, if canned tuna consumption is segmented by wealth, poorer households might exclusively buy a generic brand, whereas richer households might opt for a gourmet brand. Each firm has market shares of one half, but represents the entire market for each household, and hence the oligopoly term is $\theta$ for the households sold to.

which we use in our quantitative applications. CRRA utility has the convenient feature that one parameter, $\sigma$, controls how price sensitivity increases or decreases with the level of consumption (i.e. if a household is rich or poor). Our quantitative applications use empirical estimates of declining price sensitivity by income from Auer et al. (2024) to calibrate $\sigma$.

To develop intuition, consider the case with $\sigma > 1$ where higher consumption (richer) households are less elastic. The reason richer households are less elastic follows the logic of income and substitution effects. Consider a household buying a high priced $jm$ variety, contemplating substitution to a lower priced variety $jm'$. Substitution to $jm'$ increases consumption, $x^i_{jm'} > x^i_{j,}$. But the marginal utility of this additional consumption is falling $u'(x^i_{jm'}) < u'(x^i_{jm})$. When income effects dominate, marginal utility of additional consumption falls more than the additional consumption gained, and hence the marginal value of substituting to a lower priced variety is lower. The price of the lower-priced variety would need to be *significantly lower* to induce substitution, which is exactly a *lower elasticity on the extensive margin*.

A nested case is logarithmic preferences, with $\sigma = 1$. With logarithmic preferences, marginal utility falls one for one with the level of consumption, making the *wealth component* of the elasticity independent of both consumption and price. Later we use the log case to provide a benchmark for quantifying the strength of the wealth component in determining markup differences across firms.

Two additional points are worth highlighting. First, the standard incomplete markets provides a best-in-class theory for the endogenous distribution of the marginal value of wealth $\lambda^i_{jm}$. And as we discussed above, the marginal utility of wealth shows up in firms' demand functions, and thus generates rich cross-equation restrictions between the household and firm side. Second, intertemporal effects of changes in prices — i.e. changes in future value functions — do not appear on the left hand side of equation (26) for two reasons: (a) the effect of prices on future value functions *through effects on asset choices* is zero through an envelope condition, (b) the effect of prices on future value functions *through future purchases* are zero because consumers purchase from many $M$ markets in the future, where each firm is infinitesimally small.

**Intensive margin elasticity.** The final component of a household's elasticity of demand is the intensive margin elasticity, $\varepsilon^{x,i}_{jm}$. We can say less in closed form about this elasticity, although we can establish bounds for it. If the household is not hand-to-mouth, then the first order condition for consumption of the differentiated good applies. In the case of CRRA utility:

$$\varepsilon^{x,i}_{jm} = \frac{\partial x^i_{jm}/x^i_{jm}}{\partial p_{jm}/p_{jm}} = \frac{1}{\sigma}\left(\frac{\partial \lambda^i_{jm}/\lambda^i_{jm}}{\partial p_{jm}/p_{jm}} + 1\right). \tag{27}$$

The key issue is how the household's multiplier on its budget constraint changes with price. An increase in the price tightens the budget constraint, and the magnitude determines how

much less consumption is demanded. As one limit, for a very rich household, $\lambda^i_{jm}$ remains unchanged, and this elasticity is just $1/\sigma$. If the household is hand-to-mouth, then consumption is determined by the budget constraint and the elasticity is one. Thus, if $\sigma > 1$, we know that $\varepsilon^{i,x}_{jm} \in [1/\sigma, 1]$. As with the extensive margin elasticity, poorer households are more elastic on the intensive margin.

Quantitatively, the intensive margin elasticity is small relative to the extensive margin. This aligns with evidence that the extensive margin is the key determinant of firm sales, not more sales per customer as found in Argente et al. (2021), Afrouzi et al. (2023), or Einav et al. (2021). Because the extensive margin is the most important, our analysis will focus on the extensive margin.

Finally, notice that quality $\psi_{jm}$ does not show up directly in either elasticity. Quality shifts the demand of each individual but does not affect is shape. However, (23) and (25) show how quality *indirectly* shapes $\varepsilon_{jm}$ through (i) altering the composition of buyers and (ii) affecting a firm's market share, and hence market power via the oligopoly term. We turn to composition and sorting issues next.

### 3.3. Sorting on price and wealth

A second determinant of $\varepsilon_{jm}$ in equation (23) are the sales weights, $\omega^i_{jm}$, on the individual elasticities, which capture the types of consumers a firm faces. The heterogeneity in price elasticities described above naturally induces sorting on the extensive margin, even without additional assumptions about consumer preferences for specific goods. Under the separable, CRRA assumption with $\sigma > 1$ the rich have a higher propensity to choose high-priced products while the poor choose low-price products.

To make this precise, we derive a log-supermodularity condition and connect it with the discussion above. Consider the comparison of two varieties within a market where $p_{j'm} > p_{jm}$ and two households where total cash on hand (assets plus labor income $y$) is higher at the rich household: $y^R > y^P$. By definition, choice probabilities are log-supermodular if $\rho^R_{j'm}/\rho^R_{jm} > \rho^P_{j'm}/\rho^P_{jm}$. In words, the propensity for the rich household to choose the high price good relative to the low priced good is higher than the propensity of the poor household. Taking logarithms, the choice probabilities satisfy log-supermodularity if

$$\Upsilon = \log\left(\frac{\rho^R_{j'm}}{\rho^R_{jm}}\right) - \log\left(\frac{\rho^P_{j'm}}{\rho^P_{jm}}\right) > 0. \tag{28}$$

We can express this condition entirely in terms of objects that appear in the extensive margin demand elasticity. Substituting the optimal choice probabilities (15), quality terms $\psi_{j'm}$ and $\psi_{jm}$ cancel when comparing individuals while holding varieties fixed, and market values $\widetilde{v}^R_m$

and $\widetilde{v}_m^P$ cancel when comparing varieties while holding individuals fixed. $\Upsilon$ is hence a linear difference-in-difference of values $v^i(p)$, which are functions only of prices. This allows us to express the differences as integrals:

$$\Upsilon = \eta\Big[v^R(p_{j'm}) - v^R(p_{jm})\Big] - \eta\Big[v^P(p_{j'm}) - v^P(p_{jm})\Big] = \eta \int_{\log p_{jm}}^{\log p_{j'm}} \left\langle -\frac{\partial v^P(p)}{\partial \log p} \right\rangle - \left\langle -\frac{\partial v^R(p)}{\partial \log p} \right\rangle d\log p.$$

The derivative terms are the *wealth component* discussed in (25) and (26). If at all prices, the marginal value of a dollar, and hence price sensitivity, of the poor household is more than that of the rich household, the pattern of sorting is log-supermodular in price and income. As we have shown, this is the case under $\sigma > 1$. Using the conditions in (26), we have $\partial v(p)/\partial \log p = -\lambda(p)x(p)p$. Applying the first-order condition we can express this in terms of differences in the marginal value of a dollar at poor versus rich households at fixed prices:

$$\Upsilon = \eta \int_{\log p_{jm}}^{\log p_{j'm}} \left[ \lambda^P(p)^{\frac{\sigma-1}{\sigma}} - \lambda^R(p)^{\frac{\sigma-1}{\sigma}} \right] p^{\frac{\sigma-1}{\sigma}}\, d\log p \;>\; 0. \tag{29}$$

This equation explicitly connects the incomplete markets environment to sorting: greater dispersion in the marginal value of wealth across households leads to stronger sorting.

An intuitive way of reading this condition is to observe that as one moves up the price distribution, the most elastic consumers drop off, leaving the least elastic consumers. This naturally gives rise to heterogeneity in firms' elasticities of demand by concentrating more elastic customers at low price firms. In contrast, if marginal values of wealth were equalized, no sorting would occur, and all firms would face identical consumer types. Consequently, any variation in firms' own-price elasticities would arise solely from differences in market shares rather than household heterogeneity.

### 3.4. Discussion

Heterogeneous price sensitivity and sorting occur (a) without requiring assumptions about differing preferences between poor and rich households, and (b) is endogenous to changes in policy and shocks. This is distinct from important preference specifications used in the literature.

First, some form of non-homotheticity is necessary to for household heterogeneity to matter for demand. Boar and Midrigan (2019) use a standard incomplete markets model but each household has preferences over a continuum of goods with a homothetic, non-CES, demand aggregator, as in Klenow and Willis (2016). Firms' demand elasticities do vary, but only through

market share. Since all households spend the same fraction of their expenditure on each good, the distribution of wealth and income is immaterial for the elasticity of demand that firms face. What we have emphasized is that household heterogeneity in elasticities and sorting is a key determinant of the elasticity of demand that firms face.

Our model of non-homotheticities differs from recent studies in trade and urban economics. These papers use novel preference aggregators over a continuum of products that generate empirical patterns of demand elasticity heterogeneity and sorting. Unlike our setting, these impose relationships between income and preferences for quality or between income and elasticities. This amounts to assuming that rich and poor households have different preferences. See for example, Faber and Fally (2022), Auer et al. (2024), or Handbury (2021). Our setting economizes on free parameters. We do not need to specify how preferences vary by individual characteristics, which would require additional parameters; instead leveraging curvature in marginal utilities via $\sigma$ which is a single parameter of the incomplete markets model. As we demonstrate below using standard parameter values, the model generates a rich set of correlations that align with numerous empirical regularities.

Our setting also differs from Fajgelbaum et al. (2011). In their model, there is unit demand of the differentiated good and complementarity between non-differentiated consumption and quality of the differentiated good. As discussed in Appendix E, applying their preferences to our characterization delivers the implication that individuals' elasticities of demand *increase* with income unless additional assumptions about variances of the taste shocks are introduced. Finally, in Appendix E we show how Fajgelbaum et al. (2011) implies that larger firms are low quality and low markup which is inconsistent with the facts in Hottman et al. (2016) or Edmond et al. (2023).

A closely related approach to our modeling of households is the mixed-logit demand system of Berry, Levinsohn, and Pakes (1995). Using our notation and Nevo's (2000) interpretation of Berry, Levinsohn, and Pakes (1995), individuals have quasi-linear random utility preferences, and are heterogeneous in income:

$$v\left(e^i, \boldsymbol{\zeta}^i\right) = \max_j \left\{ \alpha^i \left(e^i - p_j\right) + \psi_j + \zeta_j^i \right\} \quad , \quad \alpha^i = \alpha_0 + \alpha_1 \log e^i. \tag{30}$$

Under single nested logit, the extensive margin elasticity in this setting is $\varepsilon_j^i = \eta \alpha^i p_j$ and the intensive margin elasticity is zero. The appeal of this setting is the empirical flexibility to account for heterogeneity in price sensitivity. However, the sensitivity of households to prices is exogenous via $\alpha_0$ and $\alpha_1$, and cannot depend on shocks or changes to policy.[7] In our model,

[7] We nest the preferences in (30). Following our derivations, through elasticity and sorting conditions above, one would observe that elasticities of demand and sorting patterns are also exogenous. For example, using $v^i(p) = \alpha^i(e^i - p)$ in the sorting condition yields $\Upsilon = \alpha_1(\log e^R - \log e^P)(p_{j'm} - p_{jm})$, which is entirely exogenous.

policies that provide resources or insurance to households change the level and dispersion of the marginal value of wealth, which in turn affect (i) demand elasticities and (ii) sorting and from the firms perspective their demand elasticities, markups, and production.

One question is whether our assumption about consuming only one variety at a time or the time interval being large, is crucial. Our results — extensive margin demand with heterogeneous elasticities and sorting — are robust to assumptions regarding the number of goods consumed and the time interval. Appendices C and D directly answer these questions by characterizing the model under (i) the households buying one variety of each of the many goods and (ii) continuous time, where income shocks are less frequent than preference shocks. In both extensions, we find nearly exactly the same characterization of heterogeneous elasticities and sorting patterns as in the main text. Because these extensions share the same economics but with additional complexity, we keep the one-good, one-variety model as our baseline in the main text and quantitative analysis.

## 4. Calibration

This section describes how we calibrate the model, parameter estimates, and how the model performs with respect to un-targeted moments.

### 4.1. Externally calibrated parameters

**Functional forms.** The time period is a quarter. Utility depends only on the differentiated good and follows a CRRA form, with an intertemporal elasticity of substitution (IES) equal to $1/\sigma$. We impose a symmetry assumption such that each good is produced by $J$ firms. For each good the distribution of firms producing the different varieties is identical. Firms differ only in their quality, $\psi_j$, with no exogenous heterogeneity in productivity $z_j$. A common component of productivity in competitive and differentiated goods sectors $\overline{Z}$, is used to normalize mean output of differentiated goods producers to one. Quality, $\psi_j$, is Pareto distributed with tail parameter $\xi$. The focus on heterogeneity in quality is motivated by Hottman et al. (2016).

**Income process.** We follow Krueger et al. (2016) and assume that log efficiency units $\log e_t^i$, follow a first-order Markov process with an AR(1) component and a MA(1) transitory shock. The process is estimated using annual PSID data by Krueger et al. (2016) and is converted to a quarterly process following their methodology. We discretize this process into a ten-state Markov chain following Aguiar et al. (2023). The mean value of $e_t^i$ equals one, ensuring that the total supply of efficiency units of labor equals one.

**Fiscal policy.** We calibrate the fiscal side of the economy by combining the strategy of Kaplan et al. (2020) with the "liquid wealth calibration" of Kaplan and Violante (2022). Transfers $T$ and income taxes $\tau$ are set such that they represent 5 and 15 percent of GDP. The annual interest

rate is 2 percent. The borrowing constraint is zero ($\underline{a} = 0$). We then choose $\beta$ so that the ratio of average household assets to average annual household income equals 0.56, consistent with the ratio of liquid assets to income in the Survey of Consumer Finances.[8]

### 4.2. Internally calibrated parameters and moments

The six remaining parameters to calibrate are as follows. On the firm side: the number of firms $J$, the Pareto tail parameter of quality $\xi$, and decreasing returns $\alpha$. On the household side: across-good dispersion in tastes $\theta$, within-good across-variety dispersion in tastes $\eta$, and curvature in utility $\sigma$. We highlight that this is a relatively standard set of parameters. The parameters $J$, $\xi$, $\alpha$, $\theta$, and $\eta$ are essentially the same as in Edmond et al. (2023), so for comparability, we follow their approach, targeting firm concentration and markup statistics.

The only additional parameter is $\sigma$, which is standard in the incomplete markets model. For this, we replicate regressions in Auer et al. (2024). Hence, we have combined standard ingredients from two workhorse models—discrete choice and the standard incomplete markets model—without adding extra parameters to either. Parameter estimates are provided in Table 1. All parameters are jointly determined, and the model is exactly identified and matches all moments exactly. We describe this in detail below.

**Firm concentration, markups and market shares.** The number of firms $J$ and Pareto tail parameter $\xi$ are chosen to match industry concentration statistics as in Edmond et al. (2023). The concentration statistics that we use are from Amiti and Heise (2022), who compute concentration statistics in the U.S. at the NAICS 5 industry level, accounting for imports. For each industry, they compute a $HHI$ for sales and the share of sales accounted for by the top four firms. Taking an unweighted average across industries, their main finding is that measures of average concentration are flat between 1992 and 2012. The $HHI$ for sales is 0.0525, and the share of sales at the top four firms is 30.5%, which are our calibration targets.[9]

Parameters $\eta$ and $\theta$ determine the average markup and slope of markups with respect to market share. Conditional on all other parameters, changing $\eta$ by itself shifts the average markup in the economy. As in Edmond et al. (2023), we target a cost-weighted average markup of 1.25. Conditional on $\eta$, a lower $\theta$ steepens the slope of the markup with respect to market share. Edmond et al. (2023) argue that there is a modest, positive correlation between a firm's markup and its market share. Unlike in Edmond et al. (2023), this relationship cannot be derived in

[8] As argued by Kaplan and Violante (2022), targeting the level of liquid assets rather than total assets generates realistic marginal propensities to consume, consistent with a less tractable model with illiquid and liquid assets. In our calibrated economy the MPC out of a \$500 transfer in the quarter received is 24.4 percent, consistent with empirical evidence cited in Kaplan and Violante (2022).

[9] These are obtained from their Figures 1(d), 1(e), respectively. In 2002, both statistics are at a maximum of 0.054 and 31.5%. In 2007, these statistics are at a minimum of 0.051 and 30.2%. Hence targeting the average values over the period is without loss. When across-industry averages of these statistics are weighted by industry sales, both statistics decline only very slightly (Figure 1(d), 1(f)).

closed form. In fact, it is polluted by dual reasons why a larger firm has a larger markup — it's high share as in Edmond et al. (2023) and that it may sell to high income customers who are less elastic. Nevertheless, we can use the empirical relationship in Edmond et al. (2023) to inform $\theta$. We replicate their regression in our model:

$$\log \mu_j = \beta_0 + \beta_{EMX} \log \left( \frac{Sales_j}{\sum_{k=1}^{J} Sales_k} \right) + \eta_j, \tag{31}$$

We target their estimate of $\beta_{EMX}$ equal to 0.03.[10]

**Heterogeneity in elasticities.** Auer et al. (2024, henceforth, ABLV) document that poorer households have higher elasticities of substitution. They arrive at this conclusion by studying microdata on Swiss household purchases of Swiss versus French varieties within the same goods category following exogenous relative price changes due to the 2015 Swiss Franc appreciation. Subject to the same, exogenous, decline in prices of French varieties, poor households substituted spending toward French varieties at a significantly higher elasticity. We show how this evidence pins down the curvature of the utility function, $\sigma$.

The estimating equation in ABLV is as follows. Let $\mathcal{M}$ and $\mathcal{D}$ partition varieties of good $m$ into those that are i$\mathcal{M}$ported or $\mathcal{D}$omestic. Their empirical specification is:

$$\log \left( \frac{b^i_{\mathcal{M}t}}{b^i_{\mathcal{D}t}} \right) = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 \log \left( \frac{p_{\mathcal{M}t}}{p_{\mathcal{D}t}} \right) + \boldsymbol{\beta}_{ABLV} \log e^i \times \log \left( \frac{p_{\mathcal{M}t}}{p_{\mathcal{D}t}} \right) + \varepsilon^i_t, \tag{32}$$

This relates household-level budget shares $b^i_{\mathcal{M}t}$, $b^i_{\mathcal{D}t}$ to relative prices and household income $e^i$. The coefficient of interest is the interaction term between income and relative prices. The exchange rate appreciation is used as an instrument for relative prices. Their main empirical result is $\widehat{\boldsymbol{\beta}}_{ABLV} = 2.2$. This means that (exogenously) lower relative prices of imports lead to higher budget shares ($\widehat{\boldsymbol{\beta}_1} < 0$), but *less so* for richer households ($\widehat{\boldsymbol{\beta}}_{ABLV} > 0$). In other words, poorer households have higher elasticities of substitution than rich households.

The regression in (32) has a meaningful relationship in our model and disciplines $\sigma$. To see this, we construct budget shares for type $(a, e)$ households by aggregating across idiosyncratic preference shocks to compute the total share of household type $i = (a, e)$ spending on variety $jm$. Call this $b^i_{jm}$, the budget share of type $i$. In Appendix F.2 we derive the cross-sectional analog to (32) by log-linearizing the budget shares and differencing across $H$igh and $L$ow income

[10] We focus on their log-linear specification presented in the Appendix of Edmond et al. (2023), Table C4. We focus on their specification with sector×year effects. Using the non-linear specification that they derived in the body of their paper, we obtain similar results.

households leading to

$$\log\left(\frac{b_{jm}^H}{b_{km}^H}\right) - \log\left(\frac{b_{jm}^L}{b_{km}^L}\right) = \underbrace{\varepsilon_{km}^L\left(\frac{\partial \log x_{km}^L}{\partial \log e^L}\right)\left(-\frac{\partial \log \varepsilon_{km}^L}{\partial \log x_{km}^L}\right)}_{\text{Coefficient estimated in Auer et al. (2024)}} \underbrace{\log\left(\frac{e^H}{e^L}\right)\log\left(\frac{p_{jm}}{p_{km}}\right)}_{\text{Interaction term}}. \tag{33}$$

The estimated coefficient of ABLV consists of three terms. First, a baseline elasticity of demand is pinned down by our target for the average markup.[11] Second, the elasticity of consumption with respect to income. The consumption-savings properties of the incomplete markets economy determine this value. Third, how the elasticity of demand varies with consumption. From Section 3, if there were no intensive margin of demand, the third term would be exactly $(\sigma - 1)$, given our formula (25): $\varepsilon_{jm}^{\rho,i} = (\theta\rho_{j|m}^i + \eta(1 - \rho_{j|m}^i))x_{jm}^{i-(\sigma-1)}$. Hence, conditional on all other parameters, $\widehat{\beta}_{ABLV}$ determines $\sigma$.

The argument above establishes a tight connection between the coefficient estimated in ABLV and $\sigma$. However, in practice, there are issues that create a gap between the estimated coefficient and the value implied by the model, such as the first-order approximations used above possibly not holding, and the model having an intensive margin of demand. Given these concerns, we implement an indirect inference procedure in the calibration routine where we simulate the experiment of ABLV in our model and choose $\sigma$ to match the regression coefficient in our model with that seen in the data, $\widehat{\beta}_{ABLV} = 2.2$. Appendix F.3 details how we implement this procedure.

**Sorting.** In addition to household-level elasticities, a firm's elasticity of demand is shaped by the composition of households that a firm faces, i.e. sorting. To measure sorting, we follow Jaimovich et al. (2019) who extend Bils and Klenow (2001) to consumer packaged goods data.

The idea is to measure how unit prices paid vary across households within very narrow product categories. Because this analysis is done within a narrow scope, the interpretation of, for example, high income households paying higher prices is that they are purchasing varieties of higher quality relative to poor households. Both Bils and Klenow (2001) and Jaimovich et al. (2019) find this in the CEX and the Kilts-Neilsen dataset.

We update and extend the Jaimovich et al. (2019) analysis by controlling for variation in prices within product modules, across geographical areas (DMA).[12] We find that within-module-DMA-

[11] As noted by ABLV, their estimates cannot be used to determine the level of the elasticity of demand, only how it varies across individuals.

[12] Our analysis extends Jaimovich et al. (2019) in the following ways. First, we use within-module-DMA-year variation, whereas they use within-module and pool across years. This takes care of the possibility that higher income households purchase higher price products that are available only in some geographic areas. Second, we use interacted fixed effects, which make precise the variation being used: within-module-DMA-year, across-household. Third, as in Faber and Fally (2022) we use total annual consumption to split households. Since this is a continuous measure—rather than the coarse bins for income in Neilsen data—we are able to construct household income quantiles by within-DMA-year rankings of households.

**Table 1: Internally Calibrated Parameters**

| Parameter | | Value | Moment | Data | Model |
|---|---|---|---|---|---|
| **A. Households** | | | | | |
| Taste dispersion — within markets | $\eta$ | 8.9 | Average markup | 1.25 | 1.25 |
| Taste dispersion — across markets | $\theta$ | 0.0 | Edmond et al. (2023) reg. coefficient | 0.03 | 0.03 |
| Coefficient of relative risk aversion | $\sigma$ | 2.57 | Auer et al. (2024) reg. coefficient | 2.20 | 2.20 |
| Discount rate | $\beta$ | 0.99 | Mean liquid assets / Mean income | 0.56 | 0.56 |
| **B. Firms** | | | | | |
| Firms per market | $J$ | 25 | Amiti and Heise (2022) sales HHI | 525 | 525 |
| Tail parameter of Pareto | $\xi$ | 10.9 | Amiti and Heise (2022) top 4 share | 30.5 | 30.5 |
| Decreasing returns | $\alpha$ | 0.63 | Jaimovich et al. (2019) reg. coefficient | 0.14 | 0.14 |
| **C. Government** | | | | | |
| Income tax rate | $\tau$ | 0.27 | Total labor income taxes to GDP | 0.15 | 0.15 |
| Transfers per capita | $T$ | 0.05 | Total transfers to GDP | 0.05 | 0.05 |

year, high expenditure households purchase products that have 14.4 percent higher average prices than the products purchased by poor households. Appendix F.4 provides details of the regression that obtains this statistic.

This pattern of sorting disciplines how marginal cost varies with size. To see this, consider how the model works with decreasing returns, $\alpha < 1$. High quality firms are relatively large because all households enjoy quality. With decreasing returns to scale, this implies that high quality (and larger firms) have higher marginal costs, leading to higher prices. Then from our sorting condition (29), rich households are proportionally more present in the customer base of large firms. Through the lens of the Jaimovich et al. (2019) statistic, this shows up as rich households paying higher prices. Thus, we choose $\alpha$ to match the fact that, relative to the lowest quintile, the top spending households purchase products that have 14.4 percent higher prices.[13]

### 4.3. Parameter estimates

Table 1 summarizes our parameter estimates and calibration targets. We view the fact that the parameter estimates are well within range of standard parameter values as a virtue of our quantitative framework. Curvature in utility $\sigma$ is 2.57, showing that we do not need substantial

[13]This generates additional dispersion in prices on top of what would be obtained from only the markup variation, so one may ask what the overall price dispersion is in our model. Kaplan and Menzio (2015) measure the standard deviation of relative prices in the Kilts-Nielsen data, which varies across their market definitions from 0.19 to 0.36. Replicating their computations in the model, we obtain 0.14.

deviations from log utility to match the heterogeneity in demand elasticities across households documented by ABLV. Decreasing returns $\alpha$ of $0.63$ is consistent with parameter estimates in the firm dynamics literature, hence given the heterogeneity in demand elasticities, we did not need that much dispersion in marginal cost to generate sorting.

The levels of $\eta$, $\theta$ and $\xi$ are not directly comparable to other studies. Because of the non-homotheticity in our model, alternative normalizations regarding units in our economy would result in different values of $\eta$ and $\theta$ to match the calibration targets. Similarly, since the quality shifter enters utility, then the tail coefficient for the quality shifters would also change.[14]

### 4.4. Non-targeted moments

This section discusses implications of the calibration for moments that we did not target and are informative about core properties of our model. Most compelling is that our model is able to rationalize and tie together a set of seemingly disparate facts from scanner datasets, as well as explain how plausibly exogenous variations in prices or household income/wealth have led to new evidence on the interaction between household heterogeneity and pricing behavior.

**The extensive margin is the dominant determinant of firm sales.** An important feature of our paper is modeling discrete choice, and hence the focus on the extensive margin elasticity of demand across firms. Recent empirical work, such as Afrouzi et al. (2023) and Einav et al. (2021), motivates this choice. The analysis in Afrouzi et al. (2023) is most easily comparable to our model. They find that, within industry and year, across firms, those with high sales have high sales primarily due to having more customers, rather than higher sales per customer.

In our model, we can exactly replicate the Afrouzi et al. (2023) analysis. Log firm sales are:

$$\log p_j x_j = \log p_j \overline{x}_j + \log \overline{\rho}_j \quad , \quad \underbrace{p_j \overline{x}_j = \frac{\sum_i \rho_j^i p_j x_j^i \Lambda^i}{\sum_i \rho_j^i \Lambda^i}}_{\text{Average sales per customer}} \quad , \quad \underbrace{\overline{\rho}_j = \sum_i \rho_j^i \Lambda^i}_{\text{Customers}} . \tag{34}$$

A regression can then be used to decompose the contribution of total sales into (i) sales per customer, and (ii) customers.[15] Afrouzi et al. (2023) find that 86.1 percent of the variance is attributable to customers. Einav et al. (2021) document that close to 80 percent of firms' sales variation is due to differences in customer bases. Using the exact same approach as Afrouzi et al. (2023), we obtain a value of $53.4$ percent in our model. We conclude that, in reduced form regressions, our model is quantitatively consistent with the fact that the extensive margin

[14] Equation (12) can be written $\rho_{j|m}^i = \phi_{jm}(\exp\{v_{jm}^i\}/\exp\{\widetilde{v}_{jm}^i\})^\eta$, where $\phi_{jm} = \eta^{-1}\log\psi_{jm}$. Hence, different values of $\eta$, imply different values for the effective shifter in demand, $\phi_{jm}$, conditional on the distribution of $\psi_{jm}$ and its tail coefficient, $\xi$.

[15] In Afrouzi et al. (2023), the authors present a variance decomposition of (34) without fixed effects. We thank the authors for providing the regression results using industry-year fixed effects.

dominates in determining firm sales.[16]

**The role of quality, marginal cost, and markups in determining firm sales.** Hottman, Redding, and Weinstein (2016) provide a different perspective on firm sales. Through the lens of a demand system, they provide an accounting of within-industry dispersion in firm sales into a quality or "appeal" component, marginal costs, and markups. They find that quality is by far the dominant factor in making large firms large, while higher marginal costs and higher markups make those large firms slightly smaller.

Through introspection, one can see how our model is qualitatively consistent with these facts. First, because all households have the same valuation of quality, a higher quality firm will attract more customers (which is consistent with the extensive margin evidence above). Second, the calibrated model results in decreasing returns to scale to match patterns of sorting. Decreasing returns implies that larger firms have higher marginal costs. Third, the model targeted the Edmond et al. (2023) size-markup correlation with larger firms having higher markups. Thus, big firms are big because of quality, while higher marginal costs and higher markups make larger firms slightly smaller than otherwise.

**Sorting of households across prices of goods and sizes of firms.** Our calibration matches the fact that, relative to the lowest quintile, the top spending households purchase products that have 14.4 percent higher prices. Several more implications of our calibration on this dimension can be checked.

First, Figure 4A shows how the model performs across the entire distribution of expenditure. As the figure illustrates, throughout the distribution of households, the model generates sorting behavior consistent with the data.

Second, the model has predictions about how households sort across firms in terms of firms' total sales. Faber and Fally (2022) compute the log deviation of firms' sales from their industry and then compute the expenditure weighted average of firms' sales deviation across all households within different income quantiles. Their main result is that purchases of households in the top decile of household expenditure are from firms that are, on average, 27 percent larger than the firms purchased from by households in the bottom decile of household expenditure.

Figure 4B displays the sorting patterns of households to firms when we replicate the Faber and Fally (2022) analysis in our model. Quantitatively, the model is consistent with the data and specifically we closely match their headline statistics with purchases of households in the top decile of household expenditure being from firms that are 24.6 percent larger than the firms purchased from by households in the bottom decile of household expenditure.

[16] Applying the same decomposition to quantities, rather than sales, we find that 88.1 percent of quantity variation is due to differences in customer bases, rather than quantity per customer.

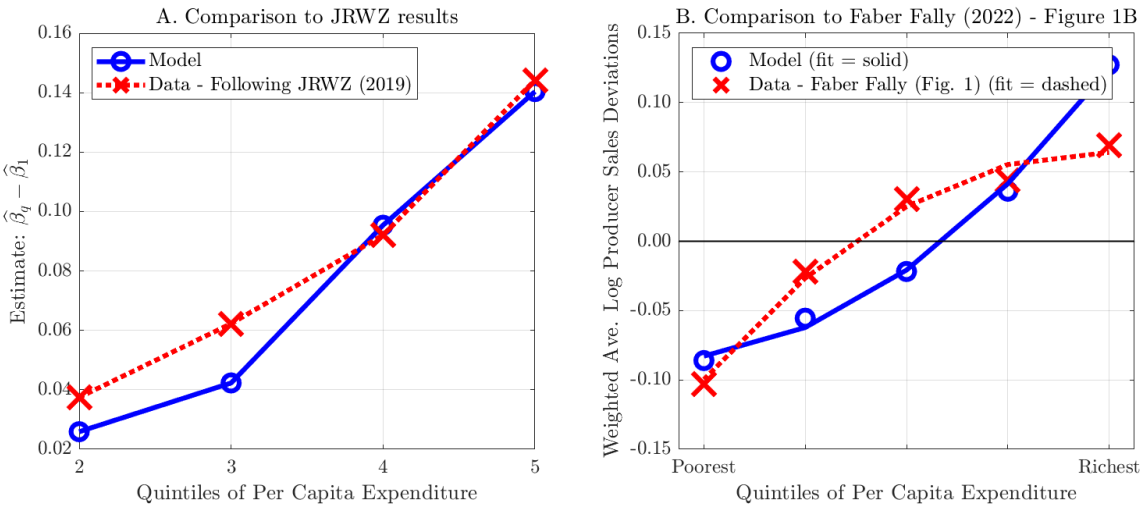


Figure 1: Sorting - Model and Data

**Markups and wealth shocks — Stroebel and Vavra (2019).** Finally, we show that the model reproduces empirical estimates of the effect of changes in wealth on markups. Stroebel and Vavra (2019, henceforth SV) show that following an increase in local housing wealth, prices of goods sold locally increase, while marginal costs remain unchanged. SV hypothesize that higher housing wealth reduces homeowners' price sensitivity, leading firms to increase markups in response. This evidence is particularly relevant as it speaks directly to the mechanisms in our model.

The approach of Stroebel and Vavra (2019) is to compare changes in markups in areas where there are large changes in wealth to areas in which there are smaller changes in wealth, where disparities across areas are due to the interaction between heterogeneity in changes in house prices and heterogeneity in areas' rate of home ownership. SV study both the housing boom from 2001-2006 and bust from 2007-2011; we focus on their results over the boom. They find that regions where initial wealth is more exposed to the increase in house prices had larger increases in markups.

To mimic their experiment we make 5,000 copies of our steady-state economy, indexed $n$ at date $t = 0$. Each of these copies is treated as a different region in the data. In each region, we randomly assign a fraction $\varphi_n$ households to be homeowners. In the next period, $t = 1$, we increase assets for homeowners by $\Delta_n$ percent, while leaving the assets of renters the same. We draw $\varphi_n$ and $\Delta_n$ from data from 2000 to 2006 provided in the replication package for SV.[17] We solve for firms' optimal prices in each location consistent with the new elasticities of demand and consumption decisions of the households in that location. We keep marginal costs fixed (consistent with observations in SV), the aggregate real interest rate fixed, and we don't clear

[17]Figure A1.A plots the distribution of ownership rates across U.S. ZIP codes in 2000 which give $\varphi_n$, and distribution of changes in U.S. ZIP code house price index between 2001 and 2006 which give $\Delta_n$.

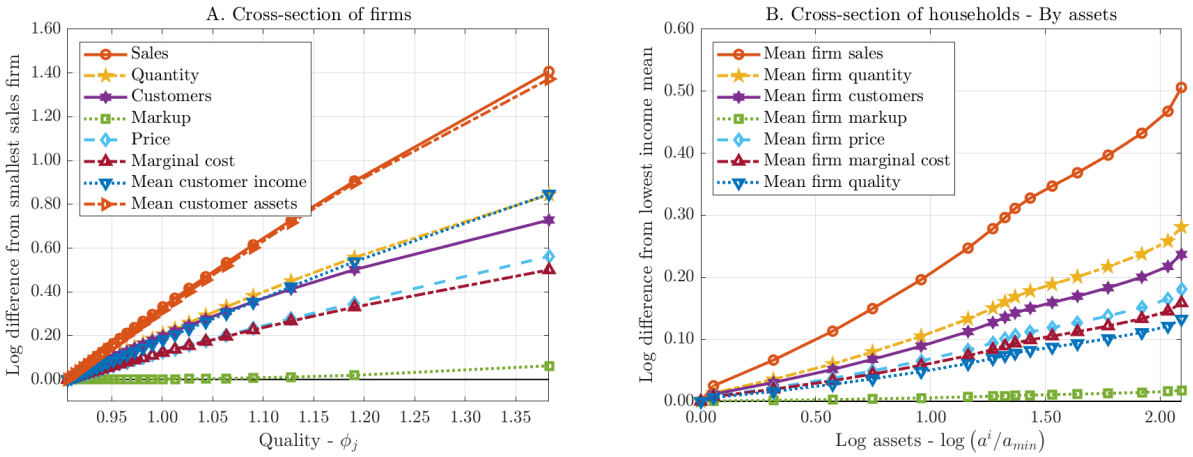


Figure 2: Cross-section of households and firms

Note: **Panel A.** Mean customer income is computed as total income of all customers of the firm, divided by total customers. Same applies for mean customer assets. **Panel B.** Let $\nu_j$ be a property of firm $j$ (e.g. sales, quantity, customers, markup, price, marginal cost, quality). We compute $\nu(a) = \int \left[\sum_j \nu_j \rho_j(a,e)\Lambda(a,e)\right] de / \int \Lambda(a,e) de$. We then plot $\log(\nu(a)/\nu(\underline{a}))$.

the asset market or government budget.

Given this simulated data we then replicate the main regression in SV (equation 3). Local price indexes $P_{nt}$ are constructed as in SV, using consumption expenditure weights updated each period. We then project the change in local price indexes on the change in log house prices, the initial home ownership rate, and their interaction:

$$\log P_{n,1} - \log P_{n,0} = \boldsymbol{\beta}_{\Delta}\Delta_n + \boldsymbol{\beta}_{\varphi}\varphi_n + \boldsymbol{\beta}_{SV}\Big(\Delta_n \times \varphi_n\Big) + \eta_n$$

Since locations in our experiment are ex-ante identical, we need no additional controls.

Our experiment in the model reproduces the key parameter estimates in Stroebel and Vavra (2019). SV's estimates of the interaction term $\boldsymbol{\beta}_{SV}$ range from 0.10 to 0.23 (SV, across Tables II, IV, A3). That is regions where initial wealth is more exposed to the increase in house prices had a larger increase in prices / markups. Replicating this exercise in our model, we obtain $\widehat{\boldsymbol{\beta}}_{SV}$ =0.112 which is consistent with the estimates of SV.[18]

### 4.5. Summary

Our first contribution is that our quantitative theory matches the moments in Table 1 and is qualitatively consistent with a range of further studies. Using a parsimonious set of parameters, the combination of a workhorse model in partial equilibrium industrial organization and a workhorse model in general equilibrium macroeconomics delivers a wide range of facts from both the household and firm sides of the economy, as well as their interaction. At a broad level, Figure 2 combines many of these observable patterns in the data. High-sales firms are high-

[18] Appendix Figure A1.B shows how the positive relationship between local price increases and home-ownership rates $(\varphi_n)$ becomes *steeper* as house price growth $(\Delta_n)$ increases, producing the positive $\boldsymbol{\beta}_{SV}$.

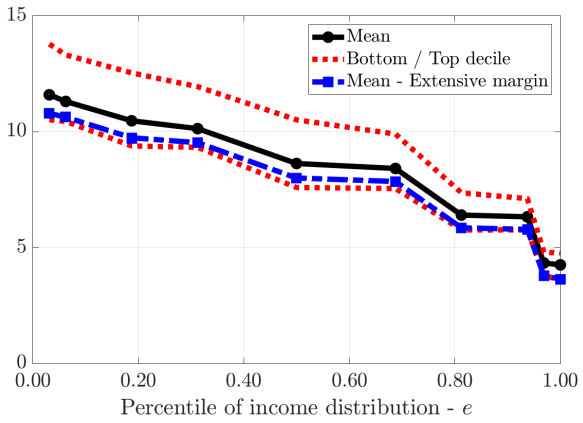


Figure 3: Household elasticities of demand by income

quality, with their high quantity owing to having many customers. They sell at high prices, with slightly higher markups, and their customers have higher incomes and more assets (Panel A). Wealthy households, on average, buy from large firms in terms of sales, customers, and quantities. They pay higher prices for higher quality and higher markups (Panel B). These empirical correlations have informed our modeling decisions. We now leverage the model for decomposition (Section 5) and counterfactual exercises (Section 6) that cannot be assessed empirically.

## 5. Results — Household Heterogeneity and Markups

This section asks the following question of the model: What is the role of household heterogeneity in determining markups in the cross-section of firms? Following equation (23), we examine the heterogeneity in household elasticities and sorting. We then use equation (25) to decompose demand elasticities across firms into components driven by market power and household heterogeneity. We find that household heterogeneity accounts for approximately 60 percent of the variation in demand elasticities across firms. We then discuss how alternative, nested models—those with no role for household heterogeneity or size-related market power—would yield different empirical patterns and answers to this question.

### 5.1. Household elasticities

Figure 3 illustrates the heterogeneity in individual demand elasticities across households as predicted by the model. Households in the bottom quintile of the income distribution have demand elasticities roughly three times higher than those in the top quintile. This decline in elasticities across income groups is consistent with estimates from Nakamura and Zerom (2010). They find that *"a household with an income one standard deviation above the mean has a price elasticity about 20% below the price elasticity of the median consumer"*. In our model, this same statistic is

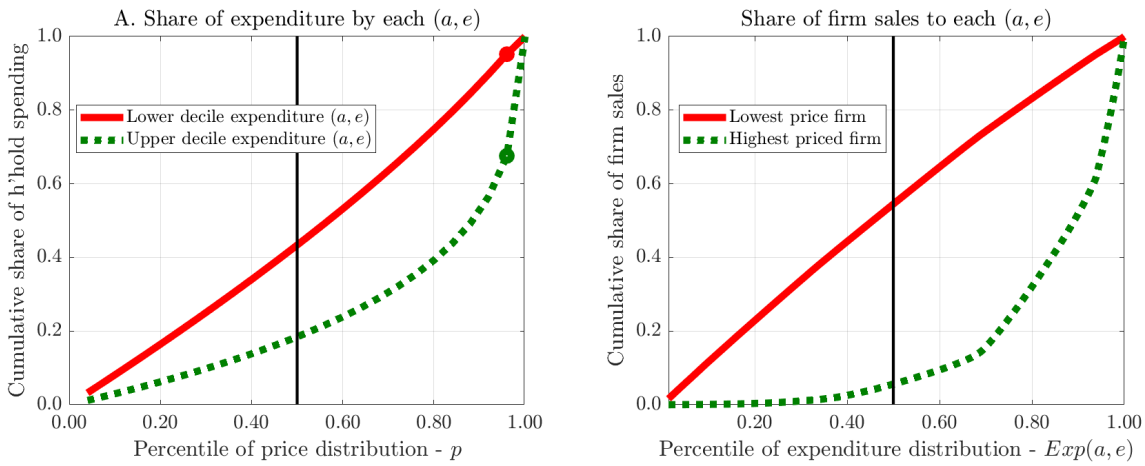


Figure 4: Sorting of **A.** Household expenditure across Firms, **B.** Firms sales across Households

18.1 percent.

Within income groups, there is significant heterogeneity in elasticities among the poor, but less so among the rich. This heterogeneity is driven by wealth differences. Among income-poor households, those that are against the borrowing constraint have the highest elasticities of demand, exceeding 14. Households with a large asset buffer have a lower marginal value of a dollar and, consequently, lower elasticities of demand. Income-rich households, whether with low or high assets, have similar elasticities, as high income provides low-asset households the opportunity to save, compressing the marginal value of a dollar across high-income households.

We find that the reduced form coefficient from the ABLV experiment correctly pin points the variation in elasticities of demand by income. This shouldn't necessarily be the case, since the experiment mimicked an empirical setting in which shocks to relative prices of bundles of goods were used, and a first order approximation was invoked. Directly regressing demand elasticities on log income yields a coefficient of $-2.19$. This result aligns with the coefficient estimated in our replication of ABLV. Thus, the experiment and associated reduced-form regression correctly identified the object that ABLV sought to estimate: how elasticities of demand vary with income.

We find that the extensive margin dominates households' demand elasticities. First, recall the relationship $\varepsilon^i_{jm} = \varepsilon^{i,\rho}_{jm} + \varepsilon^{i,x}_{jm}$. In a variance decomposition of $\varepsilon^i_{jm}$, 98.1 percent is accounted to by the variance of the extensive margin, $\varepsilon^{i,\rho}_{jm}$. Second, the blue line in Figure 3 plots the average extensive margin elasticity, which is only slightly below the overall elasticity. The dominance of the extensive margin elasticity is precisely how the model rationalizes the empirical fact that large firms are large on the extensive margin demand, while quantities sold per customer are similar to smaller firms (Afrouzi et al., 2023; Einav et al., 2021). If the intensive margin elasticity dominated, quantities sold per customer would account for level differences in sales.

### 5.2. Household sorting

The dominant role of the extensive margin elasticity shapes how customers sort themselves across firms. Panel A of Figure 4 starts from the consumer side and compares the distribution of expenditures of low expenditure households (red solid line) and high expenditure households (green dash line). While all households share identical preferences for goods' common quality and draw idiosyncratic utility from the same distributions, stark differences emerge in their spending patterns. For example, less than 20 percent of rich households' expenditure is at firms in the bottom half of the price distribution, while for poor households this is more than 40 percent. The incomplete markets models induces heterogeneity in the marginal value of wealth, which induces sorting as per equation (29). With a higher marginal value of wealth, poor households trade-off price discounts for idiosyncratic taste and focus their spending on cheaper, lower quality varieties.

From the firms' perspective, which is what matters for pricing, the distribution of spending is even more skewed. Panel B of Figure 4 shifts the focus to the firm side, illustrating the shares of a firm's sales that come from households at different points in the expenditure distribution. These shares correspond to the weights, $\omega_{jm}^{i}$, that firm $jm$ applies to households' demand elasticities in equation (23). For a low-price firm (red solid line) the consumer sorting in Panel A implies that more than half of its sales are to households in the bottom half of the expenditure distribution. In contrast, for high-price firms (green dash line), less than one-tenth of their sales come from households in the bottom half of the expenditure distribution. The strong consumer sorting implies that high priced goods will be priced to reflect the lower demand elasticities of high-income, high-wealth households, as shown in Figure 3. These elasticities are substantially lower than those of low-income households.

### 5.3. Firm demand elasticities and markups

Household heterogeneity in elasticities and sorting, combine to reduce price elasticities of high-price firms. However, as equation (25) suggested, this does not necessarily imply that high-price firms face less elastic demand overall. High-price firms could have small market shares. This would result in a more elastic oligopoly component that offsets the wealth component.

This turns out not to be the case. High-price firms in our model face *lower* price elasticities of demand because both the oligopoly and wealth components are smaller for these firms. Suppose all firms charged the same price, then the fact that quality variation drives variation in firms sales would lead to high quality firms being larger. Three forces would then cause high-quality firms to charge higher prices. First, a higher market share increases the desired markup through the oligopoly component. Second, a larger scale of production raises marginal costs. Third, with higher prices, high quality firms will also attract less elastic customers, an

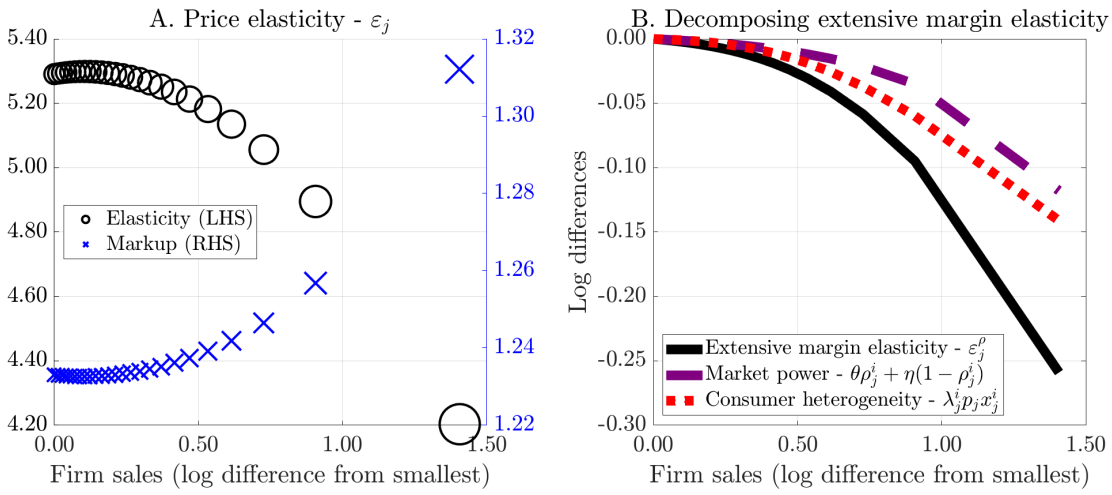


Figure 5: Decomposition of firms' elasticity of demand

additional force toward higher prices. Thus, in the cross-section of firms, both the oligopoly and wealth components reinforce each other in driving prices higher.

Which force plays the dominant role in shaping firms' markups? Figure 5A plots the firm-level elasticity of demand, $\varepsilon_j$, which we aim to decompose as it varies with firm sales. Large firms are also high price firms and have lower elasticities of demand (black circles) and hence higher markups (blue crosses). And to reemphasize the properties of this calibration: by construction we have matched properties of (a) the relationship between demand elasticities and income (Auer et al., 2024), and (b) the empirical relationship between relatives sales and markups (Edmond et al., 2023). Thus, the model successfully translates the large heterogeneity in household elasticities implied by (a) into the heterogeneity in firm elasticities implied by (b).

We decompose the extensive margin elasticity using the following approximation::

$$\varepsilon_j^\rho = \int \omega_j^i \left[\eta - (\eta-\theta)\rho_j^i\right] \lambda_j^i p_j x_j^i \, di \approx \underbrace{\int \omega_j^i \left[\theta\rho_j^i + \eta(1-\rho_j^i)\right] \, di}_{\text{Market power}} \times \underbrace{\int \omega_j^i \lambda_j^i p_j x_j^i \, di}_{\text{Household heterogeneity}} \tag{35}$$

Figure 5B presents the log difference in firms' $\varepsilon_j^\rho$ relative to the smallest firm—which is also lowest-quality and lowest-price—and the components of this log difference attributed to *Market power* (dash) and *Household heterogeneity* (dotted). Comparing firms in the top and bottom quintiles of the sales distribution, household heterogeneity accounts for 58.5 percent of the variation in demand elasticities.[19]

Overall, we find that the household heterogeneity channel plays a dominant role. We consider this result important because canonical macroeconomic approaches, such as those by Atkeson

[19]This statistic is 54.5 percent when comparing the highest- and lowest- price firm, and 63.0 when comparing the second-highest and lowest-price firm.

and Burstein (2008) and Edmond et al. (2023), attribute *all* markup variation to market power to size-based channels. In our model, we replicate key features of firm size and markups, similar to Edmond et al. (2023). However, we find that a majority of the variation in markup variation arises from the fact that different firms face different types of households, and these households are heterogeneous in their elasticities of demand.

### 5.4. Nested models

Another way to understand the role of household heterogeneity is by considering the implications of nested models where demand elasticities are independent of (a) household heterogeneity and (b) size-based market power. This also clarifies how incorporating new data on sorting and heterogeneity in household elasticities of substitution contributes to our results. In the absence of household heterogeneity, market power alone can explain the evidence from Edmond et al. (2023); however, the model fails to account for demand elasticity heterogeneity and sorting. In contrast, without size-based market power, household heterogeneity can explain the evidence from Edmond et al. (2023), but results in counterfactually large variation in demand elasticities by income and excessive sorting.

**Log model.** As discussed earlier, when $u(c) = \log c$ household heterogeneity does not influence firms' elasticities of demand. We therefore set $\sigma = 1$ and recalibrate the model to the same moments except for the statistics from Auer et al. (2024) and Jaimovich et al. (2019). Without variation in elasticities, there is no sorting, and consequently, both statistics are zero. Since dropping one parameter removed two moments, we select $\alpha$ to keep the standard deviation of log prices consistent with our benchmark model.[20] The second column of Table 2 presents the results and shows that the model can match the relationship between firm size and markups without household heterogeneity, just as effectively as models such as Edmond et al. (2023). Obviously, the model fails to deliver either statistic we dropped, and each are precisely zero.

**No oligopoly.** If we set $\eta = \theta$, then the only source of heterogeneity driving markups across firms is consumer heterogeneity. We recalibrate the model under this restriction and instead of choosing $\sigma$ to match the statistics from Auer et al. (2024), we chose $\sigma$ to match the relationship between firm size and markups in Edmond et al. (2023).[21] The third column of Table 2 presents the results, showing that (i) household heterogeneity *alone* can match the relationship between firm size and markups with a considerably higher $\sigma$, but (ii) this comes at the cost of generating a much more negative relationship between incomes and demand elasticities, leading to excessive sorting relative to the data.

[20] The parameters to calibrate are $\{\xi, \eta, \theta, \alpha\}$. These are chosen to match the concentration, average markup, $\beta_{EMX}$ and Std.$[\log p_j]$ from the baseline model.

[21] The parameters to calibrate are $\{\xi, \eta, \sigma, \alpha\}$. These are chosen to match the concentration, average markup, $\beta_{EMX}$ and Std.$[\log p_j]$ from the baseline model.

**Table 2: Alternative Calibrations**

| | | Baseline | Log model | Monopolistic Competition |
|---|---|---|---|---|
| **A. Household parameters** | | | | |
| Taste dispersion — within markets | $\eta$ | 8.9 | 2.12 | 11.7 |
| Taste dispersion — across markets | $\theta$ | 0.0 | 0.0 | — |
| Coefficient of relative risk aversion | $\sigma$ | 2.6 | — | 3.4 |
| **B. Firm parameters** | | | | |
| Tail parameter of Pareto | $\xi$ | 10.9 | 4.1 | 14.7 |
| Decreasing returns | $\alpha$ | 0.63 | 0.66 | 0.64 |
| **C. Moments** (* denotes targeted) | | | | |
| Average markup | | 1.25* | 1.25* | 1.25* |
| Amiti and Heise (2022) top 4 share | | 0.30* | 0.30* | 0.30* |
| Edmond et al. (2023) regression coefficient | | 0.03* | 0.03* | 0.03* |
| Auer et al. (2024) regression coefficient | | 2.20* | 0.0 | 2.62 |
| Jaimovich et al. (2019) regression coefficient | | 0.14* | 0.0 | 0.17 |
| Standard deviation of log prices | | 0.14 | 0.14* | 0.14* |
| **D. Elasticity Decomposition** | | | | |
| Share of elasticity variation due to household heterogeneity | | 58 | 0 | 100 |

### 5.5. Summary—Importance of household heterogeneity

Section 5 has established two conclusions. First, in our model, which matches a wide range of facts about households and firms, we find that household heterogeneity is the dominant factor in accounting for markup heterogeneity across firms, not size-based market power. Second, when comparing nested models, explaining the empirical patterns observed in the data requires that both household heterogeneity and market power forces shape markups, and these data discipline our decomposition of markups into these two forces.

## 6. Results—Fiscal Transfers and Markups

This section studies how changes in policy — a one-time fiscal transfer to households — affect markups and prices. We find that a transfer equivalent to one percent of GDP raises the aggregate markup in the economy by 0.3 percentage points, alongside a 0.4 percent increase in prices. These effects are driven by the core mechanisms emphasized throughout this paper: the transfer alters the distribution of the marginal value of wealth, making some households

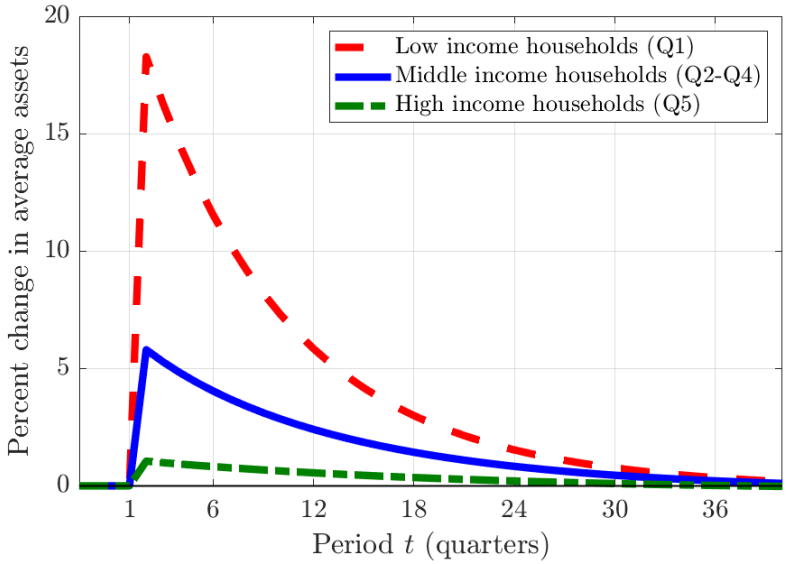

Figure 6: Fiscal transfer shock - Excess savings

more inelastic, and these households re-sort across firms. The result is higher markups for some firms, lower markups for others, and an overall increase in the aggregate markup.

### 6.1. Fiscal policy

The economy is in steady state in period $t = 0$. In period $t = 1$, the government announces and implements a path of transfers $\{T_t\}_{t=1}^{\infty}$. Transfers are deficit financed and paid off with increases in future income taxes $\{\tau_t\}_{t=1}^{\infty}$, while government spending remains fixed at $\overline{G}$, held at its originally calibrated value. The Government's budget constraints are

$$\overline{G} + T_t + R_{t-1}B_{t-1} = \tau_t W_t N_t + B_t \quad , \quad \tau_t = \tau \left( \frac{B_{t-1}}{B_0} \right)^{\phi_\tau} \quad , \quad t = 1, 2, \ldots \tag{36}$$

We follow Faria-e Castro (2024), and assume the path for $\tau_t$ endogenously depends on the excess of Government debt $B_{t-1}$ relative to steady-state $B_0$. In period 1, this implies that $\tau_t = \tau$. We choose $\phi_\tau$ to deliver a half-life of debt of 10 years.

We solve the transition of the economy away from and back to steady-state including the full distribution of prices $p_{jt}$. To focus on the direct effects of the transfer, we analyze a small open economy version where $R_t$ fixed at $R_0$ and the asset market does not clear. We consider a one-time increase in transfers equal to 1 percent of GDP.

Figure 6 plots the path for savings of households across the income distribution. Since households do not spend the entire transfer immediately, savings increase initially and are then slowly worn down. This pattern is reminiscent of the cross-section of household excess sav-

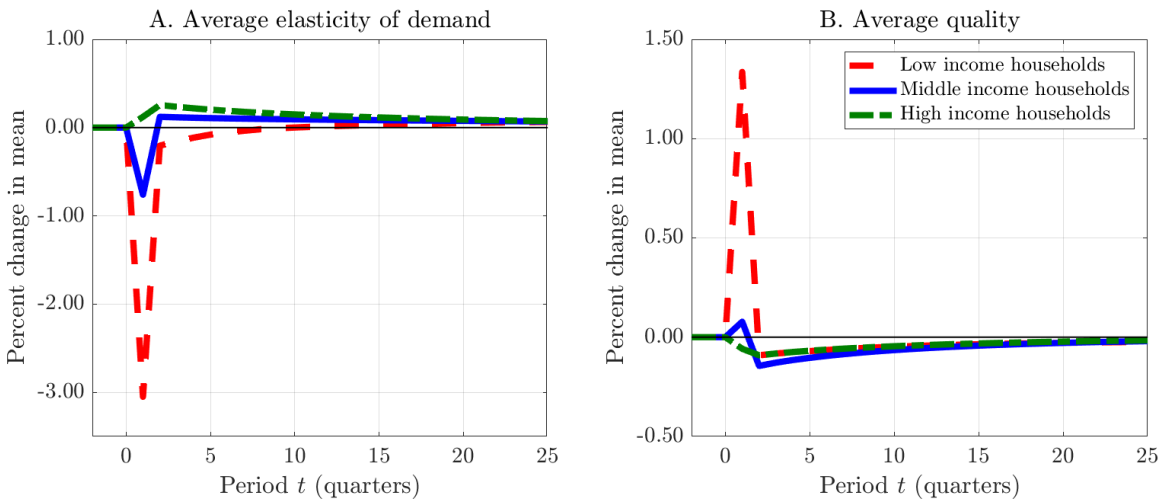


Figure 7: Fiscal transfer shock - Household level outcomes - Elasticities and sorting

ing during the pandemic period.[22] Given the lump-sum nature of the transfer, excess savings increase significantly more for poorer households, as the transfer constitutes a much larger fraction of their assets.

### 6.2. Markups and prices

Figure 7 plots how the transfer affects outcomes at the household level.Panel A shows how, for most agents, demand elasticities initially fall and then gradually rise over time. Thinking through equation (25), a fiscal transfer has the direct effect of making households wealthier, reducing the marginal utility of wealth, and hence these households become more inelastic. For wealthier agents, this effect goes in the opposite direction because of the fiscal consequences of the transfer. Unlike the poor households, rich households are essentially Ricardian. They reduce their consumption and save to pay for the higher tax burden, making them slightly more elastic. From the perspective of the firm, these results imply that their customer base becomes more or less elastic depending upon the types of households they serve.

These shifting elasticities of demand lead to a reallocation of households across the quality distribution of products. Panel B plots a measure of average quality consumed by the type of household. As poor households become less price elastic, they strongly reallocate their spending to higher quality, higher price products and hence "trade up" in response to the transfer. From the perspective of a low-quality firm, its customer base will be changing as it loses its least elastic customers to higher quality firms. From the perspective of a high-quality firm, its customer base will be becoming more elastic as it attracts lower income households.

[22]To put our fiscal transfer shock into context, the JP Morgan Chase Household Pulse Survey [link] found an increase in median monthly real cash balances of first income quartile households of 90 percent, and top quartile households of 50 percent. This lines up with the fact that pandemic period excess savings were around 7.56% of GDP (Abdelrahman and Oliveira, 2023), which is seven times our transfer. By comparison, poor households' excess savings increases by 90 percent in the data, which is around 5 times the response in the model (Figure 6).

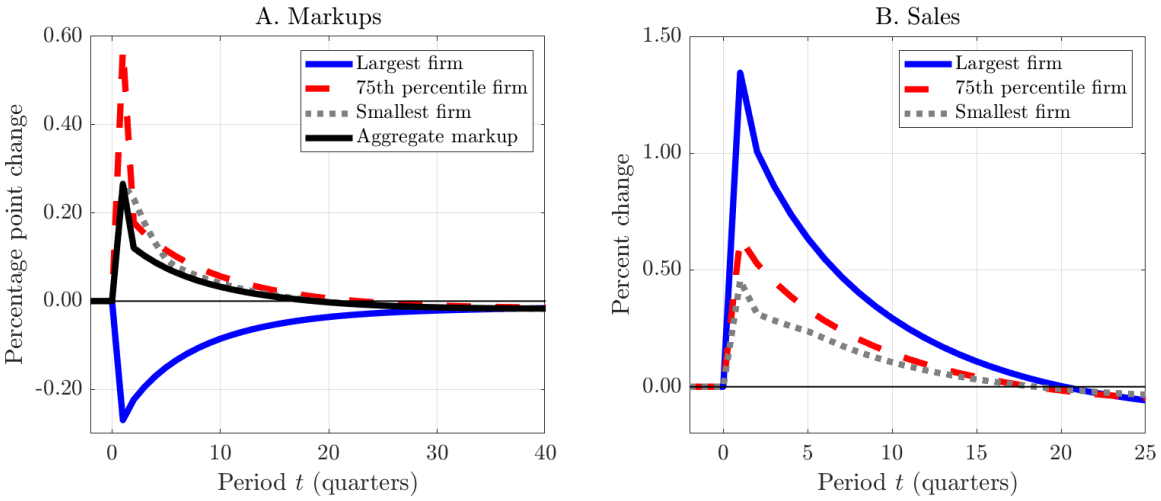


Figure 8: Fiscal transfer shock - Firm-level outcomes - Markups and sales

Figure 8 plots how firms' markups and size change in response to the fiscal transfer. The blue line depicts the largest firm (highest quality), the red dashed line depicts the 75th percentile of the distribution, and the grey dashed line is the smallest firm (lowest quality). The black line in Panel A is the cost-weighted aggregate markup.

Several things stand out. First, in response to the one-time fiscal transfer equivalent to one percent of GDP, the aggregate markup increases by 0.3 percentage points, from 1.25 to 1.28. Not surprisingly, overall demand increases for all firms, but more so for the largest firms given the "trading-up" behavior described above. As the largest firm gains market share, the market becomes more concentrated. However, this is not the primary reason for rising markups. Instead, markups increase because of declining demand elasticities and the sorting of poor and middle-income households.

For example, the largest firm lowers its markup despite gaining market share. From the perspective of a size-based theory of market power, the increasing relative share of the largest firm would typically lead to a higher markup. However, two forces arising from consumer heterogeneity push in the opposite direction. First, in terms of new customers, the firm's size increases due to a broadening customer base, comprising more elastic new customers. Second, in terms of existing customers, the largest firm has a relatively rich customer base, and as we saw in Figure 7A, their elasticities slightly increase. Both forces act in the direction of a higher firm-level demand elasticity and lower markup, which dominate the impetus toward higher markups due to increasing market share.

Second, given the declining markup of the largest firm, the aggregate response is shaped by the behavior of small and medium-sized firms, whose markups increase. These firms lose market share but raise their markups. Again, this is the opposite of what a purely size-based theory of market power would predict. These firms gain more elastic customers from lower-quality

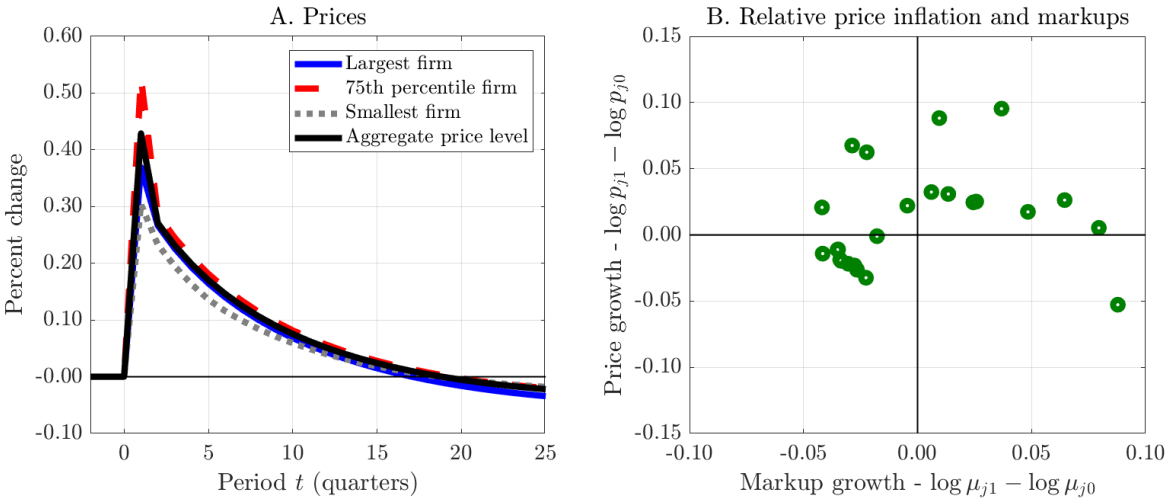


Figure 9: Fiscal transfer shock - Firm level outcomes - Prices

firms while losing their least elastic customers to higher-quality firms. While this represents offsetting forces, their inframarginal customer base is becoming more inelastic due to the transfer. This makes sense given that in our calibration, small and medium-sized firms are oriented to poor and middle-income households (Figure 4, sorting) for which demand elasticities are more sensitive to changes in wealth (Figure 3, elasticities).

Figure 9A plots the effect on prices. All firms raise their prices and the aggregate price level increases by about 0.4 percent, with heterogeneity across firms. Prices incorporate both changes in markups and marginal cost, which explains their co-movement. As the transfer increases overall demand (Figure 8B), marginal costs for all firms also rise. This increase is sufficient at the largest firm to lead to an increasing price, despite a declining markup. Overall, the increase in markups dominates the response of prices. From an accounting perspective, three-quarters of the increase in the price level is attributed to increases in markups rather than marginal costs.

Finally, we connect with empirical evidence that questions how increases in markups and increases in prices may or may not be correlated. To inform the policy debate about the connection between changes in market power and the inflation, Conlon et al. (2023) studies changes in markups and relative prices during the Covid-19 pandemic. Conlon et al. (2023) find that, in the cross-section of U.S. publicly traded firms, there was little correlation between changes in prices (PPI growth relative to the aggregate price level) and changes in markups. Figure 9B replicates the Conlon et al. (2023) exercise in our model, generating a similarly weak correlation.

The above highlights how household heterogeneity is crucial to understanding the increase in markups following a fiscal transfer. To reiterate this point, Figure 10 shows what happens to markups and prices in an economy where the role of household heterogeneity in demand elasticities is eliminated. For this, we use the model calibrated under log preferences, as shown in Table 2. As Figure 10A shows, nothing happens to markups. A fiscal transfer decreases the

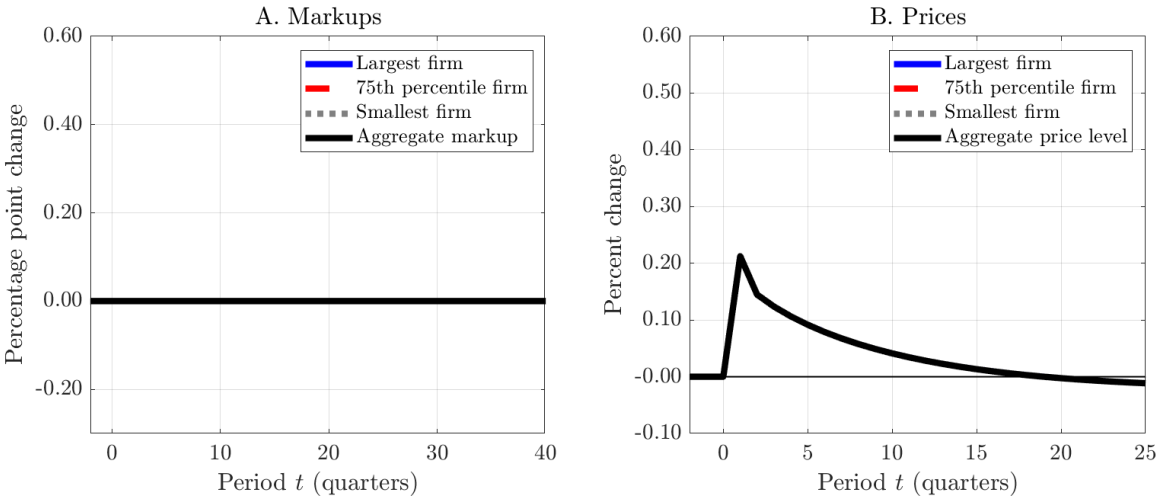


Figure 10: Fiscal transfer shock - Log model - Markups and Prices

marginal value of wealth for households, and leads households to consume more (especially among poor, high-MPC households), but this has *no* effect on elasticities of demand. With no effect on elasticities of demand, there is no change in sorting, and thus no shifts in firms' market shares. Hence shutting down the effect of household heterogeneity on elasticities of demand also kills changes in market shares, which are also a motive for changes in markups. Demand for all firms increases uniformly, marginal costs rise equally, and prices increase by the same amount (Figure 10B).

**Summary.** In Section 5 we demonstrated how household heterogeneity plays a critical role in shaping the cross-sectional distribution of markups. This section reinforces this message by showing—in response to an economic policy widely studied in heterogeneous agent economies—that household heterogeneity is essential for shaping the response of markups, thereby contributing to aggregate inflation.

## 7. Conclusion

This paper develops, quantifies, and tests a framework that enables macroeconomists to study how heterogeneity in income and wealth influences firms' pricing decisions. Our model aligns with empirical evidence on how poor versus rich households substitute between goods (Auer et al., 2024), how firm prices respond to changes in household wealth (Stroebel and Vavra, 2019), and a broad set of facts from household spending and firm pricing data. We demonstrate that this behavior is quantitatively important for understanding the cross-section of markups across firms and the aggregate effects of a canonical fiscal policy shock. Our model features limited free parameters and generates endogenous, policy-dependent notions of household sorting across prices, demand elasticities, and markups.

The core ideas of this paper are twofold: (i) the endogenous distribution of wealth shapes

the distribution of the marginal value of a dollar, and hence how individuals choose between goods with different prices; (ii) in imperfectly competitive markets, this affects pricing and hence production decisions. We believe these ideas have broad applicability. For instance, Waugh (2023) applies (i) showing how household heterogeneity shapes the pattern and gains from trade. Similarly, Berger et al. (2023) apply (i) and (ii) in a labor market context to show how progressive income taxes reduce individual elasticities of labor supply across firms, which distorts wages and employment when internalized by firms.

From a macroeconomic perspective, our model has several implications for the transmission of macroeconomic shocks and policy changes. We have been careful to nest the canonical incomplete markets model, enabling extensions that build on this widely used model. For example, are the positive effects of unemployment insurance Hansen and İmrohoroğlu (1992) or universal basic income (Daruich and Fernández, 2024) partially undone by reduction in households' price sensitivity and firms' markup responses? If integrated into a nominal economy, how do the distributional impacts of monetary policy (Kaplan et al., 2018) shape markup responses of firms? Additionally, the model could help explain the cyclicality of markups observed by Nekarda and Ramey (2020).

From a microeconomic perspective, our model provides insights into how firms may choose alternative strategies. For instance, one could investigate how innovation targets different market segments due to the heterogeneity in sales and optimal markups across the household distribution. Incorporating multi-product firms into the model could reveal how firms design product offerings at various price points to segment buyers. While we have abstracted from firm entry and dynamics, future research could explore how a firm's customer base and hence markup evolve over the lifecycle. Crucially, these questions can be addressed within a general equilibrium framework.

## References


ABDELRAHMAN, H. AND L. E. OLIVEIRA (2023): "The Rise and Fall of Pandemic Excess Savings," *FRBSF Economic Letter*, 2023, 1–6.

ACHDOU, Y., J. HAN, J.-M. LASRY, P.-L. LIONS, AND B. MOLL (2021): "Income and Wealth Distribution in Macroeconomics: A Continuous-Time Approach," *The Review of Economic Studies*, 89, 45–86.

AFROUZI, H., A. DRENIK, AND R. KIM (2023): "Concentration, Market Power, and Misallocation: The Role of Endogenous Customer Acquisition," NBER Working Papers 31415, National Bureau of Economic Research, Inc.

AGUIAR, M. A., M. AMADOR, AND C. ARELLANO (2023): "Pareto Improving Fiscal and Monetary Policies: Samuelson in the New Keynesian Model," NBER Working Papers 31297, National Bureau of Economic Research, Inc.

AIYAGARI, S. R. (1994): "Uninsured Idiosyncratic Risk and Aggregate Saving," *The Quarterly Journal of Economics*, 109, 659–684.

AMITI, M. AND S. HEISE (2022): "U.S. Market Concentration and Import Competition," Working Papers 22-34, Center for Economic Studies, U.S. Census Bureau.

ARGENTE, D., D. FITZGERALD, S. MOREIRA, AND A. PRIOLO (2021): "How do entrants build market share? The role of demand frictions," *American Economic Review: Insights*.

ATALAY, E., E. FROST, A. SORENSEN, C. SULLIVAN, AND W. ZHU (2023): "Scalable Demand and Markups," Working Papers 23-15, Federal Reserve Bank of Philadelphia.

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

AUER, R., A. BURSTEIN, S. LEIN, AND J. VOGEL (2024): "Unequal Expenditure Switching: Evidence from Switzerland," *The Review of Economic Studies*, 91, 2572–2603.

BAQAEE, D. R., E. FARHI, AND K. SANGANI (2024a): "The Darwinian Returns to Scale," *The Review of Economic Studies*, 91, 1373–1405.

——— (2024b): "The Supply-Side Effects of Monetary Policy," *Journal of Political Economy*, 132, 1065–1112.

BERGER, D., K. HERKENHOFF, AND S. MONGEY (2023): "Labor market power and optimal tax progressivity," .

BERRY, S., J. LEVINSOHN, AND A. PAKES (1995): "Automobile Prices in Market Equilibrium," *Econometrica*, 63, 841–90.

BEWLEY, T. (1979): "The optimum quantity of money," *Discussion Paper*.

BILS, M. AND P. J. KLENOW (2001): "Quantifying Quality Growth," *American Economic Review*, 91, 1006–1030.

BOAR, C. AND V. MIDRIGAN (2019): "Markups and Inequality," NBER Working Papers 25952, National Bureau of Economic Research, Inc.

BURDETT, K. AND K. L. JUDD (1983): "Equilibrium price dispersion," *Econometrica: Journal of the Econometric Society*, 955–969.

COMIN, D., D. LASHKARI, AND M. MESTIERI (2021): "Structural Change With Long-Run Income and Price Effects," *Econometrica*, 89, 311–374.

CONLON, C., N. H. MILLER, T. OTGON, AND Y. YAO (2023): "Rising Markups, Rising Prices?" *AEA Papers and Proceedings*, 113, 279–283.

DARUICH, D. AND R. FERNÁNDEZ (2024): "Universal Basic Income: A Dynamic Assessment," *American Economic Review*, 114, 38–88.

DÖPPER, H., A. MACKAY, N. H. MILLER, AND J. STIEBALE (2024): "Rising Markups and the Role of Consumer Preferences," NBER Working Papers 32739, National Bureau of Economic Research, Inc.

EDMOND, C., V. MIDRIGAN, AND D. Y. XU (2023): "How Costly Are Markups?" *Journal of Political Economy*, 131, 1619–1675.

EINAV, L., P. J. KLENOW, J. D. LEVIN, AND R. MURCIANO-GOROFF (2021): "Customers and Retail Growth," NBER Working Papers 29561, National Bureau of Economic Research, Inc.

ESLAVA, M., J. HALTIWANGER, AND N. URDANETA (2024): "The Size and Life-Cycle Growth of Plants: The Role of Productivity, Demand, and Wedges," *The Review of Economic Studies*, 91, 259–300.

FABER, B. AND T. FALLY (2022): "Firm Heterogeneity in Consumption Baskets: Evidence from Home and Store Scanner Data [Measuring Trends in Leisure: The Allocation of Time over Five Decades]," *Review of Economic Studies*, 89, 1420–1459.

FAJGELBAUM, P., G. M. GROSSMAN, AND E. HELPMAN (2011): "Income distribution, product quality, and international trade," *Journal of political Economy*, 119, 721–765.

FARIA-E CASTRO, M. (2024): "The St. Louis Fed DSGE Model," Working Papers 2024-014, Federal Reserve Bank of St. Louis.

GUPTA, A. (2024): "Demand for Quality, Variable Markups and (Mis)allocation: Evidence from India," Working paper, Dartmouth College.

HANDBURY, J. (2021): ""Are Poor Cities Cheap for Everyone? Non-Homotheticity and the Cost of Living across US Cities," *Econometrica*, 89, 2679–715.

HANSEN, G. D. AND A. İMROHOROĞLU (1992): "The role of unemployment insurance in an economy with liquidity constraints and moral hazard," *Journal of political economy*, 118–142.

HOTTMAN, C. J., S. J. REDDING, AND D. E. WEINSTEIN (2016): "Quantifying the Sources of Firm Heterogeneity," *The Quarterly Journal of Economics*, 131, 1291–1364.

HUGGETT, M. (1993): "The Risk-free Rate in Heterogeneous-agent Incomplete-insurance Economies," *Journal of Economic Dynamics and Control*, 17, 953–969.

IMROHOROĞLU, A. (1989): "Cost of business cycles with indivisibilities and liquidity constraints," *Journal of Political Economy*, 97, 1364–1383.

JAIMOVICH, N., S. REBELO, A. WONG, AND M. B. ZHANG (2019): "Trading Up and the Skill Premium," in *NBER Macroeconomics Annual 2019, volume 34*, National Bureau of Economic Research, Inc, NBER Chapters, 285–316.

KAPLAN, G. AND G. MENZIO (2015): "The Morphology Of Price Dispersion," *International Economic Review*, 56, 1165–1206.

——— (2016): "Shopping Externalities and Self-Fulfilling Unemployment Fluctuations," *Journal of Political Economy*, 124, 771–825.

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

——— (2020): "The Great Lockdown and the Big Stimulus: Tracing the Pandemic Possibility Frontier for the U.S," NBER Working Papers 27794, National Bureau of Economic Research, Inc.

KAPLAN, G. AND G. L. VIOLANTE (2010): "How Much Consumption Insurance beyond Self-Insurance?" *American Economic Journal: Macroeconomics*, 2, 53–87.

KAPLAN, G. AND G. L. V. VIOLANTE (2022): "The Marginal Propensity to Consume in Heterogeneous Agent Models," *Annual Review of Economics*, 14, 747–775.

KLENOW, P. J. AND J. L. WILLIS (2016): "Real Rigidities and Nominal Price Changes," *Economica*, 83, 443–472.

KRUEGER, D., K. MITMAN, AND F. PERRI (2016): "Macroeconomics and Household Heterogeneity," *Handbook of Macroeconomics*, 2, 843–921.

LOECKER, J. D., J. EECKHOUT, AND S. MONGEY (2021): "Quantifying Market Power and Business Dynamism in the Macroeconomy," NBER Working Papers 28761, National Bureau of Economic Research, Inc.

MCFADDEN, D. (1974): "Conditional logit analysis of qualitative choice behavior," in *Frontiers in Econometrics, edited by P. Zarembka, 105-142*, Academic Press.

MONGEY, S. AND M. E. WAUGH (2024): "Discrete Choice, Complete Markets, and Equilibrium," NBER Working Papers 32135, National Bureau of Economic Research, Inc.

NAKAMURA, E. AND D. ZEROM (2010): "Accounting for Incomplete Pass-through," *The Review of Economic Studies*, 77, 1192–1230.

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

NEVO, A. (2000): "A Practitioner's Guide to Estimation of Random-Coefficients Logit Models of Demand," *Journal of Economics & Management Strategy*, 9, 513–548.

NORD, L. (2023): "Shopping, Demand Composition, and Equilibrium Prices," Working paper, Federal Reserve Bank of Minneapolis.

OLIVI, A., V. STERK, AND D. XHANI (2024): "Optimal Monetary Policy during a Cost-of-Living Crisis," .

PYTKA, K. (2018): "Shopping Effort in Self-Insurance Economies," Working paper, University of Mannheim.

SANGANI, K. (2023): "Markups Across the Income Distribution: Measurement and Implications," Working paper, Harvard University.

STROEBEL, J. AND J. VAVRA (2019): "House Prices, Local Demand, and Retail Prices," *Journal of Political Economy*, 127, 1391–1436.

VERBOVEN, F. (1996): "The Nested Logit Model and Representative Consumer Theory," *Economics Letters*, 50, 57–63.

WAUGH, M. E. (2023): "Heterogeneous Agent Trade," Working Paper 31810, National Bureau of Economic Research.

# Appendix For Online Publication

This Appendix is organized as follows. Section A provides additional figures and tables. Section B derives all results for a simple 2 firm 2 household-type model. Section C derives elasticity and sorting results for an alternative model in which households buy one variety of every good each period. Section D derives elasticity and sorting results for an alternative model in which households buy one variety every instant, in continuous time. Section E derives results using our approach for the model in Fajgelbaum et al. (2011) and contrasts our approaches and results. Section F provides additional computational and mathematical details.

## A. Additional figures and tables

### A.1. Data used in Stroebel Vavra (2019) replication and model simulation

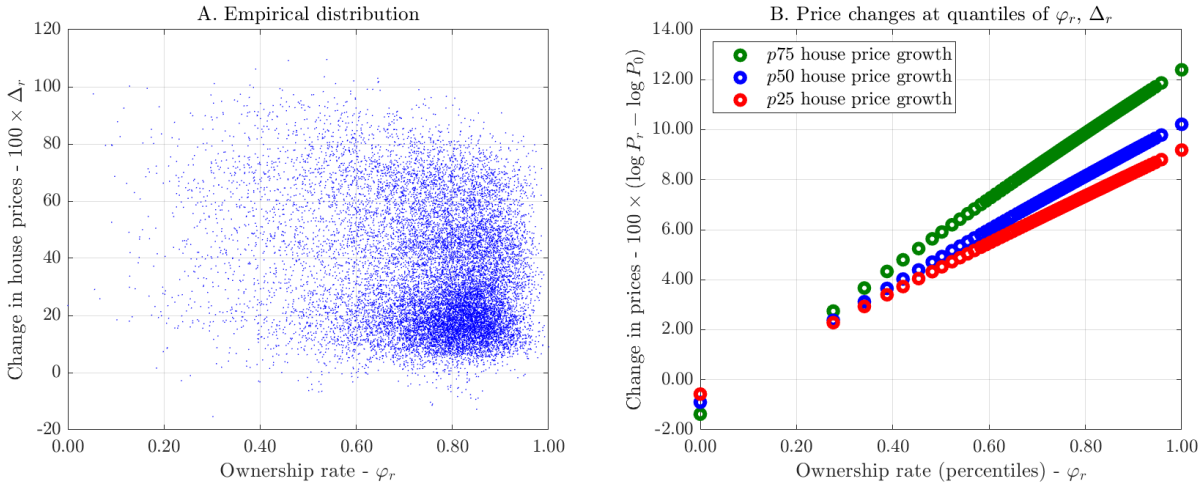


Figure A1: Home ownership rates and changes in house price

Notes: **Panel A.** Home ownership rate data from replication package for Stroebel and Vavra (2019), taken from 2000 Census. House price data from *Federal Housing Finance Agency - House Price Index Datasets*. **Panel B.** We produce the change in log prices in the model for an example set of draws of $\varphi_n$ and $\Delta_n$. The example set considers $\Delta_n$ at the 25th percentile (red), median (blue) and 75th percentile (green), and then draws of $\varphi_n$ at each of 100 percentiles of the empirical distribution.

## A.2. Marginal propensities to consume

Figure A2 shows that the model generates an empirically realistic distribution of $mpc$'s, with an average $mpc$ of 24.4 percent, and a fat tail of households with low income and high $mpc$'s.

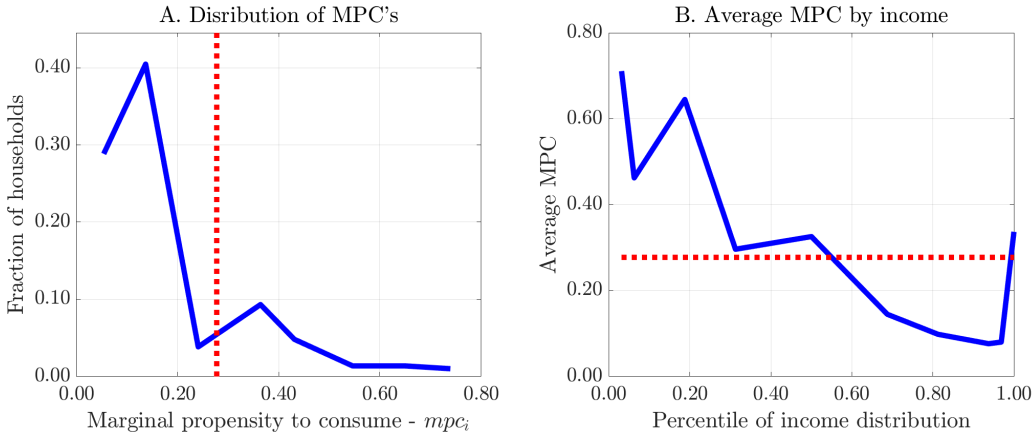


Figure A2: Household heterogeneity in the marginal propensity to consume

Note: The $mpc^i$ is computed out of an unexpected one-time payment of \$500. We measure the counterfactual increase in total consumption expenditure in the quarter after the payment, and divide by the size of the payment.

## A.3. Markup and price growth following fiscal transfer shock

This figure applies the methodology of Conlon et al. (2023, Figure 2) to data from the model. We compute markup and price growth at each firm $j$ using data from steady-state $(t = 0)$, to period $t = 2$ after the fiscal transfer policy shock.

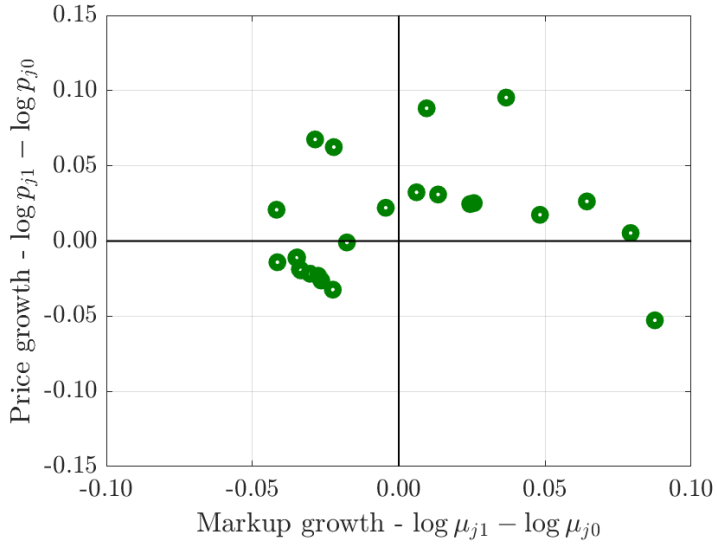


Figure A3: Markup and price growth across firms following a fiscal transfer

## B. Simple model — Two household types, two goods

In this section, we lay out a simple model for pedagogical purposes. The aim is to show how the combination of discrete choice, market power, quality heterogeneity and consumer heterogeneity simultaneously allow us to match five key features of the data: (i) firms are large because of the extensive margin, (ii) large firms are large because of quality, and smaller because of higher markups, (iii) high income households have lower price elasticities of demand, (iv) high income households sort into high quality firms. We also clarify the role of $\sigma > 1$. In Appendix E we contrast this to a model with multiplicative quality. We apply the same derivations and show that a model with multiplicative quality delivers (i) and (iv) but has opposite implications for (ii), (iii).

**Model.** There is a single good. There are two firms $j \in \{1, 2\}$ that produce different varieties of this good, with different qualities $\psi_1 > \psi_2$. Production requires only labor and has a constant marginal cost equal to the wage, $W$. There is a unit mass of two types of households $i \in \{L, H\}$ with different endowments of skills $e^i$, and income $y^i = We^i$, so $y^H > y^L$. Preferences are:

$$U^i = \max_j \; V_j^i + \psi_j + \zeta_j^i \tag{B1}$$

$$V_j^i = \max_{x_j^i} \; u\Big(x_j^i\Big) \quad \text{subject to} \quad p_j x_j^i = y^i. \tag{B2}$$

where the taste shock $\zeta_j^i$ is distributed Type 1 Extreme Value with tail parameter $\eta$. In this simple model, we abstract from the nesting structure, however still maintain the assumption that the firms behave oligopolistically when setting prices. In this simple model, all consumers buy one good which practically leads to an elasticity 'across markets' of zero.

Attached to the budget constraint, $p_j c_j^i = y^i$ which varies good by good, there is the multiplier $\lambda_j^i$. Then given the $V_j^i$s the household chooses the utility maximizing good $j$.

**Demand.** Like in the main model, demand is determined by aggregating across the two types of households as in (22). Individual demands are $\rho_j^i x_j^i$, which has an extensive margin component $\rho_j^i$ and intensive margin component $x_j^i$. The extensive margin component is the choice probability, and the intensive margin is read off the budget constraint. Let $v_j^i = \exp\{V_j^i\}$, and take a monotonic transformation of $\frac{1}{\eta}\log\phi_j = \psi_j$. Then demand is

$$\rho_j^i = \phi_j \left(\frac{v_j^i}{\widetilde{v}^i}\right)^\eta \quad , \quad \widetilde{v}^i = \left[\phi_1\Big(v_1^i\Big)^\eta + \phi_2\Big(v_2^i\Big)^\eta\right]^{1/\eta} \quad , \quad x_j^i = \frac{y^i}{p_j}. \tag{B3}$$

**Demand Elasticities.** Same idea as in the main model, the elasticity of demand of type $i$ for

good $j$ is:

$$\varepsilon_j^i = \left\langle -\frac{\partial \log \rho_j^i}{\partial \log p_j} \right\rangle + \left\langle -\frac{\partial \log x_j^i}{\partial \log p_j} \right\rangle = \left\{ \frac{\partial \log \rho_j^i}{\partial \log v_j^i} \right\} \left\{ -\frac{\partial \log v_j^i}{\partial \log p_j} \right\} + 1 \tag{B4}$$

$$= \left\{ \eta\Big(1 - \rho_j^i\Big) \right\} \left\{ \lambda_j^i p_j x_j^i \right\} + 1 \tag{B5}$$

$$= \eta\Big(1 - \rho_j^i\Big) x_j^{i-(\sigma-1)} + 1 \quad \text{with u being CRRA.} \tag{B6}$$

The first equation splits the elasticity into an extensive and intensive margin component, which is equal to one. The extensive margin component has two pieces.

The first piece is the elasticity of the choice probability with respect to the value. If firm $j$ is very low quality, changes in its value have very little impact on the market value $\widetilde{v}^i$, and hence this term is just the tail coefficient on the Gumbel distribution, $\eta$. This is seen as the first term approaches $\eta$ as $\rho_j^i$ becomes small. If firm $j$ is very high quality, changes in its value move the market value almost one-for-one, and the elasticity approaches zero. In the full model in the text, this instead approaches the across-market dispersion parameter $\theta$. We call this the *Market power effect*.

The second piece is the elasticity of the value with respect to the price. This depends on the marginal value of a dollar when purchasing good $j$, $\lambda_j^i$, times the expenditure on good $j$, $p_j x_j^i$. When $\sigma > 1$, this term is decreasing in consumption. We call this the *Consumer heterogeneity effect*.[23]

We can understand how the consumer heterogeneity effect varies with consumption through the language of income and substitution effects. By way of analogy, consider a model of labor supply, where an individual facing wage $w$ chooses hours $h$ to maximize $u(wh) - v(h)$. The first order condition is $u'(wh)w = v'(h)$. Changing $w$ has a substitution effect: holding $u'(c)$ constant, a higher $w$ increases labor supply. Changing $w$ has an income effect: holding $h$ constant, a higher $w$ increases consumption and reduces the marginal value of additional labor. Following a wage increase, if increasing hours increases consumption less than the marginal utility of consumption falls, then optimal behavior is to cut hours. With CRRA utility over consumption and a constant Frisch this becomes $w^{-(\sigma-1)} = h^{1/\varphi+\sigma}$, and hence $\sigma > 1$ is sufficient for a wage increase to lead to an hours reduction. This is similar to our above condition. In our case,

[23]The economics here is general and extends to models of unit demand. Suppose that the household obtained utility from some 'outside good' $c^i$, consumed one unit of $j$ and hence had a budget constraint $c^i + p_j = y^i$. In this case the market power effect is the same, what changes is that $\partial \log v_j^i / \partial \log p_j = \lambda^i p_j$. Then from the first order condition for $c^i$, $\lambda^i p_j = u'(c_j^i) p_j = (y^i - p_j)^{-\sigma} p_j$. This is *decreasing* in income for any $\sigma > 0$, and hence does not rely on $\sigma > 1$.

however, the question the consumer asks is: *Faced with a lower price variety that I like less, is the increase in consumption from swapping to that variety larger than the decline in the marginal utility of consumption*.

We can make this intuition clear by considering the decision of an individual. Suppose $\psi_1 = \psi_2$, $p_1 = p_2$, and that for some individual $\zeta_1^i > \zeta_2^i$. This consumer buys Good 1. Now suppose the price of Good 2 falls. In percentage terms—which is what is required of the elasticity of demand which is central to the firm's pricing decisions—the consumer asks *How much lower does the price of Good 2 have to be for me to swap to the lower priced variety that I like less?* The consumer swaps to Good 2 if:

$$\begin{aligned} u(x_2^i) + \zeta_2^i &> u(x_1^i) + \zeta_1^i \\ u(x_2^i) - u(x_1^i) &> \zeta_1^i - \zeta_2^i. \end{aligned}$$

Working with the left side of this inequality, we can approximate $u(x_1^i)$ around $u(x_2^i)$:

$$-u'(x_2^i)(x_1^i - x_2^i) = -u'(x_2^i)x_2^i\left(\frac{x_1^i - x_2^i}{x_2^i}\right) = -x_2^{i\,-(\sigma-1)}\left(\frac{y^i/p_1 - y^i/p_2}{y^i/p_2}\right) = x_2^{i\,-(\sigma-1)}\log\left(\frac{p_1}{p_2}\right).$$

Now let's compare individuals of type $L$ and type $H$ for which $\zeta_1^i - \zeta_2^i$ is the same. This expression says that the price discount on Good 2, will have to be larger for the high income household to consider swapping from Good 1 to Good 2. This is precisely a *lower elasticity of demand on the extensive margin*. For the high income consumer, the gain in consumption must be a lot larger—obtained via a lot larger price discount—to offset the fall in the marginal utility of consumption.[24]

This simple theory, therefore, generates a declining elasticity of demand by income, which is parameterized only by $\sigma$. This implies that (a) we have not added additional parameters to the consumption savings model in which $\sigma$ usually appears, (b) $\sigma$ may be calibrated to match the relationship between demand elasticities and income.

**Sorting.** The two firm, two household-type example makes statements about sorting simple as we can compute measures of sorting in closed form. Using the standard definition, sorting of type $H$ households to Good 1 is log-supermodular (and hence supermodular) if $\rho_1^H/\rho_2^H > \rho_1^L/\rho_2^L$.

[24] We can extend this approximation to derive the choice probability itself. Suppose that $\zeta_2^i = 0$. The above gives an expression for the threshold utility $\zeta_1^{i*}$, such that $i$ swaps to Good 2 if $\zeta_1^i < \zeta_1^{i*}$. This is given by $\zeta_1^{i*} = x_2^{i\,-(\sigma-1)}\Big[\log p_1 - \log p_2\Big]$. The choice probability is then simply $\rho_2^i = G(\zeta_1^{i*})$, and the elasticity of this with respect to $p_2$ is:

$$-\frac{\partial \log \rho_1^i}{\partial \log p_2} = \frac{g(\zeta_1^{i*})}{G(\zeta_1^{i*})}x_2^{i\,-(\sigma-1)}.$$

In the case of the Gumbel distribution the first term simplifies to $\eta$, and we obtain the expression derived above.

This is satisfied if the following is positive:

$$\Upsilon = \log\left(\frac{\rho_1^H/\rho_2^H}{\rho_1^L/\rho_2^L}\right) > 0. \tag{B7}$$

Since Good 1 is higher quality, we would expect $\rho_1^i/\rho_2^i > 1$ for both household types. The measure $\Upsilon$ is positive if the gradient of these choices is steeper for higher income households. Given the choice probabilities, this becomes a matter of comparing values:

$$\frac{\Upsilon}{\eta} = \left(V_1^H - V_2^H\right) - \left(V_1^L - V_2^L\right) = \left(V_1^H - V_1^L\right) - \left(V_2^H - V_1^L\right). \tag{B8}$$

Notice that quality terms do not *directly* enter this comparison. Quality shifts demand equally for both types and hence cancel when comparing within-good, across-households, as $\Upsilon$ requires. Recall that $V_j^i = V(y^i, p_j) = \max_{c_j^i} u(c_j^i)$ subject to $p_j c_j^i = y^i$. We can simplify the above expression by approximating $V_j^H$ around $V_j^L$:

$$\frac{\Upsilon}{\eta} = \left.\frac{\partial V(y,p_1)}{\partial y}\right|_{y=y^L}\left(y^H - y^L\right) - \left.\frac{\partial V(y,p_2)}{\partial y}\right|_{y=y^L}\left(y^H - y^L\right) = \left(\lambda_1^L - \lambda_2^L\right)\left(y^H - y^L\right). \tag{B9}$$

This says that sorting is positive if, holding consumer type fixed, the marginal value of a dollar is higher when consuming Good 1 relative to Good 2. Using the first order condition for consumption, $u'(x_j^i) = \lambda_j^i p_j$, along with the budget constraint, $p_j x_j^i = y^i$:

$$\frac{\Upsilon}{\eta} = \left(x_1^{L\,-(\sigma-1)} - x_2^{L\,-(\sigma-1)}\right)\left(\frac{y^H - y^L}{y^L}\right) = \left(p_1^{\sigma-1} - p_2^{\sigma-1}\right)\log\left(\frac{y^H}{y^L}\right)y^{L\,-(\sigma-1)}. \tag{B10}$$

Hence, if $p_1 > p_2$, then sorting will be log-supermodular, and so also supermodular. Moreover, the larger the price differences across the goods, the more the marginal value of wealth is changing across possible purchases, and hence the steeper the sorting. This relationship informs our use of sorting statistics to pin down the decreasing returns to scale parameter $\alpha$. Conditional on markups, a lower $\alpha$ increases price dispersion, which increases sorting.

The condition for positive sorting and a declining elasticity of demand in income are clearly linked. Intuitively, as we move up the price distribution from the lowest price to the highest price good, those with the highest elasticities of demand drop off first. This leaves only the least elastic consuming the high price good, which, under $\sigma > 1$ are the rich consumers. More formally, we can see this by constructing $\Upsilon$ via an alternative approach used in the text where

we apply the Fundamental Theorem of Calculus. Let $\widehat{p}_j = \log p_j$, then

$$\begin{aligned}
\Upsilon &= \left[\log\rho\Big(y^H,\widehat{p}_1\Big) - \log\rho\Big(y^H,\widehat{p}_2\Big)\right] - \left[\log\rho\Big(y^L,\widehat{p}_1\Big) - \log\rho\Big(y^L,\widehat{p}_2\Big)\right] \\
\Upsilon &= \int_{\widehat{p}_2}^{\widehat{p}_1} \frac{\partial \log\rho(y^H,\widehat{p})}{\partial\widehat{p}} d\widehat{p} - \int_{\widehat{p}_2}^{\widehat{p}_1} \frac{\partial \log\rho(y^L,\widehat{p})}{\partial\widehat{p}} d\widehat{p} \\
\Upsilon &= \int_{\widehat{p}_2}^{\widehat{p}_1} \left\langle -\frac{\partial \log\rho(y^L,\widehat{p})}{\partial\widehat{p}} \right\rangle - \left\langle \frac{\partial \log\rho(y^H,\widehat{p})}{\partial\widehat{p}} \right\rangle d\widehat{p}.
\end{aligned}$$

This says that if at all prices, the demand elasticity of the low income household is more than that of the high income household, then sorting will be positive. As we have shown, this is the case under $\sigma > 1$. Hence our theory delivers demand elasticities that are declining in income, that then in turn imply positive sorting of rich versus poor households across the price distribution. Quality does not enter either of these conditions directly, but nonetheless shapes prices and sorting.

Lastly, note that in this simple model sorting in customer bases carries over to sorting in expenditures. Expenditure of group $i$ on good $j$ is $Exp^i_j = \rho^i_j p_j x^i_j = \rho^i_j y^i$. If we apply the same sorting measure to expenditure, then in $\Upsilon$ for expenditure the $y^H$ and $y^L$ terms cancel, leaving exactly the same expressions as above. In the data, it may be the case that expenditures by different groups of households are more easily measured than quantities.

**Prices.** Now that we understand demand elasticities and sorting, we can answer the question: *Which firm has the higher price?*. Price is a markup over marginal cost, and with identical marginal costs the firm with the highest price must be the one with the highest markup and hence the lowest elasticity of demand. Suppose that $\sigma > 1$, then the high quality firm has a higher price: $p_1 > p_2$. A constructive way to see this is as follows. Suppose, instead, that $p_1 = p_2 = \mu_0 W$, for some $\mu_0 > 1$. Constructing consumer demand for both households $\rho^i_1 > \rho^i_2$, since prices are equal but $\psi_1 > \psi_2$. The *Consumer heterogeneity* component of the demand elasticity $\lambda^i_j p_j x^i_j$ will be equal within households across goods, since $p_1 = p_2$ and hence $x^i_1 = x^i_2$. Via the *Market power* component of the demand elasticity $\eta(1-\rho^i_1)$, Good 1 will have less elastic demand due to a larger market share. This would lead Firm 1 to wish to deviate to a higher markup, and Firm 2 to deviate to a lower markup. Suppose Firm 1 raises its price by $\Delta$, and Firm 2 lowers their price by a small amount $\Delta$. This time around, Firm 1 still has a higher market share, but given its higher price, it will also disproportionately sell to richer, less elastic customers, while Firm 2 sells to more elastic, poorer customers. This *Consumer heterogeneity* effect leads to a further increase in desired prices at Firm 1.

This exercise makes clear that high quality firms will have higher prices both due to selling to less elastic customers and due to having higher market share. While it may be the case that at

equilibrium prices markups are mainly determined by consumer heterogeneity it is worth noting that the market power effect is nonetheless necessary. Market power effects give the initial impetus to price differences which are then magnified by sorting and consumer heterogeneity. With decreasing returns to scale, the market power channel is no longer necessary, as under $p_1 = p_2$, the higher quantity sold of Good 1 would increase its marginal cost, which would alone lead to an increase in price and sorting.

**Relationship with key empirical facts.** The two firm, two household type example under $\sigma > 1$ delivers the following correlations, consistent with recent empirical evidence:

1. Richer households are have lower demand elasticities - This requires $\sigma > 1$, and gives us a way to calibrate the value of $\sigma$ by matching just how steep this negative relationship is
2. Large firms have higher markups - This is due to both market power and consumer heterogeneity effects
3. Large firms are bigger primarily due to higher quality, and are slightly smaller due to higher marginal cost - We have assumed only differences in quality. When we additionally assume decreasing returns in the quantitative model, then high quality firms also have higher marginal costs which is a force toward relatively higher prices which choke off some demand.
4. Large firms are bigger primarily due to higher customers, not sales per customer - In the theory, the intensive margin elasticity of demand is equal to one. Hence, if markups are around 1.05 to 1.30, then the overall elasticity of demand must be between 4.3 and 21, hence the elasticity of demand on the extensive margin dominates, making firms much bigger or smaller on this dimension rather than the intensive margin.
5. Richer households buy from firms that are on average larger than the firms that poorer households buy from - This is reflected in the positive sorting, driven by the heterogeneity in demand elasticities
6. Richer households pay higher markups - This is true because of sorting plus the fact that high quality, large firms have higher markups

Note that while each of these is the key point from a different empirical paper, pairs of these imply the others. This should not be surprising since they co-exist in similar sets of data.

A key feature of the theory is that these results operate through the first point: richer households have lower demand elasticities. This also operates through a single parameter: $\sigma$. Hence in turning this simple model into a quantitative theory we can use $\sigma$ to match the first item, and then check how the theory does with matching the remaining empirical regularities. An important point of this approach is that it does not require making households heterogeneous in their preferences for different goods.

## C. "Shopping Cart" Model — Purchase All Goods

In this appendix we derive our main formulas for demand elasticities in a model in which the household buys one variety $j \in \{1, \ldots, J\}$ of every good $m \in [0, 1]$ each period. This is in contrast to the model in the main text, households choose a single good $m \in [0, 1]$, and a single variety of that good $j \in \{1, \ldots, J\}$ to consume each period.

**Setup.** There are a continuum of goods indexed $m \in [0, 1]$. For each good there are $J$ varieties, that differ in quality $\psi_{jm}$. Each period, individuals draw their idiosyncratic utility $\zeta_{jm}^i$ according to a Type 1 Extreme Value distribution with parameter $\eta$. Denote the joint distribution $F(\boldsymbol{\zeta}^i)$. We assume the household does not value the outside, competitive good.

**Bellman equation.** Let $V(a, e)$ be the expected present discounted value of utility of an individual that has entered the period with assets $a$, productivity $e$, **but is yet to draw** its vector of shocks $\boldsymbol{\zeta}^i$:

$$V(a,e) = \int \left\{ \max_{\{\iota_{jm}(\boldsymbol{\xi}), x_{jm}(\boldsymbol{\xi}), a'(\boldsymbol{\xi})\}} \left\langle \int_m \sum_{j=1}^{J} \iota_{jm}(\boldsymbol{\xi}) \left[ u\left(x_{jm}(\boldsymbol{\xi})\right) + \psi_{jm} + \zeta_{jm} \right] dm \right\rangle + \beta \mathbb{E}_{e'} V\left(a'(\boldsymbol{\zeta}), e'\right) \right\} dF(\boldsymbol{\zeta})$$

where for each draw of $\boldsymbol{\zeta}$, the following budget constraint must hold:

$$\underbrace{\int_m \sum_{j=1}^{J} \iota_{jm}(\boldsymbol{\zeta})\, p_{jm} x_{jm}(\boldsymbol{\zeta})\; dm}_{\text{Expenditure: } y(\boldsymbol{\zeta})} + a'(\boldsymbol{\zeta}) = We + Ra,$$

and every vector of preferences $\boldsymbol{\zeta}^i$, the household uses the indicator $\iota_{jm}(\boldsymbol{\zeta}) \in \{0, 1\}$ to choose a single good to consume. To recap, for every vector of preferences $\boldsymbol{\zeta}$, the household fills their shopping cart with one variety of each of the many goods on offer, chooses the amount to consume of each variety, and the amount to save $a'(\boldsymbol{\zeta})$.

The key issue in the problem above is that the budget constraints are varying shock by shock, for every $\boldsymbol{\zeta}$. One way to deal with this is to make the assumption that there is two stage budgeting. That is the household chooses total expenditure $y$ *before* the realization of preference shocks, and then chooses how to allocate that spending across varieties of goods in a second stage.

One set of assumptions that leads to allocations which are equivalent to two stage budgeting are if goods are symmetric in their distribution of quality across markets $m$. Then they will also be symmetric in their distribution of prices. This then implies that a household's overall expenditure $y^i(\boldsymbol{\xi}^i)$ is independent of the particular draw of preference shocks. On some shelves it like the expensive varieties, on some shelves it likes the cheaper varieties. Given that

there are a large number of shelves, these differences wash out, and overall expenditure will be equal regardless of the shock. Finally, this implies that from the budget constraint, $a'(\boldsymbol{\zeta}^i)$ is also independent of the exact vector of shocks. This then implies that the continuation value is independent of today's draw of preferences.

These observations allow us to write the Bellman equation as:

$$V(a,e) = \max_{a',y} \int \left\{ \max_{\{\iota_{jm}(\boldsymbol{\zeta}),x_{jm}(\boldsymbol{\zeta})\}} \left\langle \int_m \sum_{j=1}^{J} \iota_{jm}(\boldsymbol{\zeta}) \left[ u\left(x_{jm}(\boldsymbol{\zeta})\right) + \psi_{jm} + \zeta_{jm} \right] dm \right\rangle \right\} dF(\boldsymbol{\zeta}) + \beta \mathbb{E}_{e'} V(a',e'),$$

where the following budget constraint must hold:

$$y + a' = We + Ra,$$

and for each draw of $\boldsymbol{\zeta}^i$:

$$\int_m \sum_{j=1}^{J} \iota_{jm}(\boldsymbol{\zeta})\, p_{jm} c_{jm}(\boldsymbol{\zeta})\; dm = y.$$

A critical observation here is that the inner decision problem over varieties is identical to that of the planning problem in Mongey and Waugh (2024). In that setting there is one market with finite number of varieties and a continuum of individuals indexed by $m \in [0,1]$ with idiosyncratic preferences for each good. The planner chooses for each person $m$, which good they consume and how much of it, given some total resources and preference shocks $\zeta_j^m$ of each person for each good.

Here we have a single individual taking the role of the planner. Given some total resources, and preference shocks $\zeta_{jm}$ for each good $m \in [0,1]$ across varieties $j$, the individual chooses for each good, which variety they consume and how much of it. We approach solving this problem in the same way we solve the planning problem in Mongey and Waugh (2024).

### C.1. Optimality and Elasticities

Below, we discuss the implications of this model and how this model has almost identical implications for heterogeneity in the price elasticity of demand across households and sorting as the baseline model. As in the main body of the text, we suppress $(a,e)$ as arguments and use superscript $i$ to capture household heterogeneity.

**Intensive Margin.** The first order condition for consumption of variety $jm$, conditional on

consuming it is:

$$u'(x^i_{jm}(\boldsymbol{\zeta}^i)) = \lambda^i p_{jm} \quad , \quad (x^i_{jm})^{-\sigma} = \lambda^i p_{jm} \quad , \quad \varepsilon^{i,x}_{jm} = -\frac{\partial \log x^i_{jm}}{\partial \log p_{jm}} = \frac{1}{\sigma}.$$

This delivers two results. First, note that the Lagrange multiplier $\lambda^i$ is independent of preference shocks. This follows from assumptions leading to two stage budgeting which implies that, with many goods, the marginal value of a dollar is independent of the exact shocks drawn. This result, plus the additivity of preference shocks implies that $x^i_{jm}(\boldsymbol{\zeta}^i)$ is independent of the preference shock.

This also implies that the intensive margin elasticity of demand is equal to $1/\sigma$ for all goods. Recall our discussion of the intensive margin elasticity in the main text. We noted that $\varepsilon^{i,x}_{jm} \in [1/\sigma, 1]$. The value of one occurs if the individual cannot move resources around and is hand-to-mouth. The value of $1/\sigma$ occurs when the individual can liquidate savings in such a way as to keep their multiplier constant. In the shopping cart model, if the household really likes variety $jm$, but it has a high price, the household can move resources away from the infinitely many other goods toward variety $jm$, keeping $\lambda^i$ constant.

**The Extensive Margin.** Following the approach in Mongey and Waugh (2024) we use a variational inequality approach to solving for the optimal choice rule $\iota^i_{jm}(\boldsymbol{\zeta}^i)$. Conditional on a vector of preferences $\boldsymbol{\zeta}^i$ and expenditure $y^i$, consider setting up the Lagrangean:

$$\mathcal{L}(\boldsymbol{\zeta}^i) = \int_m \sum_{j=1}^J \iota^i_{jm}\left(\boldsymbol{\zeta}^i\right) \left[u\left(x^i_{jm}\left(\boldsymbol{\zeta}^i\right)\right) + \psi_{jm} + \zeta^i_{jm}\right] dg + \lambda^i \left[y^i - \int_m \sum_{j=1}^J \iota^i_{jm}\left(\boldsymbol{\zeta}^i\right) p_{jm} x^i_{jm}\left(\boldsymbol{\zeta}^i\right)\, dm\right]$$

The optimal choice rule for $x^i_{jg}(\boldsymbol{\zeta}^i)$ is therefore:

$$\iota^i_{jm}(\boldsymbol{\zeta}^i) = \begin{cases} 1 & \text{if} \quad u\left(x^i_{jm}\left(\boldsymbol{\zeta}^i\right)\right) + \psi_{jm} + \zeta^i_{jg} - \lambda^i p_{jm} x^i_{jm}(\boldsymbol{\zeta}^i) \geq \max\limits_{j' \in J} \left\{u\left(x^i_{j'm}\left(\boldsymbol{\zeta}^i\right)\right) + \psi_{j'm} + \zeta^i_{j'm} - \lambda^i p_{j'm} x^i_{j'm}(\boldsymbol{\zeta}^i)\right\} \\ 0 & \text{otherwise} \end{cases}$$

This looks like a standard discrete choice optimality condition with added terms that subtract off the marginal cost associated with choosing $jm$ via removing resources available for other purchases, these are the marginal value of extra resources $\lambda^i$ times expenditure on $jm$. Using the first order condition on the intensive margin and simplifying:

$$\iota^i_{jm}(\boldsymbol{\zeta}^i) = \begin{cases} 1 & \text{if} \quad u\left(x^i_{jm}\right) + \psi_{jm} + \zeta^i_{jm} - \lambda^i p_{jm} x^i_{jm} \geq \max\limits_{j' \in J} \left\{u\left(x^i_{j'm}\right) + \psi_{j'm} + \zeta^i_{j'm} - \lambda^i p_{jm} x^i_{jm}\right\} \\ 0 & \text{otherwise} \end{cases}$$

Then the probability that individual $i$ purchases good $jm$ is

$$\rho^i_{j|m} = \int \iota^i_{jm}(\boldsymbol{\zeta}^i) dF(\boldsymbol{\zeta}^i) = \frac{\exp\left\{\eta\left[u(x^i_{jm}) + \psi_{jm} - \lambda^i p_{jm} x^i_{jm}\right]\right\}}{\sum_{j'=1}^{J} \exp\left\{\eta\left[u(x^i_{j'm}) + \psi_{j'm} - \lambda^i p_{j'm} x^i_{j'm}\right]\right\}}$$

We can then compute the extensive margin elasticity, i.e., how the choice probability changes respect to price.

$$\begin{aligned}
\varepsilon^{\rho,i}_{j|m} = -\frac{\partial \log \rho^i_{j|m}}{\partial \log p_{jm}} &= \eta(1 - \rho^i_{j|m})\lambda^i x^i_{jm} p_{jm} \\
&= \eta(1 - \rho^i_{j|m}) u'(x^i_{jm}) x^i_{jm} \\
&= \eta(1 - \rho^i_{j|m}) x^{i\,-(\sigma-1)}_{jm} \quad \text{when u is CRRA}
\end{aligned}$$

This is exactly the same expression for the extensive margin choice probability obtained in the main text, with one difference. That difference is that the first term is $\eta(1 - \rho^i_{j|m})$ instead of $[\theta\rho^i_{j|m} + \eta(1 - \rho^i_{j|m})]$. In the shopping cart model, the household must buy one variety of every good, and hence there is a zero elasticity of substitution *across goods*. While the shopping cart model cannot be generalized to the full model, the full model can replicate the shopping cart model when $\theta = 0$. In this case preferences are so dispersed across goods that there is effectively a zero elasticity of substitution between goods. Quantitatively, the fact that the calibrated value of $\theta$ is close to zero (Table 1), implies that the quantitative implications of the shopping cart model and one-good-one-variety model are very similar.

**What is the same?** When households can consume all goods, what still comes out of this model is heterogeneous price elasticities and sorting like in our one-good-one-variety model in the body of the text. In this model, different individuals choose different levels of expenditure. Conditional on expenditure, individuals fill their shopping cart as follows. Higher income households have a lower marginal value of a dollar, $\lambda^i$, and are less price elastic, good-by-good they have a higher probability of buying the most expensive variety. Their shopping cart tilts more toward high quality varieties, but they also buy low quality varieties as well. Lower income households have a higher $\lambda^i$ and are more price elastic, good-by-good they have a lower probability of buying the most expensive variety. Their shopping cart tilts more toward low quality varieties, but they also buy some high quality varieties as well. Incomplete markets are still crucial to shaping choice probabilities via the link between the marginal value of a dollar to the household, and the $\lambda^i$ that appears in the individual's choice probability for any variety.

**What is different?** There are two key differences between the shopping cart model and the

one-good-one-variety model. In the shopping cart model, the intensive margin elasticity is constant for all consumers and all goods. However, since this is between $[1/\sigma, 1] \approx [0.5, 1]$ in the main text—while elasticities of demand overall are governed mostly by the extensive margin which is in the range of 2 to 15—then this difference is of little practical importance. Second, the extensive margin elasticity loses the $\theta\rho_j^i$ term, however, again, this is approximately zero in the calibrate one-good-one-variety model.

The benefit of the shopping cart model is that it retains the three important features of our model (i) heterogeneous price elasticities (ii) sorting, and (iii) a meaningful extensive margin of demand consistent with (Argente et al. (2021); Afrouzi et al. (2023); Einav et al. (2021)). And yet this extension preserves the intuitive feature that households consume many goods simultaneously. Thus, it forms a data generating process that can be mapped directly to microdata that itemizes expenditure and prices of all goods in shopping trips. The drawback of the shopping cart model is that, pedagogically, it is cumbersome in terms of presentation relative to the simpler one-good-one-variety model. Thus, we use the later in the main text and emphasize how it is robust to the inclusion of many goods with reference to the shopping cart model in this Appendix.

## D. Continuous Time Model

In this section, we consider a different environment in which the time interval is infinitesimal. In this setting, there are still heterogeneous price elasticities and sorting as in the body of the text.

**Setup** We consider the exact same environment as described in the text with the only difference being that productivity shocks arrive at some Poisson rate. Specifically, start with our discrete time model with time interval $\Delta$, and then take the limit as $\Delta \to 0$. For simplicity, we assume individual productivity states are discrete with $N_e = 2$ states. The discount rate is $\beta(\Delta) = e^{-\rho\Delta}$, which is approximately equal to $\beta(\Delta) \approx 1 - \rho\Delta$. As in the text, the individual draws a new vector $\boldsymbol{\zeta}^i$ from the nested Gumbel distribution $F(\boldsymbol{\zeta})$ every time interval $\Delta$. The only difference is that we now assume that productivity shocks do not arrive every period. The probability of staying at the same productivity state depends on $e$, and is $k(\Delta, e) = e^{-\kappa(e)\Delta} \approx 1 - \kappa(e)\Delta$.

**Derivation of Hamilton-Jacobi-Bellman (HJB) equation.** Our presentation follows the derivation of the HJB equation in Appendix B of Achdou et al. (2021).

Given the environment, the discrete time Bellman equation is:

$$
\begin{aligned}
V(a_t, e_t, \boldsymbol{\zeta}) = \max_{\iota_{jmt}, x_{jmt}} \int_m \sum_{j=1}^{J} \iota_{jmt} \Bigg\langle & \left[u(x_{jmt}) + \psi_{jm} + \zeta_{jmt}\right]\Delta \\
& + \beta(\Delta)\left[k(\Delta, e_t)\,\mathbb{E}\left[V(a_{t+\Delta}, e_t, \boldsymbol{\zeta}')\right] + (1 - k(\Delta, e_t))\,\mathbb{E}\left[V(a_{t+\Delta}, e_t', \boldsymbol{\zeta}')\right]\right] \Bigg\rangle dm, \\
a_{t+\Delta} &= \Delta(We_t + ra_t - p_{jmt}x_{jmt}) + a_t, \\
a_{t+\Delta} &\geq \underline{a}, \\
\iota_{jmt} &= \begin{cases} 1 & \text{for only one good-variety } jm \\ 0 & \text{for all other good-varieties .} \end{cases}
\end{aligned}
$$

In each period, the household chooses $\iota_{jmt}$, which is an indicator equaling one for only one good-variety $jm$, and how much to consume of this variety. Given this, next period assets $a_{t+\Delta}$ are determined via the budget constraint. Note that the entire term $\langle \cdot \rangle$ depends on the choice of good $\iota_{jgt}$, including continuation values via $a_{t+\Delta}$. In the continuation values, the expectation is only with respect to the draw of $\boldsymbol{\zeta}'$.

We simplify this in a number of steps to obtain the HJB equation:

1. We first replace $\beta(\Delta)$ and $k(\Delta, e_t)$ with the above approximations:

$$V(a_t, e_t, \boldsymbol{\zeta}) = \max_{\iota_{jmt}, x_{jmt}} \int_m \sum_{j=1}^{J} \iota_{jmt} \Bigg\langle [u(x_{jmt}) + \psi_{jm} + \zeta_{jmt}]\Delta + \\ + (1-\rho\Delta)\left[(1-\kappa(e)\Delta)\,\mathbb{E}\left[V(a_{t+\Delta}, e_t, \boldsymbol{\zeta}')\right] + \kappa(e)\Delta\mathbb{E}\left[V(a_{t+\Delta}, e_t', \boldsymbol{\zeta}')\right]\right]\Bigg\rangle dm.$$

2. Subtracting $(1-\rho\Delta)V(a_t, e_t, \boldsymbol{\zeta})$ from both sides, we obtain the following. Note that since $\langle\cdot\rangle = 0$ for all $jm$ apart from the one chosen, we are able to bring the last term inside $\langle\cdot\rangle$, and then re-arrange expressions

$$\rho\Delta V(a_t, e_t, \boldsymbol{\zeta}) = \max_{\iota_{jmt}, x_{jmt}} \int_m \sum_{j=1}^{J} \iota_{jmt} \Bigg\langle [u(x_{jmt}) + \psi_{jm} + \zeta_{jmt}]\Delta \\ + (1-\rho\Delta)\Bigg[(1-\kappa(e_t)\Delta)\left(\mathbb{E}\left[V(a_{t+\Delta}, e_t, \boldsymbol{\zeta}')\right] - V(a_t, e_t, \boldsymbol{\zeta})\right) \\ + \kappa(e_t)\Delta\left(\mathbb{E}\left[V(a_{t+\Delta}, e_t', \boldsymbol{\zeta}')\right] - V(a_t, e_t, \boldsymbol{\zeta})\right)\Bigg]\Bigg\rangle dm.$$

3. Dividing through by $\Delta$:

$$\rho V(a_t, e_t, \boldsymbol{\zeta}) = \max_{\iota_{jmt}, x_{jmt}} \int_m \sum_{j=1}^{J} \iota_{jmt} \Bigg\langle [u(x_{jmt}) + \psi_{jm} + \zeta_{jmt}] + \\ + (1-\rho\Delta)\Bigg[(1-\kappa(e_t)\Delta)\left(\frac{\mathbb{E}\left[V(a_{t+\Delta}, e_t, \boldsymbol{\zeta}')\right] - V(a_t, e_t, \boldsymbol{\zeta})}{\Delta}\right) \\ + \kappa(e_t)\left(\mathbb{E}\left[V(a_{t+\Delta}, e_t', \boldsymbol{\zeta}')\right] - V(a_t, e_t, \boldsymbol{\zeta})\right)\Bigg]\Bigg\rangle dm.$$

4. Taking limits as $\Delta \to 0$:

$$\rho V(a_t, e_t, \boldsymbol{\zeta}) = \max_{\iota_{jmt}, x_{jmt}} \int_m \sum_{j=1}^{J} \iota_{jmt} \Bigg\langle [u(x_{jmt}) + \psi_{jm} + \zeta_{jmt}] \\ + \Bigg[\lim_{\Delta\to 0} \frac{\mathbb{E}\left[V(a_{t+\Delta}, e_t, \boldsymbol{\zeta}')\right] - V(a_t, e_t, \boldsymbol{\zeta})}{\Delta} \\ + \kappa(e_t)\left(\lim_{\Delta\to 0}\mathbb{E}\left[V(a_{t+\Delta}, e_t', \boldsymbol{\zeta}')\right] - V(a_t, e_t, \boldsymbol{\zeta})\right)\Bigg]\Bigg\rangle dm.$$

5. Note that we can write the first limit as the expectation of the limit,

$$\lim_{\Delta\to 0}\frac{\mathbb{E}\left[V\left(a_{t+\Delta},e_t,\boldsymbol{\zeta}'\right)\right]-V\left(a_t,e_t,\boldsymbol{\zeta}\right)}{\Delta}=\mathbb{E}\left[\lim_{\Delta\to 0}\frac{V\left(a_{t+\Delta},e_t,\boldsymbol{\zeta}'\right)-V\left(a_t,e_t,\boldsymbol{\zeta}\right)}{\Delta}\right],$$

which implements the chain rule:

$$\lim_{\Delta\to 0}\frac{V\left(\Delta\left[We_t+ra_t-p_{jmt}x_{jmt}\right]+a_t,e_t,\boldsymbol{\zeta}'\right)-V\left(a_t,e_t,\boldsymbol{\zeta}\right)}{\Delta}=\frac{\partial V\left(a_t,e_t,\boldsymbol{\zeta}'\right)}{\partial a_t}\left(We_t+ra_t-p_{jmt}x_{jmt}\right)$$

6. From the budget constraint, the second limit is equal to $a_t$:

$$\lim_{\Delta\to 0}a_{t+\Delta}=\lim_{\Delta\to 0}\left\{\Delta We_t+\left(1+\Delta r\right)a_t-\Delta p_{jmt}x_{jmt}\right\}=a_t.$$

7. Substituting these results back in, and noting that the expectation with respect to $\boldsymbol{\zeta}'$ does not need to apply to the terms in $(\cdot)$ in the above limit:

$$\begin{aligned}\rho V\left(a_t,e_t,\boldsymbol{\zeta}\right)=\max_{\iota_{jmt},x_{jmt}}\int_m\sum_{j=1}^{J}\iota_{jmt}\Bigg\langle&\left[u\left(x_{jmt}\right)+\psi_{jm}+\zeta_{jmt}\right]\\&+\mathbb{E}\left[V_a\left(a_t,e_t,\boldsymbol{\zeta}'\right)\right]\left(We_t+ra_t-p_{jmt}x_{jmt}\right)+\kappa\left(e_t\right)\left(\mathbb{E}\left[V\left(a_t,e_t',\boldsymbol{\zeta}'\right)\right]-V\left(a_t,e_t,\boldsymbol{\zeta}\right)\right)\Bigg\rangle dm.\end{aligned}$$

8. We can remove terms from the integral and sum $\int_m\sum_j\iota_{jmt}\left\langle\cdot\right\rangle dm$ that do not depend on the choice of good-variety, this includes the continuation values associated with productivity shocks.

$$\begin{aligned}\rho V\left(a_t,e_t,\boldsymbol{\zeta}\right)=\max_{\iota_{jmt},x_{jmt}}\int_m\sum_{j=1}^{J}\iota_{jmt}\Bigg\langle&\left[u\left(x_{jmt}\right)+\psi_{jm}+\zeta_{jmt}\right]-\mathbb{E}\left[V_a\left(a_t,e_t,\boldsymbol{\zeta}'\right)\right]p_{jmt}x_{jmt}\Bigg\rangle dm\\&+\mathbb{E}\left[V_a\left(a_t,e_t,\boldsymbol{\zeta}'\right)\right]\left(We_t+ra_t\right)+\kappa\left(e_t\right)\left(\mathbb{E}\left[V\left(a_t,e_t'\boldsymbol{\zeta}'\right)\right]-V\left(a_t,e_t,\boldsymbol{\zeta}\right)\right).\end{aligned}$$

Finally, note this can easily be generalized to the case of $N_e$ income states, with $\pi\left(e,e'\right)$ being the arrival rates of transitions from $e$ to $e'$. We add this in the final step.

Given these steps, the HJB equation is:

$$\begin{aligned}\rho V\left(a_t,e_t,\boldsymbol{\zeta}\right)=\max_{\iota_{jmt},x_{jmt}}\int_m\sum_{j=1}^{J}\iota_{jmt}\Bigg\langle&\left[u\left(x_{jmt}\right)+\psi_{jm}+\zeta_{jmt}-\mathbb{E}\left[V_a\left(a_t,e_t,\boldsymbol{\zeta}'\right)\right]p_{jmt}x_{jmt}\right]\Bigg\rangle dm\\&+\mathbb{E}\left[V_a\left(a_t,e_t,\boldsymbol{\zeta}'\right)\right]\left(We_t+ra_t\right)+\sum_{e'\neq e}\pi\left(e,e'\right)\left(\mathbb{E}\left[V\left(a_t,e',\boldsymbol{\zeta}'\right)\right]-V\left(a_t,e_t,\boldsymbol{\zeta}\right)\right).\end{aligned}$$

To emphasize this again, this is the HJB associated with the model in the main text, under the only alternative assumption that income shocks arrive according to a Poisson process.

### D.1. Optimality and Elasticities

Like in the text, we compute the optimality conditions for intensive and extensive margin and the associated elasticities.

**Intensive margin.** Given the HJB, the optimality condition for intensive margin consumption for the good consumed in positive quantities is as follows.

$$\frac{u'\left(x_{jmt}\left(a_t, e_t, \boldsymbol{\zeta}_t\right)\right)}{p_{jmt}} = \mathbb{E}\left[V_a\left(a_t, e_t, \boldsymbol{\zeta}'\right)\right].$$

Now two more observations imply that optimal quantity $x_{jmt}$ does not depend upon the shock $\boldsymbol{\zeta}$. One the additivity and thus the shock does not appear on the left hand side of the first order condition. Second, the shocks are independent across time so independence of $\boldsymbol{\zeta} \perp \boldsymbol{\zeta}'$, and thus does not affect the right hand side of the first order condition.

Finally, under CRRA utility, the intensive margin elasticity of demand is:

$$\varepsilon^x_{jm}\left(a_t, e_t\right) = -\frac{\partial \log x_{jm}\left(a_t, e_t\right)}{\partial \log p_{jm}} = \frac{1}{\sigma}.$$

This is a common constant for all households for all goods and varieties.

**Extensive margin.** Given the HJB, we can apply a similar variational argument as in Mongey and Waugh (2024) or in the shopping cart model above. One chooses good $jm$ if the term $\langle\cdot\rangle$ is largest

$$\iota_{jmt}\left(a_t, e_t, \boldsymbol{\zeta}_t\right) = \begin{cases} 1 & \text{if} \quad u\left(x_{jmt}\right) + \psi_{jm} + \zeta_{jmt} - \mathbb{E}\left[V_a\left(a_t, e_t, \boldsymbol{\zeta}'\right)\right] p_{jmt} x_{jmt} \quad \geq \\ & \quad \max_{j',m'} \left\{ u\left(x_{j'm't}\right) + \psi_{j'm'} + \zeta_{j'm't} - \mathbb{E}\left[V_a\left(a_t, e_t, \boldsymbol{\zeta}'\right)\right] p_{j'm't} x_{j'm't} \right\} \\ 0 & \text{otherwise} \end{cases}$$

The choice probability for $jm$ of individuals with $(a, e)$ is $\rho_{jm}\left(a, e\right) = \int_m \sum_{j=1}^J \iota_{jm}\left(a, e, \boldsymbol{\zeta}\right) dF\left(\boldsymbol{\zeta}\right)$. Integrating over the preferences of individuals leads to the same structure of choice probabilities as in the main text:

$$\rho_{jm}(a_t, e_t) = \left( \frac{\exp\left\{v_{jm}(a_t, e_t) + \psi_{jm}\right\}}{\exp\left\{\widetilde{v}_m(a_t, e_t)\right\}} \right)^{\eta} \left( \frac{\exp\left\{\widetilde{v}_m(a_t, e_t)\right\}}{\exp\left\{\overline{V}(a_t, e_t)\right\}} \right)^{\theta}$$

Here the value functions inside the $\exp$ functions are the continuous time analogs to those in the main text. The extensive margin demand elasticity is then obtained as in the main text:

$$\varepsilon_{jm}^{\rho}\left(a_t, e_t\right) = -\frac{\partial \log \rho_{jm}\left(a_t, e_t\right)}{\partial \log p_{jmt}} = \left\{\frac{\partial \log \rho\left(a_t, e_t\right)}{\partial \log v\left(a_t, e_t\right)}\right\}\left\{-\frac{\partial \log v\left(a_t, e_t\right)}{\partial \log p_{jmt}}\right\}$$

and then expanding out each of the terms in brackets gives

$$\varepsilon_{jm}^{\rho}\left(a_t, e_t\right) = \left[\theta \rho_{j|m}\left(a_t, e_t\right) + \eta\left(1 - \rho_{j|m}\left(a_t, e_t\right)\right)\right]\left\{-\frac{\partial}{\partial p_{jmt}}\left[u\left(x_{jmt}(a_t, e_t)\right) - \mathbb{E}\left[V_a\left(a_t, e_t, \boldsymbol{\zeta}'\right)\right] p_{jmt} x_{jmt}\left(a_t, e_t\right)\right]\right\} p_{jmt}$$

$$= \left[\theta \rho_{j|g}\left(a_t, e_t\right) + \eta\left(1 - \rho_{j|g}\left(a_t, e_t\right)\right)\right]\left\{-\underbrace{\left[u'\left(x_{jmt}\right) - \mathbb{E}\left[V_a\left(a_t, e_t, \boldsymbol{\zeta}'\right)\right] p_{jmt}\right]}_{\text{Equals zero from } FOC\left(c_{jmt}\right)} \frac{\partial x_{jmt}}{\partial p_{jmt}} + \mathbb{E}\left[V_a\left(a_t, e_t, \boldsymbol{\zeta}'\right)\right] x_{jmt}\left(a_t, e_t\right)\right\} p_{jmt}$$

Then simplifying leaves

$$\varepsilon_{jm}^{\rho}\left(a_t, e_t\right) = \left[\theta \rho_{j|m}\left(a_t, e_t\right) + \eta\left(1 - \rho_{j|m}\left(a_t, e_t\right)\right)\right] \mathbb{E}\left[V_a\left(a_t, e_t, \boldsymbol{\zeta}'\right)\right] x_{jmt}\left(a_t, e_t\right) p_{jm},$$

$$= \left[\theta \rho_{j|m}\left(a_t, e_t\right) + \eta\left(1 - \rho_{j|m}\left(a_t, e_t\right)\right)\right] u'\left(x_{jmt}\left(a_t, e_t\right)\right) x_{jmt}\left(a_t, e_t\right).$$

The last line follows from inserting the first order condition from the quantity choice. This elasticity is exactly as in the main text.

**What is the same?** Essentially everything: heterogeneous price elasticities and sorting which are all fundamentally connected with the marginal value of wealth and the functional relationship is exactly the same as in the body of the text. The one minimal difference is that the intensive margin elasticity of demand is fixed. This difference is minimal because this is the quantitatively less important elasticity as the extensive margin elasticity dominates.

## E. Comparison to quality model of FGH (2011)

Our model has additive separability between quality and consumption. With CRRA and $\sigma > 1$, this delivers (a) price elasticities of demand that are *declining* in consumption and (b) which induce positive sorting of high income households into high price goods. These features plus the firm's technology and pricing strategy imply that high quality goods are high price goods both because of their higher markups and higher marginal cost. That they have higher markups is due to lower elasticities of demand due to more market power, conditional on consumer type, and sorting of low elasticity customers into high price goods.

In this section, we illustrate that quasi-linear utility and unit demand with *multiplicative quality* as in Fajgelbaum et al. (2011, henceforth FGH) delivers (a) the same sorting implications, but (b) has the opposite implication for the relationship between quality and demand elasticities.

In the model of FGH, households consume one unit of the differentiated good (in our notation, $x_j^i = 1$), so consumption $c_j^i$ is of an outside good and determined via the budget constraint $c_j^i = y^i - p_j$. Preferences, and the implied choice probabilities are as follows:

$$V_j^i = \max_{c_j^i} \; u\Big(c_j^i, \phi_j\Big) + \zeta_j^i \quad , \quad u\Big(c_j^i, \phi_j\Big) = \phi_j c_j^i \quad , \quad \zeta_j^i \sim \text{Type 1 EV}(\eta)$$

It is important to notice that consuming a high quality good increases the marginal utility of consuming the outside goods. The implied choice probabilities are:

$$\rho_j^i = \frac{\exp\left(\eta\phi_j(y^i - p_j)\right)}{\sum\limits_{j' \in J} \exp\left(\eta\phi_k(y^i - p_k)\right)}$$

**Elasticities.** Unit demand implies that the elasticity on extensive margin is the only relevant elasticity $\varepsilon_j^i$. As in the main text we can apply the chain rule to make explicit the piece that depends on the distribution of shocks and the piece that depends on utility:

$$\frac{\partial \log \rho_j^i}{\partial \log p_j} = \frac{\partial \log \rho_j^i}{\partial u(c_j^i, \phi_j)} \times \frac{\partial u(c_j^i, \phi_j)}{\partial \log p_j} = \eta\Big(1 - \rho_j^i\Big) \times \lambda_j^i p_j,$$

which is generically the same as in our setting. Then inserting the FGH functional form we have

$$\begin{aligned} \frac{\partial \log \rho_j^i}{\partial \log p_j} &= \eta\Big(1 - \rho_j^i\Big) \times u_c\Big(c_j^i, \phi_j\Big) p_j \\ &= \eta\Big(1 - \rho_j^i\Big) \phi_j p_j. \end{aligned}$$

The elasticity of demand is *increasing* in quality. The intuition is straight-forward: via the budget

constraint, changes in price $p_j$, pass-through one-to-one into consumption $c_j^i$ which has a much larger effect on utility $u(c_j^i, \phi_j)$, when consuming a high quality $\phi_j$ good. The marginal value of a dollar, given by the Lagrange multiplier $\lambda_j^i$, increases in $\phi_j$.

In the atomistic case with $\rho_j^i \to 0$, the elasticity becomes $\varepsilon_j^i = \eta\phi_j p_j$. This is independent of $i$, hence $\varepsilon_j = \eta\phi_j p_j$, exactly as in FGH (their equation 6). Optimal pricing is then a markup over marginal cost:

$$p_j = \frac{\varepsilon_j}{\varepsilon_j - 1} mc_j \quad \Rightarrow \quad p_j = mc_j + \frac{1}{\eta\phi_j}.$$

This is "cost-plus" pricing, where the plus part is smaller for higher quality goods, owing to their less elastic demand. Put simply, high quality firms have *lower* markups.

**Sorting.** First, the model delivers sorting only on quality, not price. To see this consider the case of two goods and two households, with $\phi_1 > \phi_2$, and $y^H > y^L$. By definition, demand is log-supermodular (and hence supermodular) if $\rho_1^H \rho_2^L > \rho_2^H \rho_1^L$. Re-arranging, and taking logs, this is true if $\Upsilon = \log(\rho_1^H/\rho_1^L) - \log(\rho_2^H/\rho_2^L) > 0$.

In the FGM model, the log-supermodularity condition holds regardless of price. Comparing across goods and households, relative prices cancel leaving $\Upsilon = \eta(\phi_1 - \phi_2)(y^H - y^L) > 0$. Hence, regardless of prices, there is positive sorting. In our model, the sufficient condition for sorting can be written without reference to quality, and depends only on price differences: since high income households are less elastic, they sort toward higher priced varieties.

Now recall the facts that we are aiming for: High quality firms are larger, sell at higher prices, charger higher markups. In the FGH setting, If marginal cost $mc_j$ and quality $\phi_j$ are positively correlated, which firms have high prices? From here it is unclear. Holding quality fixed, higher marginal costs lead to higher prices. But holding marginal cost fixed, higher quality leads to lower markups and lower prices. Note that we have already established that sorting is independent of relative prices, hence this question of prices can be handled independently without undoing positive sorting.

What about markups? If high quality firms are to have high prices and high markups, then one needs to assume something additional about $\eta$. Assumption 1 of FGH is that the dispersion of preferences $1/\eta$ increases in quality. If further assumed that $\eta_j = g(\phi_j)$ and the function $g$ is such that $g'(\phi_j) < -1$, and hence decreasing more than one-for-one with as $\phi_j$ increases, then the equilibrium would feature higher markups on high quality goods, which combined with higher marginal cost yields higher prices.

Substantively, the assumption is that each individual has more dispersed preferences for high quality goods. In the main text, a key part of our exercise is to assume a symmetry of households' preferences both (a) they do not systematically differ by income, (b) they do not system-

atically differ over goods, apart from the additive quality term.

# F. Computation & Additional Mathematical Details

## F.1. Computation

A key contribution of the paper is to demonstrate that this economy is a feasible laboratory for quantitative work. The equilibrium consists of a Nash equilibrium in every market, where demand elasticities depend on policy functions of households as well as the stationary distribution of households, and firms are arbitrarily heterogeneous in productivity and quality, $(z_{jm}, \phi_{jm})$. Despite this, the structure of the equilibrium can be exploited to maintain a degree of tractability on par with the standard Bewley model. To see this, fix the bond price $Q$ and wage $W$ and consider solving for the Nash equilibrium in all markets.

- Set $k = 0$. Guess prices at all firms $p_{jm}^{(k)}$.

  1. Set $l = 0$. Guess a continuation value function $\overline{V}^{(k,l)}(a, e)$.

     1a. Solve for $v(a, e, p_n)$ on a grid of points $p_n \in \left[\min\{p_{jm}^{(k)}\}, \max\{p_{jm}^{(k)}\}\right]$.

     $$v(a, e, p_n) = \max_{a' \in [\underline{a}, \infty)} u\Big( \left((1-\tau)We + T + \Pi + a - Qa'\right)/p_n \Big) + \beta \int \overline{V}(a', e') d\mathcal{P}(e, e')$$

     Fit an interpolant to $v(a, e, p_n)$ across $p_n$, and denote this $\widehat{v}(a, e, p)$. Since this problem is so well behaved, small number of points $p_n$ can be used.

     1b. Interpolate $\widehat{v}(a, e, p)$ on guessed prices $p_{jm}^{(k)}$ to construct average values, which gives an update of $\overline{V}^{(k,l+1)}(a, e)$

     $$\widetilde{v}_m(a, e) = \frac{1}{\eta} \log \left[ \sum_{j \in J_m} \phi_{jm} e^{\eta \widehat{v}(a, e, p_{jm})} \right] \quad , \quad \overline{V}^{(k,l+1)}(a, e) = \frac{1}{\theta} \log \left[ \sum_{m \in M} e^{\theta \widetilde{v}(a, e)} \right]$$

     1c. Iterate on $l$ to convergence of $\overline{V}^{(k,l)}(a, e)$

  2. Use the above values to construct choice probabilities $\rho_{jm}^{(k)}(a, e)$, and combine with an interpolant of the asset policy function $\widehat{a}'^{(k)}(a, e, p)$ to solve for the stationary distribution $\Lambda^{(k)}(a, e)$.

  3. Use the stationary distribution, choice probabilities and an interpolant of the consumption policy function $\widehat{c}^{(k)}(a, e, p)$ to compute demand and demand elasticities:

  $$\begin{aligned} x_{jm}^{(k)} &= \int \rho_{jm}^{(k)}(a, e) d\Lambda^{(k)}(a, e) \quad , \\ \varepsilon_{jm}^{(k)} &= \int \left( \frac{\rho_{jm}^{(k)}(a, e)}{x_{jm}^{(k)}} \right) \left[ \eta(1 - \rho_{jm|m}^{(k)}(a, e)) + \theta \rho_{jm|m}^{(k)}(a, e) \right] u'\left( \widehat{c}^{(k)}(a, e, p_{jm}) \right) d\Lambda^{(k)}(a, e) \end{aligned}$$

  4. Update firms' optimal price.

  $$p_{jm}^{(k+1)} = \frac{\varepsilon_{jm}^{(k)}}{\varepsilon_{jm}^{(k)} + 1} \times mc_{jm}^{(k)} \quad , \quad mc_{jm}^{(k)} = \frac{1}{\alpha} \frac{W}{z_{jm}} \left( \frac{x_{jm}^{(k)}}{z_{jm}} \right)^{\frac{1-\alpha}{\alpha}} .$$

- Iterate on $k$ to convergence of $p_{jm}^{(k)}$.

A few key observations yield tractability. First, we only have to solve the choice problem on a

small set of prices $p_n$ in (1a), and can interpolate in (1b). Second, (1c) uses all firms' qualities, but is fast. Combined, these imply that solving the Bellman equation does not take substantially longer than a usual consumption saving problem. Construction of the stationary distribution is standard (2), and the update of $p_{jm}$ is done in closed form (7, 8). Third, in our experience, $p_{jm}^{(k)}$ can be updated along with $Q$ and $W$. Hence the overall equilibrium takes not much longer than a standard Bewley model.

## F.2. Derivation of Auer et al. (2024) coefficient

$$
\begin{aligned}
b_{jm}^{i} &= \frac{p_{jm} c_{jm}^{i} \rho_{jm}^{i}}{\sum_{m} \sum_{j \in m} p_{jm} c_{jm}^{i} \rho_{jm}^{i}} \\
\frac{b_{jm}^{i}}{b_{km}^{i}} &= \frac{p_{jm} c_{jm}^{i} \rho_{jm}^{i}}{p_{km} c_{km}^{i} \rho_{km}^{i}} \\
\log\left(\frac{b_{jm}^{H}}{b_{km}^{H}}\right) &= \log\left(\frac{\rho_{jm}^{H} c_{jm}^{H}}{\rho_{km}^{H} c_{km}^{H}}\right) + \log\left(\frac{p_{jm}}{p_{km}}\right) \\
\rho_{jm}^{H} c_{jm}^{H} &\approx \rho_{km}^{H} c_{km}^{H} + \frac{\partial \rho_{km}^{H} c_{km}^{H}}{\partial p_{km}} \left(p_{jm} - p_{km}\right) \\
\log\left(\frac{\rho_{jm}^{H} c_{jm}^{H}}{\rho_{km}^{H} c_{km}^{H}}\right) &\approx \frac{\partial \rho_{km}^{H} c_{km}^{H}}{\partial p_{km}} \frac{p_{km}}{\rho_{km}^{H} c_{km}^{H}} \log\left(\frac{p_{jm}}{p_{km}}\right) \\
\log\left(\frac{\rho_{jm}^{H} c_{jm}^{H}}{\rho_{km}^{H} c_{km}^{H}}\right) &\approx -\varepsilon_{km}^{H} \log\left(\frac{p_{jm}}{p_{km}}\right) \\
\log\left(\frac{b_{jm}^{H}}{b_{km}^{H}}\right) &= -\varepsilon_{km}^{H} \log\left(\frac{p_{jm}}{p_{km}}\right) + \log\left(\frac{p_{jm}}{p_{km}}\right) \\
\log\left(\frac{b_{jm}^{H}}{b_{km}^{H}}\right) - \log\left(\frac{b_{jm}^{L}}{b_{km}^{L}}\right) &= -\left(\varepsilon_{km}^{H} - \varepsilon_{km}^{L}\right) \log\left(\frac{p_{jm}}{p_{km}}\right) \\
\varepsilon_{km}^{H} &\approx \varepsilon_{km}^{L} + \frac{\partial \varepsilon_{km}^{L}}{\partial e^{L}} \left(e^{H} - e^{L}\right) \\
\varepsilon_{km}^{H} - \varepsilon_{km}^{L} &\approx \varepsilon_{km}^{L} \left(\frac{\partial \varepsilon_{km}^{L}}{\partial c_{km}^{L}} \frac{c_{km}^{L}}{\varepsilon_{km}^{L}}\right) \left(\frac{\partial c_{km}^{L}}{\partial e^{L}} \frac{e^{L}}{c_{km}^{L}}\right) \left(\frac{e^{H} - e^{L}}{e^{L}}\right) \\
\varepsilon_{km}^{H} - \varepsilon_{km}^{L} &\approx \varepsilon_{km}^{L} \left(\frac{\partial \log \varepsilon_{km}^{L}}{\partial \log c_{km}^{L}}\right) \left(\frac{\partial \log c_{km}^{L}}{\partial \log e^{L}}\right) \log\left(\frac{e_{H}}{e_{L}}\right) \\
-\left(\varepsilon_{km}^{H} - \varepsilon_{km}^{L}\right) &\approx -\varepsilon_{km}^{L} \left(\frac{\partial \log \varepsilon_{km}^{L}}{\partial \log c_{km}^{L}}\right) \left(\frac{\partial \log c_{km}^{L}}{\partial \log e^{L}}\right) \log\left(\frac{e_{H}}{e_{L}}\right) \\
\log\left(\frac{b_{jm}^{H}}{b_{km}^{H}}\right) - \log\left(\frac{b_{jm}^{L}}{b_{km}^{L}}\right) &= \varepsilon_{km}^{L} \left(\frac{\partial \log c_{km}^{L}}{\partial \log e^{L}}\right) \left(-\frac{\partial \log \varepsilon_{km}^{L}}{\partial \log c_{km}^{L}}\right) \log\left(\frac{e_{H}}{e_{L}}\right) \log\left(\frac{p_{jm}}{p_{km}}\right) \\
\varepsilon_{km}^{h} &= \left[\rho_{j|m}^{i} \theta + \left(1 - \rho_{j|m}^{i}\right) \eta\right] c_{jm}^{i-(\sigma-1)} \\
\frac{\partial \log \varepsilon_{km}^{L}}{\partial \log c_{km}^{L}} &= -\left(\sigma - 1\right) \\
-\frac{\partial \log \varepsilon_{km}^{L}}{\partial \log c_{km}^{L}} &= \left(\sigma - 1\right) \\
\log\left(\frac{b_{jm}^{H}}{b_{km}^{H}}\right) - \log\left(\frac{b_{jm}^{L}}{b_{km}^{L}}\right) &= \left(\sigma - 1\right) \left(\frac{\partial \log c_{km}^{L}}{\partial \log e^{L}}\right) \log\left(\frac{e_{H}}{e_{L}}\right) \log\left(\frac{p_{jm}}{p_{km}}\right)
\end{aligned}
$$

## F.3. Replication of Auer et al. (2024) regression

We simulate the experiment in ABLV and choose $\sigma$ to match the regression coefficient $\widehat{\beta}_{ABLV} = 2.2$. Consistent with import shares reported in their Table 1, we randomly choose 30% of varieties in each market to be called *Imports*. Consistent with their Figure 1, we reduce the price of

*Imports* by 4 percent, and *Domestics* by 1 percent. We then recompute demand. For each household type $i = (a, e)$, we compute budget shares pre- $(t-1)$ and post $(t)$ the change in prices. We then estimate their regression. We first write the regression out with individual fixed effects, and then first difference to obtain the regression we estimate on data from the model:

$$\log\left(\frac{b^i_{Mt}}{b^i_{Dt}}\right) - \log\left(\frac{b^i_{Mt-1}}{b^i_{Dt-1}}\right) = \beta_0 + \beta_{ABLV}\log e^i \quad \underbrace{\left[\log\left(\frac{p_{Mt}}{p_{Dt}}\right) - \log\left(\frac{p_{Mt-1}}{p_{Dt-1}}\right)\right]}_{=\Delta\log p_{Mt} - \Delta\log p_{Dt} = (-0.04)-(0.01) = -0.03} + \eta^i_t. \qquad \text{(F1)}$$

Here $\beta_0$ captures the common effect of changes in prices. We choose $\sigma$ to deliver $\widehat{\beta}_{ABLV} = 2.2$.

### F.4. Empirical implementation of Jaimovich et al. (2019) regression

To measure sorting, we follow Jaimovich et al. (2019) who extend the approach of Bils and Klenow (2001) to consumer packaged goods data. Using Kilts-Neilsen data, Jaimovich et al. (2019) find that, within product modules, higher income households buy goods that have higher prices. We extend their analysis as follows. Let $\overline{P}_{umt}$ be the average price of a UPC $u$, in market $m$ in year $t$.[25] We define a market as the interaction of a product group $g$ and geographical area (DMA) $d$. We define groups by *product modules*.[26] Let $q^i_{dt}$ be the within-DMA-year, across-household quantile $\mathtt{q} \in \{1, \ldots, Q\}$ of individual $i$'s total expenditure in the Neilsen data. Let $\omega_{umt}$ be the household's within-market-year expenditure share on UPC $u$. We estimate equation (5) from Jaimovich et al. (2019):

$$\log P^i_{mt} = \lambda_{mt} + \sum_{\mathtt{q}=1}^{Q} \beta^{\mathtt{q}}_{JRWZ} \mathbb{1}\left[q^i_{dt} = \mathtt{q}\right] + \eta^i_{mt} \quad \text{, where} \quad \log P^i_{mt} = \sum_{u\in\{m,t\}} \omega^i_{umt} \log \overline{P}_{umt}. \qquad \text{(F2)}$$

When we use $Q = 5$ quantiles, we find that $\widehat{\beta}^5_{JRWZ} - \widehat{\beta}^1_{JRWZ} = 14.4$.[27] Within-module-DMA-year, high expenditure households purchase products that have $14.4$ percent higher average prices than the products purchased by poor households. In constructing the same moment in the model, we take a cross-section of households, simulate them for one year, construct the regression inputs the same way, and estimate the same regression.

[25] Let $\tau$ be a transaction in the Kilts-Neilsen *Consumer Scanner Data*. We observe the expenditure $x_\tau$, and the units purchased $units_\tau$ (for example, 200mL of shampoo). We compute $\overline{P}_{umt} = \sum_{\tau\in\{u,m,t\}} x_\tau / \sum_{\tau\in\{u,m,t\}} units_\tau$.

[26] Product modules are a categorization of goods within the Neilsen data, however one may be concerned that they are too broad (for example, combining ketchup and mustard). We are currently updating this analysis to group firms using market definitions from Kaplan and Menzio (2015), which are narrower than product modules.

[27] Our analysis extends Jaimovich et al. (2019) in the following ways. First, we use within-module-DMA-year variation, whereas they use within-module and pool across years. This takes care of the possibility that higher income households purchase higher price products that are available only in some geographic areas. Second, we use interacted fixed effects, which make precise the variation being used: within-module-DMA-year, across-household. Third, as in Faber and Fally (2022) we use total annual consumption to split households. This is a continuous measure, rather than the coarse bins for income in Neilsen data, which allows us to construct household income quantiles by within-DMA-year rankings of households. We do not include household observables as controls as they are found to have almost no impact on regression coefficients in Jaimovich et al. (2019).